FISHER INFORMATION AND ITS TEMPORAL DECOMPOSITION FOR DIFFUSION MODEL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligning diffusion models with human preferences requires effective use of costly comparison labels. To assess candidate queries by their effects on generation, we study how learning from an answer changes the output distribution and how information about the resulting update is distributed over generation time. We formulate directional Fisher information for updates of a terminal potential, which reweights the model's output distribution. Using Doob's h-transform, we connect the sensitivity of the score function to a temporal decomposition of this information. The decomposition characterizes when the terminal value of an update becomes predictable from intermediate states. Terminal Fisher information provides a criterion for query selection, while the decomposition supports analysis of generation. Numerical estimates agree with analytic solutions in a low-dimensional setting. On CIFAR-10, color becomes predictable earlier than car-likeness, and the temporal distribution of update information differs between query selection strategies. In direct preference optimization (DPO) with a fixed diffusion generator, Fisher-based query selection on ImageRewardDB yields lower mean test cross-entropy with 128 labels than random selection with 256 across the evaluated settings. Performance is comparable to D-optimal selection at matched label budgets. When the learned potentials select among shared generated candidates, those learned with Fisher-based query selection yield larger gains than those learned with random selection under the automatic evaluator ImageReward.

1 INTRODUCTION

Pretrained diffusion models generate diverse images (Ho et al., 2020; Song et al., 2020), but aligning their outputs with human preferences requires learning which images people prefer. Direct Preference Optimization (DPO) uses pairwise comparisons to adjust the relative probabilities of model outputs (Rafailov et al., 2023; Wallace et al., 2024). More broadly, alignment can be formulated as optimization over output distributions (Kawata et al., 2025). We adopt this distributional perspective and represent the output distribution as $q_f \propto e^{-f} r_T$, where r_T is a reference distribution and f is a terminal potential. Outputs with smaller potential values f receive greater weight relative to other outputs. This representation provides a common basis for studying how updates to f change the output distribution, whether those updates arise from a preference loss or another distributional objective.

Human comparisons are costly to collect, making the selection of informative queries a central challenge in preference learning (Xu et al., 2023; Settles, 2009). Prior work has explored feedback-efficient adaptation of diffusion models through online reward learning and human-guided fine-tuning (Uehara et al., 2024b; Hiranaka et al., 2025). Active DPO methods also select comparisons based on information about preference parameters or uncertainty in reward differences (Kveton et al., 2025; Lin et al., 2026). These approaches motivate a complementary question: how much would learning from each possible answer to a comparison change the generated output distribution?

The importance of a comparison for alignment depends on how learning from its answer changes the generated outputs. An ambiguous comparison may induce an update with little impact on generation. Assessing this impact for selecting a query feedback requires accounting for both the learner's update and the distribution on which it acts. For each candidate comparison, we therefore consider the update induced by each possible answer and quantify the resulting local change in the output

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.

distribution. Averaging over predicted answers yields a query selection criterion based on expected model change (Settles, 2009). We quantify distributional change through the local Kullback–Leibler (KL) divergence, whose second-order coefficient is given by *directional Fisher information*. This quantity captures how an update redistributes probability across outputs and complements information about the comparison labels themselves. We empirically evaluate whether prioritizing this distributional change improves label efficiency in preference learning.

The total amount of distributional change does not reveal when information about an update becomes available during generation. Two updates with similar total Fisher information may differ in when their terminal values become predictable from intermediate states. Neither a terminal loss nor an aggregate measure of change captures this distinction. A temporal account of update information is therefore needed to relate learning updates to the emergence of output attributes and to select comparisons that emphasize particular stages of generation. We investigate both when output attributes become predictable and how the temporal distribution of update information depends on the labels acquired during learning. This connects the effects of learning to the information available at intermediate stages of generation (Stančević & Ambrogioni, 2026; Handke et al., 2026).

Connecting terminal updates to generation time requires a distribution over trajectories, or a path law, which is not uniquely determined by the terminal distribution. We use Doob's h-transform to realize the target terminal distribution while minimizing the path-space KL divergence from the reference process (Léonard, 2013; Uehara et al., 2024a; Kawata et al., 2025). Under the resulting path law, the conditional expectation of the terminal update given an intermediate state characterizes its predictability at that stage. The variance of this conditional expectation quantifies the information already available, while its negative state gradient gives the corresponding score sensitivity. These relationships connect terminal Fisher information to its contributions over generation time.

Our framework supports both comparison selection based on terminal Fisher information and analysis of generation through its temporal decomposition. We make three contributions:

- **Analysis of distributional updates** (§3). We establish connections among score sensitivity, directional Fisher information, and the predictability of terminal potential updates, providing a unified formulation in continuous and discrete generation time.
- **Computation and temporal analysis** (§4, §5.1–5.2). We estimate score sensitivity using a Gaussian transition derivative and cumulative information using independent continuations from a shared intermediate state. The estimates agree with analytic solutions in a low-dimensional setting and predict changes in expected terminal quantities induced by perturbations of intermediate states late in CIFAR-10 generation. On CIFAR-10, color information becomes predictable earlier than car-likeness, and the temporal distribution of update information varies across query selection strategies.
- **Preference learning and image selection** (§5.3–5.4). With a fixed diffusion generator and learned terminal potentials, query selection based on terminal Fisher information achieves lower mean test cross-entropy on ImageRewardDB with 128 labels than random selection with 256 labels. At matched label budgets, its performance is comparable to that of D-optimal query selection. The learned potentials also improve image selection as measured by the automatic evaluator ImageReward, with larger gains when queries are selected using Fisher information.

2 Related Work

Distributional optimization and diffusion alignment Reinforcement learning from human feedback (RLHF) learns rewards from preferences to align language models (Ouyang et al., 2022). DPO learns directly from comparisons (Rafailov et al., 2023), and Diffusion-DPO extends this approach to diffusion models (Wallace et al., 2024). Kawata et al. (2025) connect distributional optimization by Dual Averaging (DA) (Nesterov, 2009; Nitanda et al., 2021) to generation through Doob's h-transform. Uehara et al. (2024a) formulate diffusion fine-tuning as entropy-regularized control, relating terminal rewards to time-dependent value functions and optimal drifts. We build on this correspondence to analyze a given update: how it changes the distribution and how its information is distributed over time.

Feedback-efficient adaptation and active preference learning SEIKO uses reward uncertainty to explore during online diffusion fine-tuning (Uehara et al., 2024b). HERO learns from human evaluations and uses preferred images to guide subsequent generation (Hiranaka et al., 2025), while PAPA adapts diffusion models from online binary feedback on individual images (Sarkar et al., 2026). Generative active learning also selects synthetic training data for image synthesis personalization (Zhang et al., 2024). These methods establish feedback efficiency as a diffusion alignment problem, while coupling data collection to generator adaptation. Our experiment instead isolates the selection of pairwise labels under a common potential learner and fixed generator. For this comparison, D-optimal design (Kveton et al., 2025) provides a criterion based on information in comparison labels that directly fits the linear parameterization (§5.3). ActiveDPO measures uncertainty in the current language model's implicit reward differences using their gradients with respect to model parameters (Lin et al., 2026). This is distinct from the predictive ambiguity used by our uncertainty-sampling baseline. Our criterion evaluates the update following each possible answer using the Fisher information of the output distribution. With linear potentials, this is a quadratic form in the feature covariance under that distribution, rather than the information matrix accumulated from comparisons (Appendix D).

Sensitivity and information of distributional updates Scarvelis & Solomon (2025) differentiate diffusion scores under perturbations of the data distribution and propagate these changes to generated samples. Here, terminal potential updates perturb the output distribution, and their temporal effects are defined through a reference generative path law. Infinitesimal reward tilts (Lu, 2026) and reward responses to drift changes (Jiao et al., 2025b) provide related sensitivity formulas. Path-space Fisher information decomposes into initial and transition contributions (Arampatzis et al., 2015). We specialize these relations to terminal updates and connect their temporal information to finite-step computation and the analysis of learning updates.

Generation time and preference optimization Information dynamics and class emergence have been studied through entropy and score-based quantities along diffusion trajectories (Stančević & Ambrogioni, 2026; Handke et al., 2026). We extend this viewpoint to updates produced by learning, comparing their predictability across generation times and training stages alongside that of individual attributes. Time also enters preference optimization: step-by-step preference optimization (SPO) ranks intermediate candidates for training (Liang et al., 2025), and Yang et al. (2024) discount contributions along the trajectory in a preference objective. These methods change the training objective or its supervision along the trajectory. Our time-weighted extension keeps the learner and terminal comparison labels fixed and weights interval Fisher information to select which comparisons to request (§5.2, Appendix F.6).

Computation and image selection Our estimators use a Gaussian likelihood-ratio derivative and the pick-freeze covariance identity (Janon et al., 2014). The continuous-time derivative is related to the Bismut–Elworthy–Li formula (Bismut, 1984; Elworthy & Li, 1994). Gradient estimation via quadratic covariation (Guo et al., 2026) is discussed in Appendix C.2. These tools estimate either the direction of a local response or cumulative information, as required by the analysis. We assess learned potentials by selecting generated images, using the connection between tilted distributions and soft best-of-N (Jiao et al., 2025a) and systematic resampling (Douc et al., 2005).

3 Setting and Main Theory

We ask two questions about a given learning update: how much does it change the output distribution locally, and when does information about that change become available during generation? The learning algorithm supplies an update of the terminal potential. The following analysis measures its effect independently of how that update was computed.

3.1 Potentials, updates, and path laws

Generation proceeds from noise to output with states X_t at times $t \in [0, T]$, or X_j at steps $j = 0, \ldots, N_{\mathrm{gen}}$. The training stage k counts comparison labels used for learning and is distinct from j. We initially fix and omit the prompt c.

Let R be the reference Markov path law, with initial distribution μ_R and terminal distribution r_T. For a finite-valued measurable potential f with $0 < Z_f = \mathbb{E}_R e^{-f(X_T)} < \infty$, define

$$q_f(\mathrm{d}y) = Z_f^{-1} e^{-f(y)} r_T(\mathrm{d}y), \qquad P^f(\mathrm{d}\omega) = Z_f^{-1} e^{-f(X_T(\omega))} R(\mathrm{d}\omega). \tag{1}$$

Thus the output and the entire trajectory ω receive the same terminal weight. Write p_t^f and r_t for the time-t marginals of P^f and R. We study $f + \varepsilon g$ for fixed f and update direction $g = f_{\text{after}} - f_{\text{before}}$, writing $P = P^f$ and $\bar{g} = \mathbb{E}_P g(X_T)$. The scalar ε defines the local analysis, whereas the actual learning update corresponds to $\varepsilon = 1$. In latent-space generation, f and g include evaluation through the deterministic decoder.

Condition L requires P to be Markov and $g(X_T) \in L^2(P)$; Condition E requires $\mathbb{E}_P e^{\eta|g(X_T)|} < \infty$ for some $\eta > 0$. Conditions S and N provide sufficient regularity for continuous-time and Gaussian-transition differentiation, respectively (Appendices A–B).

3.2 Path laws and directional Fisher information

A terminal distribution can be realized by many path laws. When $D_{\mathrm{KL}}(q_f \| r_T) < \infty$, P^f uniquely minimizes $D_{\mathrm{KL}}(Q\|R)$ over those with terminal distribution q_f. It changes the probability of each terminal outcome while preserving the reference law of trajectories conditional on that outcome. This is Doob's h-transform (Léonard, 2013; Chopin et al., 2023), with $h_t^f(x) = \mathbb{E}_R[e^{-f(X_T)} \mid X_t = x]$ and initial distribution $\mu_f = h_0^f \mu_R / Z_f$ (Appendix A.2). Under Condition E, the statistical score and directional Fisher information of the family $P^{f+\varepsilon g}$ at $\varepsilon = 0$ are

$$S_T = -(g(X_T) - \bar{g}), \qquad I_T^f(g) = \mathbb{E}_P S_T^2 = \mathrm{Var}_{q_f}(g). \tag{2}$$

Here S_T differentiates with respect to ε, unlike the state-space score $\nabla_x \log p_t$. The information measures the local change of the distribution:

$$D_{\mathrm{KL}}(P^{f+\varepsilon g} \| P^f) = \tfrac{1}{2}\varepsilon^2 I_T^f(g) + O(|\varepsilon|^3). \tag{3}$$

This path-space KL equals the corresponding terminal KL (Appendix A). Thus the path representation adds a temporal description without adding to the total local information of the output update. Averaging this quantity over the possible answers to a comparison gives the acquisition criterion defined in §5.2 and evaluated in §5.3.

3.3 Conditional mean and temporal decomposition

Under Condition L, define

$$m_t^f(x) = \mathbb{E}_P[g(X_T) \mid X_t = x], \qquad M_t = m_t^f(X_t), \qquad I_t = \mathrm{Var}_P(M_t).$$

The conditional mean is the best predictor of the terminal update value from X_t under squared error. Accordingly, I_t is the reduction in prediction error relative to using the constant $\bar{g}$. By the Markov property, $M_t = \mathbb{E}_P[g(X_T) \mid \mathcal{F}_t]$ is a martingale, where $\mathcal{F}_t$ is the information carried by the trajectory up to t. For the continuous-time reference process, write

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W_t, \qquad a(t) = \sigma(t)\sigma(t)^\top,$$

where b is the drift, σ the noise coefficient, and W_t standard Brownian motion.

Proposition 1 (Sensitivity and temporal decomposition)**.**

(i) Under Condition L, $I_T - I_t = \mathbb{E}_P[\mathrm{Var}_P(g(X_T) \mid X_t)] \geq 0$, and $I_s \leq I_t$ for $s \leq t$. Under Condition E, the score of the path observed up to t is $-(M_t - \bar{g})$, with Fisher information I_t.

(ii) Under Condition S,

$$G_t^f[g](x) := \partial_\varepsilon \nabla_x \log h_t^{f+\varepsilon g}(x)\big|_{\varepsilon=0} = -\nabla_x m_t^f(x). \tag{4}$$

The drift of P^f is $b + a\nabla \log h_t^f$, with update derivative $aG_t^f[g]$, and

$$I_T^f(g) = I_0 + \int_0^T \iota_t\,\mathrm{d}t, \qquad \iota_t = \mathbb{E}_P[G_t^\top a(t) G_t]. \tag{5}$$

(iii) For a Markov chain with N_{gen} steps satisfying Condition L, define $M_j = \mathbb{E}_P[g(X_{N_{\mathrm{gen}}}) \mid X_j]$, $I_j = \mathrm{Var}_P(M_j)$, and $J_j = \mathbb{E}_P[(M_{j+1} - M_j)^2]$. Then

$$J_j = I_{j+1} - I_j \geq 0, \qquad I_{N_{\mathrm{gen}}} = I_0 + \sum_{j=0}^{N_{\mathrm{gen}}-1} J_j. \tag{6}$$

Under Condition E, J_j is the mean conditional Fisher information of transition j.

To understand the discrete decomposition, consider how each transition revises the conditional prediction of the terminal update. Each revision has zero conditional mean given the preceding trajectory, so revisions at different steps are orthogonal. Their variances therefore add without cross terms, while I_0 accounts for information already present in the initial state.

Where the marginal densities are positive and differentiable, $p_t^f = r_t h_t^f / Z_f$ identifies G_t as the sensitivity of the state-space score to the potential update. It also predicts how perturbing a state changes the expected terminal quantity: for a unit vector v, the directional derivative is $-v^\top G_t$. We test this prediction in the experiment on local responses in §5.1.

The decomposition gives a complementary interpretation: I_t/I_T (when $I_T > 0$) is the fraction of the update's variation predictable by time t, and J_j is the additional variation resolved by the next step. These quantities allow updates to be compared by their temporal profiles, as in §5.2. Our image analysis uses the exact finite-step identity (6). It requires no limit of vanishing step size and retains information already present in the initial state through I_0.

4 Computing Sensitivity and Information

Sensitivity of the score function The vector $G_j^f[g] = -\nabla_x m_j^f$ is needed to choose a state perturbation and predict its effect. We estimate it with a Gaussian likelihood-ratio derivative, analogous to the Bismut–Elworthy–Li integration-by-parts formula (Appendix C.1). Under Condition N, for $f = 0$ the calculation averages the product of the first step's Gaussian noise and the terminal value g. Only the first mean map is differentiated. Subsequent transitions and the terminal evaluator enter through samples. The precise formula and its extension to nonzero f are given in Appendix B.1. A quadratic form of G_j need not equal the finite-step information J_j, which also includes nonlinear dependence on the step's noise (Appendix B.2).

Cumulative information Temporal analysis requires $I_j = \mathrm{Var}(m_j(X_j))$, rather than a perturbation direction. Taking the variance of a noisy estimate of m_j adds estimation error. Instead, for $Z \sim p_j^f$, we use two conditionally independent, square-integrable estimates A, B with $\mathbb{E}[A \mid Z] = \mathbb{E}[B \mid Z] = m_j^f(Z)$. The pick-freeze identity gives $\mathrm{Cov}(A, B) = I_j$ (Janon et al., 2014). The sample covariance over independent states is unbiased (Appendix B.3). For $f = 0$, A, B are sample means from independent continuations. This needs neither G nor derivatives. For a fixed terminal feature vector Φ and linear directions $g_\theta = \theta^\top \Phi$, a shared matrix gives $\widehat{I}_j(g_\theta) = \theta^\top \widehat{\mathcal{C}}_j \theta$ for all updates. Appendix B.3 also treats weighting and bias for nonzero f.

5 Experiments

We first validate the computations, then analyze temporal structure and evaluate preference learning and image selection. Both image generators are pretrained and remain fixed. Detailed settings are in Appendix E.

5.1 Numerical validation

Figure 1 compares estimates of G and ι_t with analytic solutions for a two-dimensional Ornstein–Uhlenbeck process with a quadratic potential. Across four update directions, the discrepancy between $I_0 + \int_0^T \iota_t \, \mathrm{d}t$ and I_T is at most 0.92% with analytic G and 5.8% with estimated G. These checks validate the decomposition and the derivative calculation, respectively (Appendix F.1, Figure 5 and Table 1; settings in Appendix E).

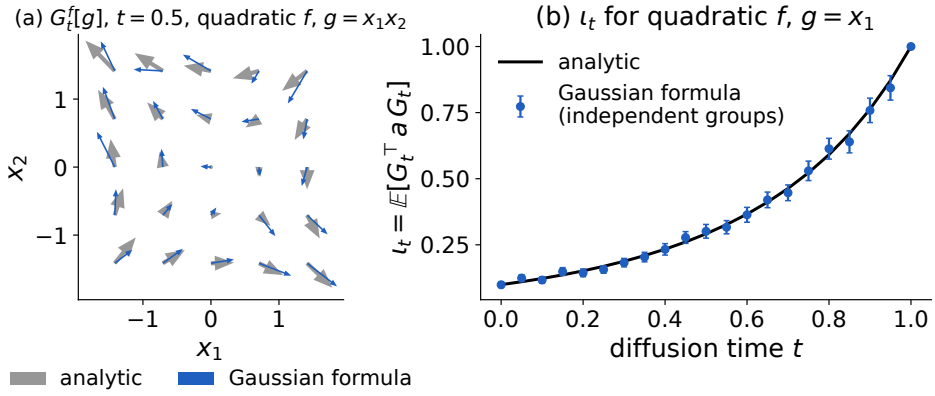


Figure 1: Validation for the two-dimensional Ornstein–Uhlenbeck process with a quadratic potential. (a) Analytic and estimated $G_t^f[g]$ for $g = x_1x_2$ at $t = 0.5$. (b) Analytic ι_t and estimates from the product of two independent estimates of G, for $g = x_1$. Error bars: 1.96 standard errors.

Local responses in image generation We test local response predictions with the public denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) checkpoint `google/ddpm-cifar10-32`, pretrained on CIFAR-10, with fixed weights and 100 generation steps. Synthetic preferences use the mean red–blue difference (COLOR), automobile probability (CLASS), proximity of the mean Sobel gradient magnitude to a target (LOCAL), or their combination (MIX), as detailed in Appendix E.2.

Predictions and measurements from independent samples agree more closely late in generation. Regression slopes at $j = 90$ are 0.80–0.82 for MIX updates and 0.95–0.97 for COLOR, versus 0.30–0.38 for MIX at $j = 60$ (Appendix F.2, Figure 6). This tests $-v^\top G_j$ under state perturbations. The following temporal analysis instead computes information directly from independent continuations.

5.2 Temporal analysis across attributes and training stages

We analyze when output attributes and learned updates become predictable during CIFAR-10 generation.

Query selection and analysis setup For a comparison $z = (c, a, b)$ of two images with prompt c, let $y = 1$ prefer a and $y = 0$ prefer b. For each response, we compute the update $g_{z,y} = f_{\text{after}}(z, y) - f_{\text{before}}$ from adding that label to the current training state, and select the comparison maximizing

$$A(z) = \sum_{y \in \{0,1\}} p_f(y \mid z)\, I_T^\nu(f; g_{z,y}), \tag{7}$$

where $p_f(y \mid z)$ is the predicted response probability and $I_T^\nu(f; g)$ averages $\mathrm{Var}_{q_f(\cdot|c)} g(c, \cdot)$ over fixed nonnegative prompt weights ν summing to one (Appendix A.3). For unconditional generation, the prompt is omitted. We call this Fisher selection and use it here and in §5.3. By (3), it prioritizes updates expected to change the generative distribution substantially. It uses terminal information under the current q_f, without computing G or I_j. Random selection uses the same learner and candidate-generation procedure.

For temporal analysis, we save the training states after k labels and compare updates from adding one more label under the common reference process $f = 0$. Using the shared matrix of §4, we estimate $I_j^{\text{ref}}(g) = \mathrm{Var}_R\{\mathbb{E}_R[g(X_{N_{\text{gen}}}) \mid X_j]\}$. Averaging each response's information with its predicted probability, then over runs and candidates, gives $\mathcal{I}_{k,j}$. The ratio $\mathcal{I}_{k,j}/\mathcal{I}_{k,N_{\text{gen}}}$ is the fraction predictable by j. The averaging weights do not depend on j, preserving the decomposition (Appendix E.2). The decomposition along j is theoretical, whereas its change with training stage k is studied experimentally.terminal step.

Attributes Figure 2(a) shows $I_j/I_{N_{\text{gen}}}$ for each attribute score as g. At $j = 40$, the fractions are 54.2% for COLOR, 11.2% for CLASS, and 10.3% for LOCAL: color is predictable earlier

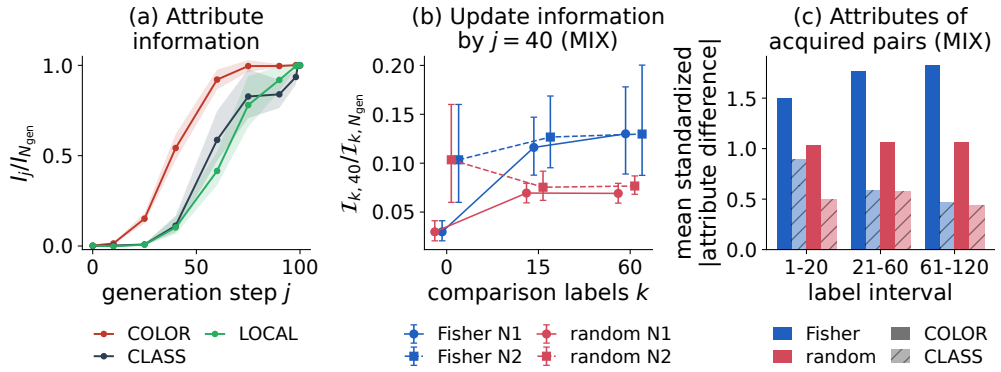


Figure 2: Information across generation times and training stages on the common reference process $f = 0$. (a) Cumulative fractions for the three attribute scores. (b) Fraction predictable by $j = 40$ for the MIX update after k labels; N1/N2 are independent reference image sets. Bands and error bars: 95% confidence intervals from resampling states. (c) Mean absolute differences in standardized attribute scores of selected MIX pairs, averaged equally over N1/N2. Blue/red denote Fisher/random query selection in (b,c).

than the other attributes. By Proposition 1(i), over half of the variation in the final color score is already predictable at this step, while most variation in the other two scores remains unresolved. The temporal profiles distinguish these attributes even though every normalized profile reaches one at the terminal step.

Training stages Figure 2(b) compares MIX updates after training with Fisher or random selection on two independent reference image sets, N1 and N2. At each $k \in \{0, 15, 60\}$, we evaluate the candidate maximizing (7). At $k = 0$, the fraction predictable by $j = 40$ is 0.03 for N1 and 0.10 for N2. After 15 and 60 labels, the fraction predictable by $j = 40$ is 0.12–0.13 for Fisher selection versus 0.07–0.08 for random selection in both sets. Thus, in this MIX setting, training with Fisher selection yields a larger share of update information predictable by $j = 40$ than random selection at the same label count.

Figure 2(c) shows that Fisher selection increasingly chooses pairs with larger color differences, whereas random selection changes little. Car-likeness differences decrease for both methods. The shift toward color comparisons is consistent with earlier predictability of color and the larger early fraction of update information. Using the same reference process for all training states lets us relate the kinds of comparisons selected to the temporal profiles of the resulting updates. Full temporal curves and alternative candidate choices are reported in Appendix F.3 (Figure 7, Table 2).

Using the temporal analysis for query selection Weighting generation intervals changes which comparisons are selected. On the reference process $f = 0$, quadrupling the early interval $j = 0$–40 for COLOR favors comparisons with larger color differences, consistent with their earlier predictability. Learning performance changes little. The same tendency appears among the ten highest-ranked candidates when the learning state and candidate pairs are held fixed (Appendix F.6). All weights equal to one recover terminal information on the same reference process, isolating the effect of time weighting. The interpretation of nonnegative weights is given in Appendix B.4.

5.3 Active preference learning

We test whether active learning (AL) with Fisher selection (§5.2) reduces the labels needed to predict unseen comparisons. The reference generator is pretrained Stable Diffusion v1.5 (Rombach et al., 2022), with fixed weights. Preference learning changes the terminal potential. DA with ridge regularization learns $f_\theta(c, x) = \theta^\top \phi(c, x)$ on compressed Bootstrapping Language-Image Pre-training (BLIP) features (Li et al., 2022). Fisher information is computed from the feature covariance of reference-generated images weighted by e^{-f} (Appendix D).

We simulate AL by revealing existing ImageRewardDB (Xu et al., 2023) comparison labels for image–text alignment (ALIGNMENT) and overall ranking (OVERALL). Three training sets A–C with three seeds each share 16 initial labels between methods and use up to 256 labels. Evaluation

covers 12 task/train/test conditions using prompt-disjoint subsets constructed within the official test split, with TEST_NEW as the primary subset (Appendix E.3).

We compare Fisher selection with random selection, uncertainty sampling, and D-optimal selection based on Algorithm 1 of Kveton et al. (2025), using the same learner, features, and candidate rule. Uncertainty sampling maximizes $\pi(1-\pi)$, where π is the predicted probability of preferring the first image. D-optimal selection maximizes the increase in the log determinant of the accumulated Fisher information matrix of comparison labels. Its regularization is selected on validation data (Appendices D.3, E.3).

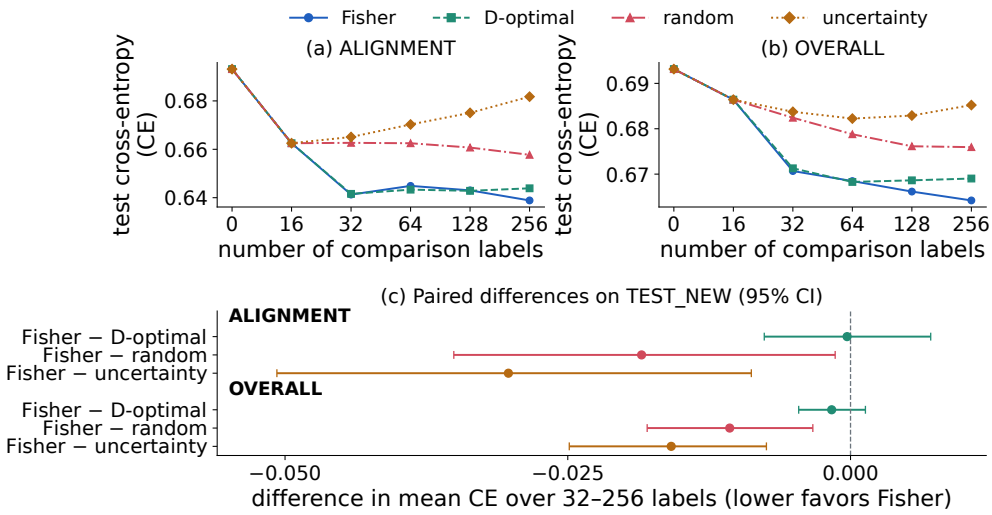


Figure 3: Preference learning on ImageRewardDB with four query selection rules. (a,b) Test CE on TEST_NEW, averaged equally over training sets A–C and three seeds per set. (c) Paired differences in mean CE over 32–256 labels, averaged over label budgets on a logarithmic scale; negative values favor Fisher selection. Intervals are 95% confidence intervals from resampling prompts with the nine training runs fixed (Appendix E.3).

Figures 3(a,b) plot test cross-entropy (CE) against acquired labels. CE compares predicted response probabilities with labels and is averaged over pairs within each prompt, then equally over prompts. Fisher and D-optimal selection both yield lower CE than random selection, whereas uncertainty sampling performs worse. Across the 12 conditions, 128 Fisher-selected labels give lower mean CE than 256 randomly selected labels, with 95% confidence intervals excluding zero in 9. At 256 labels per method, all 12 intervals are negative (Appendix F.4, Table 3). Thus Fisher selection achieves lower prediction loss with half as many labels in these experiments.

Figure 3(c) compares mean CE over 32–256 labels on TEST_NEW. Fisher-minus-D-optimal differences are -0.0003 for ALIGNMENT and -0.0017 for OVERALL. The results suggest comparable performance, with slightly lower mean loss for Fisher, although both 95% intervals include zero. Fisher also outperforms uncertainty sampling in both tasks (Appendix F.4, Table 4).

What the selection criteria measure D-optimal selection evaluates information for estimating the preference parameters from comparison labels. Our criterion instead evaluates the local change in the output distribution caused by the learner's update after each possible answer. Our criterion is based on the same information measure decomposed over generation time in §5.2.

Post-hoc feature comparisons suggest that maximizing ambiguity alone may favor similar feature vectors that provide limited information about preference parameters (Appendix F.4, Table 5).

5.4 Image selection with the learned potential

RESAMPLE uses weights proportional to $e^{-\lambda f}$, with selection strength $\lambda = 4$, to select 16 of 48 candidates per prompt from the fixed Stable Diffusion generator, using systematic resampling

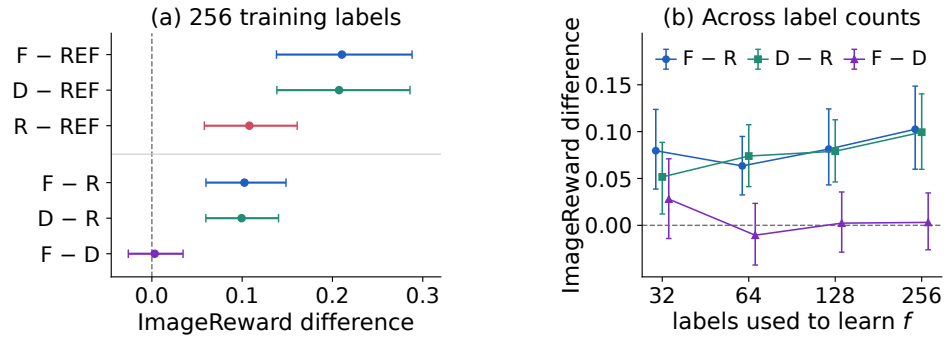


Figure 4: Image selection by RESAMPLE (64 prompts, two seeds, selection strength $\lambda = 4$). (a) ImageReward differences at 256 labels. (b) Differences across label counts. F/D/R: Fisher/D-optimal/random query selection. Error bars: 95% pointwise confidence intervals from prompt resampling.

(Douc et al., 2005). REF uses the first 16. Potentials learned by Fisher, D-optimal, and random query selection share candidates, with 64 prompts and two training seeds (Appendix E.4).

Figure 4(a) shows ImageReward gains of 0.21, 0.21, and 0.11 for Fisher, D-optimal, and random query selection, respectively, at 256 labels. Figure 4(b) shows that Fisher and D-optimal exceed random selection at 32–256 labels, with no detected Fisher–D-optimal difference. Fisher at 128 labels also exceeds random at 256 by 0.094 $[0.053, 0.140]$ (Appendix F.5, Table 6). The benefit of query selection therefore extends to finite-candidate image selection under an automatic evaluator.

With the candidate images shared across methods, this experiment evaluates how the learned potentials reweight a common set of generated outputs. It complements test CE by assessing learned preferences through image selection as well as comparison prediction.

6 Conclusion

We formulated directional Fisher information for terminal potential updates in diffusion models, connected it to the sensitivity of the score function, derived its temporal decomposition, and verified the estimators numerically. On CIFAR-10, attributes become predictable at different generation times, and the temporal distribution of update information differs between query selection strategies. In the DPO experiments, selecting queries by terminal Fisher information achieves lower comparison loss with fewer labels than random selection and performs comparably to D-optimal selection, while the learned potentials improve image selection under an automatic evaluator. This framework connects the effect of learning updates on the output distribution with the temporal structure of their information, supporting both preference query selection and analysis of the generation process.

7 Limitations

The theory concerns fixed f, g and infinitesimal updates, not repeated-training guarantees. Gaussian-transition estimates of G predict responses less accurately mid-generation (§5.1). Full-vector accuracy remains to be established. Unbiasedness for fixed g does not directly cover updates learned or selected using the estimation samples (Appendix B.3).

The temporal analysis compares updates under a common reference process, rather than tracking the trained generative process. The larger early fraction of update information and the shift toward comparisons with larger color differences (§5.2) are consistent, but a causal relation between them has not been established. Human-preference AL uses terminal Fisher information, not the temporal decomposition. In a supplementary experiment, generating images by shifting Gaussian transition means using a learned h (GUIDED) did not improve on REF. Learning errors in h, transition approximation, and a fixed initial distribution can cause deviations from the target law (Appendices C.3, D.4, F.5). Image selection is evaluated automatically by ImageReward, which is closely related to the training data. No new human evaluation was conducted.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation; developing and refining the theoretical framework and mathematical claims; identifying key proof steps and drafting proofs; designing, implementing, and running experiments; and interpreting results. We also used these tools to retrieve, summarize, and analyze related literature. Verification included comparisons with analytical solutions, numerical consistency checks, and cross-checks against source literature. The authors are responsible for the final manuscript and all accompanying claims, results, and artifacts, including those developed with AI assistance.

REFERENCES

Georgios Arampatzis, Markos A. Katsoulakis, and Yannis Pantazis. Pathwise sensitivity analysis in transient regimes, 2015. URL https://arxiv.org/abs/1502.05430. arXiv:1502.05430.

J.M. Bismut. *Large Deviations and the Malliavin Calculus*. Progress in Mathematics. Birkhäuser Boston, 1984. ISBN 9783764332204. URL https://books.google.co.jp/books?id=UXLvAAAAMAAJ.

Nicolas Chopin, Andras Fulop, Jeremy Heng, and Alexandre H Thiery. Computational Doob's h-transforms for online filtering of discretely observed diffusions. In *International Conference on Machine Learning*, pp. 5904–5923. PMLR, 2023.

Randal Douc, Olivier Cappé, and Eric Moulines. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pp. 64–69. IEEE, 2005. doi: 10.1109/ISPA.2005.195385.

K David Elworthy and Xue-Mei Li. Formulae for the derivatives of heat semigroups. *Journal of Functional Analysis*, 125(1):252–286, 1994.

Zhengyi Guo, Wenpin Tang, and Renyuan Xu. Conditional diffusion guidance under hard constraint: A stochastic analysis approach. *arXiv preprint arXiv:2602.05533*, 2026.

Florian Handke, Dejan Stančević, Felix Koulischer, Thomas Demeester, and Luca Ambrogioni. The entropic signature of class speciation in diffusion models. *arXiv preprint arXiv:2602.09651*, 2026.

Ayano Hiranaka, Shang-Fu Chen, Chieh-Hsin Lai, Dongjun Kim, Naoki Murata, Takashi Shibuya, WeiHsiang Liao, Shao-Hua Sun, and Yuki Mitsufuji. Hero: Human-feedback efficient reinforcement learning for online diffusion model finetuning. In *International Conference on Learning Representations*, volume 2025, pp. 54942–54971, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Alexandre Janon, Thierry Klein, Agnes Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two Sobol index estimators. *ESAIM: Probability and Statistics*, 18: 342–364, 2014.

Yuchen Jiao, Yuxin Chen, and Gen Li. Connections between reinforcement learning with feedback, test-time scaling, and diffusion guidance: An anthology. *arXiv preprint arXiv:2509.04372*, 2025a.

Yuchen Jiao, Yuxin Chen, and Gen Li. Towards a unified framework for guided diffusion models. *arXiv preprint arXiv:2512.04985*, 2025b.

Ryotaro Kawata, Kazusato Oko, Atsushi Nitanda, and Taiji Suzuki. Direct distributional optimization for provable alignment of diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 72792–72840, 2025.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Branislav Kveton, Xintong Li, Julian McAuley, Ryan Rossi, Jingbo Shang, Junda Wu, and Tong Yu. Active learning for direct preference optimization. *arXiv preprint arXiv:2503.01076*, 2025.

Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13199–13208. IEEE, 2025.

Xiaoqiang Lin, Arun Verma, Zhongxiang Dai, Daniela Rus, See-Kiong Ng, and Bryan Kian Hsiang Low. Activedpo: Active direct preference optimization for sample-efficient alignment. In *International Conference on Learning Representations*, volume 2026, pp. 21618–21635, 2026.

Jianfeng Lu. A mathematical introduction to diffusion models. *arXiv preprint arXiv:2607.01693*, 2026.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Particle dual averaging: Optimization of mean field neural network with global convergence rate analysis. *Advances in Neural Information Processing Systems*, 34:19608–19621, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

David Pollard. A note on insufficiency and the preservation of Fisher information, 2012. URL `https://arxiv.org/abs/1107.3797v2`. arXiv:1107.3797v2.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

L Chris G Rogers and David Williams. *Diffusions, Markov processes, and martingales*, volume 2. Cambridge University Press, 2000.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685. IEEE, 2022.

Anindya Sarkar, Nasik Muhammad Nafi, Isaac Lyngaas, Muralikrishnan Gopalakrishnan Meena, and Yevgeniy Vorobeychik. Papa: Online personalized active preference alignment. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 666–683. Springer, 2026.

Christopher Scarvelis and Justin Solomon. Sensitivity analysis for diffusion models. *arXiv preprint arXiv:2509.23092*, 2025.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, Department of Computer Sciences, 2009. URL `https://burrsettles.com/pub/settles.activelearning.pdf`.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Dejan Stančević and Luca Ambrogioni. The information dynamics of generative diffusion. *Entropy*, 28(2):195, 2026.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Tommaso Biancalani, and Sergey Levine. Fine-tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint arXiv:2402.15194*, 2024a.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Sergey Levine, and Tommaso Biancalani. Feedback efficient online fine-tuning of diffusion models. *arXiv preprint arXiv:2402.16359*, 2024b.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. *arXiv preprint arXiv:2402.08265*, 2024.

Xulu Zhang, Wengyu Zhang, Xiaoyong Wei, Jinlin Wu, Zhaoxiang Zhang, Zhen Lei, and Qing Li. Generative active learning for image synthesis personalization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10669–10677, 2024.

A Conditions and Proofs of the Propositions in the Main Text

We state the assumptions for each result and then prove the h-transform representation, the local Fisher identity, and the temporal decomposition. The path spaces are standard Borel spaces (continuous paths for diffusions and finite products for chains), so regular conditional laws are available. We use the natural filtration $\mathcal{F}_t = \sigma(X_s : s \le t)$, with its usual augmentation in continuous time. Conditional identities are understood almost surely under the relevant law. Throughout, let $N_t^f(x) = \mathbb{E}_R[e^{-f(X_T)} g(X_T) \mid X_t = x]$. Continuous-time differentiation requires stronger assumptions than the finite-step variance decomposition.

A.1 Conditions

- Condition L: $g(X_T) \in L^2(P^f)$ and P^f is Markov. It is used for the conditional mean, the discrete decomposition, and the population identity of pick-freeze.
- Condition E: $\mathbb{E}_{P^f} e^{\eta |g(X_T)|} < \infty$ for some $\eta > 0$. It is used for the local exponential family in both directions, the second-order expansion of the KL divergence, and the Fisher information with respect to the update coefficient. E does not follow from L.
- Condition S (a sufficient condition for continuous diffusions): on a finite time interval, $X_0 \sim \mu_R$ is independent of the Brownian motion and has a finite second moment. b is C^3 in space, b and its spatial derivatives of orders 1–3 are continuous in (t, x), the spatial derivatives of orders 1–3 are uniformly bounded, and $b(t, 0)$ is bounded. σ is a $d \times d$ matrix that is C^1 in time, with σ and σ^{-1} uniformly bounded. $f, g \in C_b^3(\mathbb{R}^d)$. Under this condition, a strong solution and the spatial derivative flow exist, h_t^f and N_t^f are $C^{1,2}$ functions with bounded spatial derivatives, and h_t^f is bounded above and below by positive constants. These hold uniformly in a small neighborhood of $f + \varepsilon g$, so that the derivative with respect to the state and the derivative with respect to ε can be interchanged. S is a sufficient condition for the continuous-time results. For the finite-step implementation on images, Proposition 1(iii) and Condition N are applied.
- Condition N: the j-th transition is $X_{j+1} = \mu_j(x) + L_j \xi$ with $\xi \sim N(0, I_d)$, L_j is an invertible matrix independent of x, and μ_j is C^1 in a neighborhood of the state of interest. Let $k_x(z)$ be this transition density, $\xi_x(z) = L_j^{-1}(z - \mu_j(x))$, and $H(z) = \mathbb{E}_R[e^{-f}(1 + |g|) \mid X_{j+1} = z]$, where f, g are evaluated at the terminal state. On some neighborhood U, assume $H(z) \sup_{x' \in U} \{k_{x'}(z)(1 + \|\xi_{x'}(z)\|) + \|\nabla_{x'} k_{x'}(z)\|\}$ has an integrable envelope. This ensures that the noise-weighted moments are finite and permits differentiation of both numerator and denominator. Bounded f, g satisfy this condition after taking U sufficiently small.
- Condition MC: the estimation target g is fixed, the outer states have the target marginal law, and the two estimates are conditionally independent and unbiased given the state, with finite second moments. The complete triples consisting of the state and its two estimates are independent and identically distributed across outer samples. In particular, the randomness used for continuations at distinct states is independent.

Condition L alone suffices for the martingale variance identities. It need not define an exponential family for both signs of ε: for example, $g(X) = X^4$ with Gaussian X has a finite second moment but $\mathbb{E} e^{|\varepsilon| X^4} = \infty$. This is why the statistical-score and KL statements use Condition E separately.

To justify the regularity asserted under S, the spatial derivative flow satisfies $\dot{J}_{t,s} = D_x b(s, X_s) J_{t,s}$ and $\|J_{t,s}\| \le e^{C(s-t)}$. Its higher spatial derivatives are also bounded on the finite time interval by their linear differential equations. Differentiating the expectations of the bounded smooth terminal functions e^{-f} and $e^{-f} g$ therefore gives bounded spatial derivatives of h and N. The backward Kolmogorov equations give their time derivatives. Moreover, $e^{-\|f\|_\infty} \le h \le e^{\|f\|_\infty}$, so the quotient $m = N/h$ is regular as well. The same bounds hold locally uniformly in ε, justifying the mixed derivatives used below. S is sufficient rather than necessary. The unbounded quadratic OU example is justified separately by explicit Gaussian integration in Appendix E.

A.2 h-TRANSFORM AND MINIMALITY OF THE PATH KL DIVERGENCE

The initial distribution and the transition kernel of Doob's h-transform (Rogers & Williams, 2000; Chopin et al., 2023) are given by

$$\mu_f(\mathrm{d}x) = \frac{h_0^f(x)}{Z_f}\,\mu_R(\mathrm{d}x), \qquad K_{s,t}^f(x,\mathrm{d}y) = \frac{h_t^f(y)}{h_s^f(x)}\,K_{s,t}^R(x,\mathrm{d}y), \tag{8}$$

where $K_{s,t}^R$ is the transition kernel of the reference process.

Proposition 2 (h-transform and minimality of the path KL divergence)**.** *Let $h_t^f(x) = \mathbb{E}_R[e^{-f(X_T)} \mid X_t = x]$. Then P^f is Markov, and its initial distribution and transition kernel are given by* (8)*. Moreover, if $D_{\mathrm{KL}}(q_f\|r_T) < \infty$, then among the path laws Q whose terminal distribution is q_f, P^f is the unique minimizer of $D_{\mathrm{KL}}(Q\|R)$, and $D_{\mathrm{KL}}(P^f\|R) = D_{\mathrm{KL}}(q_f\|r_T)$.*

Proof. The density of P^f restricted to $\mathcal{F}_t$ with respect to R is

$$\mathbb{E}_R\left[\frac{e^{-f(X_T)}}{Z_f}\middle|\mathcal{F}_t\right] = \frac{h_t^f(X_t)}{Z_f}.$$

At $t=0$ this gives the initial distribution in (8). By conditional Bayes' rule and the Markov property, for $s \le t$ and bounded measurable u,

$$\mathbb{E}_{P^f}[u(X_t) \mid \mathcal{F}_s] = \frac{\mathbb{E}_R[u(X_t)h_t^f(X_t) \mid X_s]}{h_s^f(X_s)}.$$

Thus the conditional law depends on the past only through X_s, proving the Markov property and the kernel formula. The tower property gives $K_{s,t}^R h_t^f = h_s^f$, so the kernel is normalized.

For minimality, it suffices to consider Q with finite $D_{\mathrm{KL}}(Q\|R)$, since the proposed minimizer has finite KL. Disintegrating such a path law with respect to its terminal value gives

$$D_{\mathrm{KL}}(Q\|R) = D_{\mathrm{KL}}(Q_T\|r_T) + \int D_{\mathrm{KL}}\big(Q(\cdot \mid X_T = y)\,\|\,R(\cdot \mid X_T = y)\big)\,Q_T(\mathrm{d}y),$$

(Léonard, 2013). The density in (1) depends only on X_T, so $P^f(\cdot \mid X_T = y) = R(\cdot \mid X_T = y)$ for q_f-almost every y. It attains the lower bound $D_{\mathrm{KL}}(q_f\|r_T)$. Equality requires these conditional laws to coincide. Together with $Q_T = q_f$, this determines $Q = P^f$ uniquely. If the constraint instead concerns the decoder output $Y = D(X_T)$ and the potential is a function of Y, the same proof uses disintegration with respect to Y. The decoder need not be invertible. □

The construction also applies with the current law P^f as the reference: for sufficiently small ε under Condition E, $P^{f+\varepsilon g}$ minimizes the KL from P^f subject to its updated terminal marginal. The canonical path update therefore describes the change from the current distribution, as well as from the original reference.

A.3 FISHER INFORMATION OF THE UPDATE DIRECTION AND KL DIVERGENCE

Proposition 3 (Fisher information of the update direction and KL divergence)**.** *Under Condition E, for sufficiently small $|\varepsilon|$, $\mathrm{d}P^{f+\varepsilon g}/\mathrm{d}P^f = \exp\{-\varepsilon g(X_T) - \psi(\varepsilon)\}$ with $\psi(\varepsilon) = \log \mathbb{E}_P e^{-\varepsilon g(X_T)}$; the score and the Fisher information of this one-parameter family are given by* (2)*, the KL divergences in both directions satisfy the same second-order expansion as* (3)*, and each of them equals the corresponding KL divergence between the terminal distributions.*

Proof. Taking the ratio of the densities in (1) yields the asserted likelihood ratio. For $|\varepsilon| \le \eta/2$, each $|g|^r e^{-\varepsilon g}$, $r \le 3$, is dominated by a constant times $e^{\eta|g|}$. Differentiation under the expectation is therefore valid. With $P_\varepsilon = P^{f+\varepsilon g}$, it gives

$$\psi'(\varepsilon) = -\mathbb{E}_{P_\varepsilon} g, \qquad \psi''(\varepsilon) = \mathrm{Var}_{P_\varepsilon}(g), \qquad \partial_\varepsilon \log \frac{\mathrm{d}P_\varepsilon}{\mathrm{d}P} = -g - \psi'(\varepsilon).$$

At zero this is $S_T = -(g - \bar{g})$, proving (2). The third derivative of ψ is bounded on a smaller neighborhood of zero. Since $\psi(0) = 0$, Taylor expansion gives

$$D_{\mathrm{KL}}(P_\varepsilon \| P) = \varepsilon\psi'(\varepsilon) - \psi(\varepsilon) = \tfrac{1}{2}\varepsilon^2 I_T + O(|\varepsilon|^3),$$
$$D_{\mathrm{KL}}(P \| P_\varepsilon) = \varepsilon\bar{g} + \psi(\varepsilon) = \tfrac{1}{2}\varepsilon^2 I_T + O(|\varepsilon|^3).$$

The density ratio is a function of the terminal state alone, so integrating either log ratio over paths gives the corresponding terminal KL exactly.

For the projection of the score used below, we also verify differentiability in quadratic mean (DQM). Write $H_\varepsilon = \exp\{-(\varepsilon g + \psi(\varepsilon))/2\}$. Then $H_0' = S_T/2$ and

$$H_\varepsilon'' = \left[\tfrac{1}{4}(g + \psi'(\varepsilon))^2 - \tfrac{1}{2}\psi''(\varepsilon)\right] H_\varepsilon.$$

Condition E bounds $\|H_\varepsilon''\|_{L^2(P)}$ uniformly for $|\varepsilon| \le \eta/4$. The integral Taylor remainder thus gives

$$\left\|H_\varepsilon - 1 - \tfrac{1}{2}\varepsilon S_T\right\|_{L^2(P)} = O(\varepsilon^2) = o(|\varepsilon|),$$

which is DQM with score S_T. □

Averaging across prompts For multiple prompts in a finite set $\mathcal{C}$, we average with fixed weights $\nu(c) \ge 0$, $\sum_c \nu(c) = 1$, as

$$I_T^\nu(f; g) = \sum_{c \in \mathcal{C}} \nu(c) \mathrm{Var}_{q_f(\cdot|c)} g(c, \cdot).$$

This is the Fisher information of the joint law $\nu(c)q_f(\mathrm{d}y \mid c)$ when the prompt law is held fixed. Its score is $-[g(c, y) - \mathbb{E}_{q_f(\cdot|c)} g(c, Y)]$, centered within each prompt. Thus differences of mean update values between prompts are excluded. In particular, adding a prompt-dependent constant to a potential leaves both the conditional output law and its information unchanged.

A.4 Proof of Proposition 1

Proof. (i) Prediction, variance, and Fisher information up to time t. Under Condition L, conditional expectation is an orthogonal projection in $L^2(P)$. The Markov property implies $M_t = \mathbb{E}_P[g(X_T) \mid \mathcal{F}_t] = m_t(X_t)$, and hence $\mathbb{E}[M_t \mid \mathcal{F}_s] = M_s$ for $s \le t$. For any square-integrable predictor $v(X_t)$,

$$\mathbb{E}_P[(g(X_T) - v(X_t))^2] = \mathbb{E}_P[(g(X_T) - M_t)^2] + \mathbb{E}_P[(M_t - v(X_t))^2].$$

The cross term vanishes on conditioning on X_t. Taking $v = \bar{g}$ proves $I_T - I_t = \mathbb{E}_P \mathrm{Var}_P(g(X_T) \mid X_t)$ and the interpretation of I_t as the reduction in prediction error. Similarly, $\mathbb{E}[(M_t - M_s)M_s] = 0$ by the tower property, so

$$I_t - I_s = \mathbb{E}_P[(M_t - M_s)^2] \ge 0.$$

Under Condition E, the likelihood ratio of the path observed up to t is

$$L_{t,\varepsilon} = \mathbb{E}_P[e^{-\varepsilon g(X_T) - \psi(\varepsilon)} \mid \mathcal{F}_t] = e^{-\psi(\varepsilon)} \mathbb{E}_P[e^{-\varepsilon g(X_T)} \mid X_t].$$

It is a function of X_t alone. The exponential moment permits its differentiation at zero, giving

$$-\mathbb{E}_P[g(X_T) - \bar{g} \mid \mathcal{F}_t] = -(M_t - \bar{g}).$$

Proposition 3 establishes DQM. The score-projection theorem (Pollard, 2012) then identifies this derivative as the statistical score of the observed path; its squared norm is I_t. The same score is obtained on observing X_t alone. This equality concerns the terminally tilted family considered here and is not an assertion that an intermediate state preserves all information about an arbitrary path perturbation.

(ii) G and the continuous-time decomposition. Conditional Bayes' rule gives $m_t^f = N_t^f / h_t^f$. Since $\partial_\varepsilon h_t^{f+\varepsilon g}|_0 = -N_t^f$,

$$\partial_\varepsilon \log h_t^{f+\varepsilon g}\big|_0 = -N_t^f / h_t^f = -m_t^f.$$

Interchanging the state and potential derivatives under Condition S gives (4). At times where the marginal densities have spatial scores, $\nabla \log p_t^{f+\varepsilon g} = \nabla \log r_t + \nabla \log h_t^{f+\varepsilon g}$; the normalization has

no state dependence. Differentiating gives the stated interpretation of G. Under S the nondegenerate diffusion has such densities for positive times; no density assumption on the initial law is needed for the information decomposition.

Next identify the drift under P^f. Let $\mathcal{L}_R u = b^\top \nabla u + \frac{1}{2}\operatorname{tr}(a\nabla^2 u)$ be the reference generator. Both h and N solve the backward equation $(\partial_t + \mathcal{L}_R)u = 0$, with terminal data e^{-f} and $e^{-f}g$, respectively. Itô's formula gives $\mathrm{d}h(t, X_t) = \nabla h^\top \sigma \, \mathrm{d}W_t$ under R. The path likelihood ratio factors as

$$\frac{h_t(X_t)}{Z_f} = \frac{h_0(X_0)}{Z_f} \exp\left\{ \int_0^t \zeta_s^\top \mathrm{d}W_s - \tfrac{1}{2}\int_0^t \|\zeta_s\|^2 \mathrm{d}s \right\}, \qquad \zeta_s = \sigma(s)^\top \nabla \log h_s(X_s).$$

The first factor changes the initial distribution to μ_f, leaving the conditional reference dynamics unchanged. Under S, ζ is bounded, so Novikov's condition holds also after this initial change. Girsanov's theorem applied to the second factor gives a Brownian motion $W_t^f = W_t - \int_0^t \zeta_s \mathrm{d}s$ under P^f and drift $b + a\nabla \log h$. Its derivative with respect to the update is aG.

Finally, the product rule for the generator and $N = hm$ give

$$0 = (\partial_t + \mathcal{L}_R)(hm) = h(\partial_t + \mathcal{L}_R)m + (\nabla h)^\top a \nabla m,$$

where the term $m(\partial_t + \mathcal{L}_R)h$ is zero. Dividing by h proves $(\partial_t + \mathcal{L}_f)m = 0$ for $\mathcal{L}_f = \mathcal{L}_R + (a\nabla \log h)^\top \nabla$. Itô's formula under P^f therefore yields

$$M_t - M_0 = -\int_0^t G_s(X_s)^\top \sigma(s) \, \mathrm{d}W_s^f.$$

The integrand is square-integrable under S. The stochastic integral has conditional mean zero given $\mathcal{F}_0$, and is consequently orthogonal to $M_0 - \bar{g}$. Itô's isometry gives

$$\operatorname{Var}_P(M_t) = \operatorname{Var}_P(M_0) + \mathbb{E}_P \int_0^t G_s(X_s)^\top a(s) G_s(X_s) \, \mathrm{d}s.$$

Substituting $M_T = g(X_T)$ at $t = T$ yields (5).

(iii) Finite-step decomposition. Let $D_j = M_{j+1} - M_j$; then $\mathbb{E}[D_j \mid \mathcal{F}_j] = 0$. For $i < j$, D_i is $\mathcal{F}_j$-measurable, so $\mathbb{E}[D_i D_j] = \mathbb{E}[D_i \mathbb{E}[D_j \mid \mathcal{F}_j]] = 0$. The same argument makes D_j orthogonal to $M_0 - \bar{g}$. Taking the variance of $M_{N_{\mathrm{gen}}} = M_0 + \sum_{j=0}^{N_{\mathrm{gen}}-1} D_j$ gives

$$I_{N_{\mathrm{gen}}} = I_0 + \sum_{j=0}^{N_{\mathrm{gen}}-1} \mathbb{E}[D_j^2],$$

Applying the same computation to one interval gives $I_{j+1} - I_j = \mathbb{E}[D_j^2]$. This proves (6) under L alone, including for deterministic or singular kernels.

For the statistical interpretation under E, set $C_{j,\varepsilon}(x) = \mathbb{E}_{P^f}[e^{-\varepsilon g(X_{N_{\mathrm{gen}}})} \mid X_j = x]$. The transition density ratio is

$$\frac{\mathrm{d}K_j^{f+\varepsilon g}(x, \cdot)}{\mathrm{d}K_j^f(x, \cdot)}(x') = \frac{C_{j+1,\varepsilon}(x')}{C_{j,\varepsilon}(x)}.$$

Its logarithmic derivative at zero is $-m_{j+1}(x') + m_j(x)$. This has conditional mean zero under $K_j^f(x, \cdot)$. The conditional Fisher information is therefore $\mathbb{E}[D_j^2 \mid X_j = x]$, whose average over p_j^f is J_j. This ratio argument uses absolute continuity with respect to the transition law itself; a Lebesgue density is unnecessary. □

Evaluation of the directional derivative Fix a unit direction v. If the absolute value of the third directional derivative of m_t is at most B on the segment, the one-variable Taylor formula gives $|(m_t(x + \delta v) - m_t(x - \delta v))/2\delta + v^\top G_t(x)| \le B\delta^2/6$. Splitting the choice of the direction, the prediction, and the measurement of the response into independent groups avoids correlations induced by the selection. Using common random numbers for the measurement of the responses at $x \pm \delta v$ does not affect the unbiasedness of the difference of the two means.

Sparse times and deterministic steps When sparse times $j_0 < j_1 < \cdots$ are measured, the interval information is $I_{j_{r+1}} - I_{j_r} = \mathbb{E}[(M_{j_{r+1}} - M_{j_r})^2]$. A deterministic final step can also be included in (6). If the whole process is deterministic, $g(X_{N_{\mathrm{gen}}})$ is determined by X_0, so $I_0 = I_{N_{\mathrm{gen}}}$ and $J_j = 0$. Hence the general statement that the initial term is always negligible does not hold.

B Details of the Finite-Step Computations

B.1 Gaussian transition derivative and its proof

Proposition 4 (Gaussian transition derivative)**.** *Let $X_{j+1} = \mu_j(x) + L_j\xi$ with $\xi \sim N(0, I_d)$, invertible state-independent L_j, and differentiable μ_j. Assume differentiation and expectation can be interchanged as in Condition N. For $f = 0$,*

$$G_j^0[g](x) = -D\mu_j(x)^\top L_j^{-\top}\mathbb{E}_R[g(X_{N_{\mathrm{gen}}})\xi \mid x]. \tag{9}$$

Subsequent transitions may be arbitrary Markov kernels.

On samples generated to the end with the reference process from the state x, let $w = e^{-f(X_{N_{\mathrm{gen}}})}$, $h = \mathbb{E}_R[w \mid x]$, and $m = \mathbb{E}_R[wg \mid x]/h$. Under Condition N,

$$G_j^f[g](x) = -D\mu_j(x)^\top L_j^{-\top}\,\frac{\mathbb{E}_R[w\,g(X_{N_{\mathrm{gen}}})\,\xi \mid x] - m\,\mathbb{E}_R[w\,\xi \mid x]}{h}, \tag{10}$$

which reduces to (9) when $w = 1$ and $\mathbb{E}\xi = 0$.

Proof. The Gaussian transition density and its derivative are

$$k_x(z) = \frac{\exp\{-\frac{1}{2}\|L_j^{-1}(z - \mu_j(x))\|^2\}}{(2\pi)^{d/2}|\det L_j|}, \qquad \nabla_x \log k_x(z) = D\mu_j(x)^\top L_j^{-\top}\xi_x(z).$$

Given the next state z, the law of subsequent transitions has no further dependence on x. Thus

$$h_j(x) = \int h_{j+1}(z)k_x(z)\,\mathrm{d}z, \qquad N_j(x) = \int N_{j+1}(z)k_x(z)\,\mathrm{d}z.$$

Condition N justifies differentiation inside these integrals. Applying the tower property to the remainder of the trajectory gives

$$\nabla_x h_j(x) = D\mu_j(x)^\top L_j^{-\top}\mathbb{E}_R[w\xi \mid x],$$
$$\nabla_x N_j(x) = D\mu_j(x)^\top L_j^{-\top}\mathbb{E}_R[wg\xi \mid x].$$

Since $h > 0$, differentiating $m = N/h$ yields $\nabla m = (\nabla N - m\nabla h)/h$. Substitution and $G = -\nabla m$ prove (10). For $f = 0$, $h = 1$ and $\mathbb{E}\xi = 0$, which gives Proposition 4. No derivative of the terminal evaluator or of the subsequent kernels is used. □

If the implementation uses isotropic noise $L_j = s_j I$, the coefficient is $1/s_j$, and the variance can increase for small noise. The formula is not applied to steps with zero noise; for such steps one uses the derivative of the deterministic map or the preceding nondegenerate step. If the noise covariance depends on x, a term for the derivative with respect to the covariance is required, and (10) is not used as it stands. For nonzero f, differentiating the unnormalized numerator instead of m gives, in general, a different quantity.

Centering and antithetic pairs For $f = 0$ and a fixed baseline b, $\mathbb{E}[(g - b)\xi] = \mathbb{E}[g\xi]$ holds, but if the mean $\bar{g}$ of the same independent sample is used as in $N^{-1}\sum(g_i - \bar{g})\xi_i$, the expectation is multiplied by $(N - 1)/N$. The implementation uses antithetic pairs $(+\xi, -\xi)$, so $\sum_i \xi_i = 0$ holds exactly within the sample and this shrinkage bias does not arise. If each member of an antithetic pair has the correct marginal law and distinct pairs are independent, the mean for $f = 0$ is unbiased. The variance estimate for independent samples, which ignores the correlation within a pair, is not used. The bias of self-normalized importance sampling for nonzero f is not removed by antithetic sampling.

B.2 Linear component of the Gaussian noise and the interval information

In continuous time the quadratic form of G gives the rate of increase of the information, but for a single finite step the square of G_j does not necessarily coincide with the interval information J_j. This is because the information also contains components beyond the linear response to the noise of that step. For $f = 0$ and a fixed state x, let

$$U(\xi) = m_{j+1}(\mu_j(x) + L_j\xi), \qquad a_x = \mathbb{E}[U(\xi)\xi].$$

Assume U is square-integrable under the standard Gaussian law. Orthogonal projection onto its linear component gives

$$\mathrm{Var}(U) = \|a_x\|^2 + \mathbb{E}\big[U - \mathbb{E}U - a_x^\top \xi\big]^2 \geq \|a_x\|^2.$$

To verify this identity, set $R_x = U - \mathbb{E}U - a_x^\top \xi$. The standard Gaussian moments imply $\mathbb{E}R_x = 0$ and $\mathbb{E}[R_x\xi] = 0$. Expanding $\mathbb{E}[(a_x^\top \xi + R_x)^2]$ eliminates the cross term and gives the expression above. The tower property gives $m_j(x) = \mathbb{E}U$, so the left side is precisely $\mathbb{E}[(M_{j+1} - M_j)^2 \mid X_j = x]$. Proposition 4 gives $G_j = -D\mu_j^\top L_j^{-\top} a_x$: it measures a transformed linear component, while the conditional interval information also includes the residual. If $D\mu_j$ is singular, even the full linear component need not be recoverable from G_j.

For example, let $X_{j+1} = x + \xi$ in one dimension and take $g(X_{j+1}) = X_{j+1}^2$ at this final step. At $x = 0$, $G_j(0) = -2x = 0$, whereas the conditional interval information is $\mathrm{Var}(\xi^2) = 2$. Thus squaring an exact estimate of G_j would already miss information at a finite step. For nonzero f the tilted noise is not standard normal, so the displayed projection identity cannot be used as it stands. In the finite-step temporal analysis, we instead estimate I_j directly by Proposition 5 and obtain J_j from differences.

B.3 Information from independent continuations and the shared matrix

Proposition 5 (Information from independent continuations)**.** *For $Z \sim p_j^f$, let $A, B \in L^2$ be conditionally independent given Z, with $\mathbb{E}[A \mid Z] = \mathbb{E}[B \mid Z] = m_j^f(Z)$. Then*

$$\mathrm{Cov}(A, B) = \mathrm{Var}(m_j^f(Z)) = I_j. \tag{11}$$

The sample covariance from independent copies of (Z, A, B) is unbiased for I_j.

For $N \geq 2$ independent copies (Z_i, A_i, B_i) satisfying Proposition 5, define

$$\widehat{I}_j = \frac{1}{N-1}\sum_{i=1}^{N}(A_i - \bar{A})(B_i - \bar{B}).$$

Proof. By the conditional independence and the conditional unbiasedness, $\mathbb{E}[AB \mid Z] = m(Z)^2$, and since $\mathbb{E}A = \mathbb{E}B = \mathbb{E}m$, (11) follows. For the sample formula, $\mathbb{E}[N^{-1}\sum_i A_iB_i] = \mathbb{E}AB$ and $\mathbb{E}[\bar{A}\bar{B}] = N^{-1}\mathbb{E}AB + (1 - N^{-1})\mathbb{E}A\,\mathbb{E}B$; multiplying the difference by $N/(N-1)$ gives the unbiasedness of $\widehat{I}_j$. □

The distinction from taking the variance of one set of estimates is explicit in

$$\mathrm{Var}(A) = \mathrm{Var}(m(Z)) + \mathbb{E}[\mathrm{Var}(A \mid Z)].$$

With two groups, conditional independence makes the covariance of their estimation errors zero. Each group may contain only one continuation or the mean of several independent continuations; the same identity applies as long as the conditional means are unbiased. Independence between the groups removes variance from estimating the conditional mean, not a bias in that mean.

A matrix shared by many directions For a fixed feature $\Phi \in L^2(P;\mathbb{R}^p)$ of the final output, let

$$\mathcal{C}_j = \mathrm{Cov}_{p_j^f}\big(\mathbb{E}_{P^f}[\Phi(X_{N_{\mathrm{gen}}}) \mid X_j]\big).$$

Write $\mathbf{M}_j = \mathbb{E}_{P^f}[\Phi(X_{N_{\mathrm{gen}}}) \mid \mathcal{F}_j]$. Since the conditional mean of $g_\theta = \theta^\top \Phi$ is $\theta^\top \mathbf{M}_j$, $I_j(g_\theta) = \theta^\top \mathcal{C}_j \theta$. The vector martingale has orthogonal increments, so

$$\mathcal{C}_{j+1} - \mathcal{C}_j = \mathbb{E}[(\mathbf{M}_{j+1} - \mathbf{M}_j)(\mathbf{M}_{j+1} - \mathbf{M}_j)^\top] \succeq 0.$$

Both $\mathcal{C}_j$ and each population increment are therefore positive semidefinite. This is the matrix version of Proposition 1(iii).

Let $A_i, B_i \in \mathbb{R}^p$ be two vectors of feature means, each computed from one of two conditionally independent sets of continuations from the i-th state, and form

$$\widehat{\mathcal{C}}_j = \frac{1}{2(N-1)} \sum_{i=1}^{N} \left\{ (A_i - \bar{A})(B_i - \bar{B})^\top + (B_i - \bar{B})(A_i - \bar{A})^\top \right\}.$$

Applying the scalar covariance proof to each pair of coordinates gives $\mathbb{E}\widehat{\mathcal{C}}_j = \mathcal{C}_j$ under Condition MC; symmetrization does not change this expectation. If the scalar estimate for g_θ is formed from the same samples, multiplying out the quadratic form gives $\widehat{I}_j(g_\theta) = \theta^\top \widehat{\mathcal{C}}_j \theta$ exactly. Thus trajectories can be reused algebraically for many update directions. In finite samples, $\widehat{\mathcal{C}}_j$ can be indefinite. To evaluate it as an unbiased estimator, our experiments do not project onto positive semidefinite matrices or truncate negative values.

Moreover, the information of $g = g_1 + g_2$ includes the covariance between the components:

$$I_j(g) = I_j(g_1) + I_j(g_2) + 2\mathrm{Cov}(m_{j,1}, m_{j,2}).$$

Therefore the information of the individual attributes cannot simply be summed and interpreted as the information of the whole update.

Estimation of intervals The difference $\widehat{I}_b - \widehat{I}_a$ is unbiased for $I_b - I_a$ if each estimate is unbiased. In computing the confidence intervals, the same states are used in the same resampling iteration, so that the correspondence between the estimates at the two times is preserved.

Two changes needed for nonzero f When generation is continued to the end with the reference process, the inner target is $m_j^f(x) = \mathbb{E}_R[e^{-f} g \mid x]/\mathbb{E}_R[e^{-f} \mid x]$, and if self-normalized importance sampling is used in each set, then for a finite number of samples $\mathbb{E}[\widehat{m}_A \mid x] = m(x) + b_A(x)$. Even if the two sets are independent, $\mathrm{Cov}(\widehat{m}_A, \widehat{m}_B) = \mathrm{Cov}_{p_j^f}(m + b_A, m + b_B)$, which in general is not I_j. Furthermore, the outer states must follow p_j^f; if the reference states r_j are used, the correction $\mathrm{d}p_j^f/\mathrm{d}r_j = h_j^f/Z_f$ is required. Neither problem arises in the main analysis with $f = 0$.

Reuse for learning and selection The population statements are correct for each fixed g, and the algebraic identity on the same arrays remains correct for a g chosen after the fact. However, when g is learned and selected from the same reference image set, the unbiasedness of the estimate for a fixed g cannot be applied directly to the g chosen afterwards. The accuracy in the presence of this dependence is assessed by comparisons with independent groups of samples and by experiments in settings where analytic solutions are available.

B.4 Nonnegative weights on generation intervals

Fix a process and a partition $0 = j_0 < \cdots < j_B = N_{\mathrm{gen}}$. For nonnegative weights w_b, define

$$I_w(g) = I_0(g) + \sum_{b=0}^{B-1} w_b \{ I_{j_{b+1}}(g) - I_{j_b}(g) \}.$$

Each interval term is $\mathbb{E}[(M_{j_{b+1}} - M_{j_b})^2]$. Consequently, I_w is a nonnegative quadratic form in the update direction. If $w_{\min} = \min\{1, w_0, \ldots, w_{B-1}\}$ and $w_{\max} = \max\{1, w_0, \ldots, w_{B-1}\}$, multiplying each nonnegative term in the temporal decomposition by these bounds gives

$$w_{\min} I_{N_{\mathrm{gen}}}(g) \le I_w(g) \le w_{\max} I_{N_{\mathrm{gen}}}(g).$$

All weights equal to one recover the terminal information exactly. For linear directions, the corresponding matrix is $\mathcal{C}_0 + \sum_b w_b(\mathcal{C}_{j_{b+1}} - \mathcal{C}_{j_b}) \succeq 0$. These are population identities. A finite-sample covariance estimate need not preserve nonnegativity. Unequal weights define a chosen measure of temporal variation, rather than the Fisher information of the original terminal family. Averaging over possible answers gives the time-weighted acquisition criterion in Appendix F.6. Increasing a weight favors updates with more variation in that interval, without asserting that the resulting labels reduce a preference loss more.

C Auxiliary Theory

C.1 Continuous-time derivative formula (Bismut–Elworthy–Li)

In continuous time, the Bismut–Elworthy–Li formula expresses the gradient as the expectation of the terminal function multiplied by a stochastic-integral weight, without differentiating the terminal function directly. Start the diffusion of Condition S at time t from the state x, and let $J_{t,s} = D_x X_s^{t,x}$ be the derivative flow (since the noise is additive, $\mathrm{d}J_{t,s} = D_x b(s, X_s) J_{t,s} \mathrm{d}s$ with $J_{t,t} = I$). For $r \in C_b^2$ and a deterministic $\alpha \in L^2([t,T])$ with $\int_t^T \alpha(s)\mathrm{d}s = 1$,

$$\nabla_x \mathbb{E}_{t,x} r(X_T) = \mathbb{E}_{t,x}\Big[r(X_T) \int_t^T \alpha(s) J_{t,s}^\top \sigma(s)^{-\top} \mathrm{d}W_s \Big],$$

(Elworthy & Li, 1994). The proof follows from $\nabla_x u(t,x) = \mathbb{E}[J_{t,s}^\top \nabla u(s, X_s)]$ for $u(s,z) = \mathbb{E}_{s,z} r(X_T)$, from the fact that the integrand of the martingale representation of $u(s, X_s)$ is $\sigma^\top \nabla u$, and from the bilinear form of Itô's isometry applied to the two stochastic integrals. For nonzero f, the formula is applied to both h and N and the derivative of the quotient is taken. Equation (9) is proved directly for a single finite step. It does not come with an evaluation of the discretization error of the continuous formula.

C.2 Gradient estimation using quadratic covariation

Guo et al. (2026) proposed methods that learn h from its martingale property and its gradient from quadratic covariation with the state process. Under Condition S, the conditional mean in this paper also satisfies, by Itô's formula,

$$\mathrm{d}\langle M, X\rangle_t = a(t) \nabla m_t(X_t) \mathrm{d}t = -a(t) G_t(X_t) \mathrm{d}t.$$

This relation motivates estimating the gradient by regressing targets derived from quadratic covariation on time and state. For numerical validation, we use the Gaussian derivative formula and the estimator based on independent continuations described in §4 (Appendices B.1, B.3).

C.3 Effect of fixing the initial distribution

Let $\widetilde{P}^f$ be the path law that uses the same transitions as P^f but keeps the initial distribution μ_R, and let $\widetilde{q}_f$ be its terminal distribution. Under absolute continuity and a finite initial KL divergence, the chain rule for the KL divergence and the data processing inequality give

$$D_{\mathrm{KL}}(P^f \| \widetilde{P}^f) = D_{\mathrm{KL}}(\mu_f \| \mu_R), \qquad D_{\mathrm{KL}}(q_f \| \widetilde{q}_f) \le D_{\mathrm{KL}}(\mu_f \| \mu_R).$$

For the infinitesimal update $f + \varepsilon g$, when the initial distribution is fixed to the reference μ_f, we have for bounded g

$$D_{\mathrm{KL}}(\mu_{f+\varepsilon g} \| \mu_f) = \tfrac{1}{2}\varepsilon^2 I_0^f(g) + o(\varepsilon^2).$$

This follows from the fact that the score of the initial state is $-(m_0(X_0) - \bar{g})$. The smallness of this local coefficient alone does not guarantee the error of generation with a finite λf, nor that of GUIDED, which also approximates the transitions.

D Implementation of the Learner and the Acquisition Function

This section describes the update of the potential used in the experiments. The theory in the main text is applied to the difference g of the functions obtained from this update.

D.1 Dual Averaging in a finite feature space

We use $f_\theta = \theta^\top \phi$ with fixed features $\phi(c, x) \in \mathbb{R}^p$. For a comparison $z = (c, a, b)$, the response probability and the estimate of the density ratio are

$$\pi_\theta(z) = \mathrm{sigmoid}\{-\gamma\theta^\top(\phi_a - \phi_b)\}, \qquad R_\theta(c, x) = \widehat{Z}_\theta(c) e^{\theta^\top \phi(c,x)},$$

where $\widehat{Z}_\theta(c)$ is the average of e^{-f_θ} over the reference image set for each prompt; with the population Z_θ, $R_\theta = r_T/q_{f_\theta}$.

The functional derivative of the DPO objective with respect to the distribution (Kawata et al., 2025) is mapped to the finite features, and the following vector is computed from the comparison set $\mathcal{D}$ ($n = |\mathcal{D}|$):

$$b(\theta, \mathcal{D}) = \frac{1}{n} \sum_{(c,a,b,y)\in\mathcal{D}} \gamma\,(\pi_\theta(c, a, b) - y)\{R_\theta(c, a)\phi(c, a) - R_\theta(c, b)\phi(c, b)\}.$$

Let $\widehat{A}$ be the Gram matrix of the features of the reference image set, ρP_{pen} the ridge regularization matrix, and $H = \widehat{A} + \rho P_{\mathrm{pen}} \succ 0$.

We write the internal update index of the learner as ℓ, distinguished from the number of acquired labels k. One internal update is

$$\begin{aligned} B_\ell &= B_{\ell-1} + \ell\, b(\theta_{\ell-1}, \mathcal{D}), \\ \Lambda_\ell &= \beta\ell(\ell + 1)/2 + \beta'(\ell + 1), \\ \theta_\ell &= H^{-1} B_\ell/\Lambda_\ell. \end{aligned}$$

At each update, all comparisons acquired so far are evaluated at the current θ. The last equation is the unique minimizer of the quadratic objective $\theta^\top H\theta - 2B_\ell^\top\theta/\Lambda_\ell$ on the finite feature space. This gives a finite-feature implementation of dual averaging with ridge regularization, using the functional derivative mapped to the feature space. In the experiments we set $\beta' = 4\beta$. Even when several internal updates are performed per label, the number of labels k increases by only one.

D.2 Computing the acquisition function from the feature covariance

Let $\delta\theta_{z,y} = \theta_{\mathrm{after}}(z, y) - \theta$ be the difference of the coefficients when the response y to the candidate z is added. Since the update direction for each response is $g_{z,y} = \delta\theta_{z,y}^\top\phi$, with the feature covariance under the current distribution

$$F_\theta = \sum_c \nu(c)\mathrm{Cov}_{q_{f_\theta}(\cdot|c)}\phi,$$

Eq. (7) becomes

$$A(z) = \sum_{y\in\{0,1\}} p_f(y \mid z)\, \delta\theta_{z,y}^\top F_\theta \delta\theta_{z,y}.$$

In the implementation, F_θ is estimated by the covariance of the features of reference images weighted by e^{-f_θ}. In settings with several internal updates, the coefficient difference after all the updates is used. The F_θ here is the feature covariance of the generative distribution and differs from the Fisher information of the logistic model of the comparison labels.

D.3 Information from comparison labels and acquisition baselines

For the Bernoulli response model above, let $d_z = \phi(c, a) - \phi(c, b)$. Its Fisher information with respect to θ for one comparison is

$$\begin{aligned} F_{\mathrm{label}}(z; \theta) &= \gamma^2\pi_\theta(z)(1 - \pi_\theta(z))d_z d_z^\top = v(z; \theta)v(z; \theta)^\top, \\ v(z; \theta) &= \gamma\sqrt{\pi_\theta(z)(1 - \pi_\theta(z))}\, d_z. \end{aligned}$$

We implement the online D-optimal criterion of Kveton et al. (2025) with the common DA learner. Let θ_{s-1} be the coefficients just before acquiring comparison z_s. The design matrix and the acquisition score are

$$V_k = \lambda_{\mathrm{des}} I_p + \sum_{s=1}^{k} v(z_s; \theta_{s-1})v(z_s; \theta_{s-1})^\top, \qquad A_{\mathrm{D}}(z) = v(z; \theta_k)^\top V_k^{-1} v(z; \theta_k).$$

Each historical vector is kept at its acquisition-time value, including the initial random comparisons. The matrix determinant lemma gives

$$\log\det\{V_k + v(z;\theta_k)v(z;\theta_k)^\top\} - \log\det V_k = \log(1 + A_{\mathrm{D}}(z)),$$

so maximizing A_{D} maximizes the increase in the log determinant. Uncertainty sampling uses only $\pi_\theta(z)(1-\pi_\theta(z))$. Both rules use the same candidate set construction as Fisher selection and do not access unacquired labels.

The two information matrices answer different questions. F_{label} describes how the probability of a comparison label changes with the preference parameters; accumulating it supports identifying those parameters. F_θ, in contrast, describes how the output distribution changes under a potential update. The latter is evaluated along the actual update $\delta\theta_{z,y}$ of the learner and averaged over answers in $A(z)$. Our temporal decomposition then relates this update's terminal information to its predictability along diffusion trajectories.

D.4 RESAMPLE AND GUIDED

Drawing in RESAMPLE and its expectation In systematic resampling, a single $U \sim \mathrm{Unif}[0, 1/n)$ is drawn, and the n candidate indices are selected from the intervals of the cumulative weights into which the grid points $U + \ell/n$ ($\ell = 0, \ldots, n-1$) fall. Since the length of the interval of candidate i is w_i, the conditional expectation of the number of selections is $\mathbb{E}[N_i \mid Y_{1:N}, w] = nw_i$. Hence the expected value of the output average of an evaluation function r equals $\sum_i w_i r(Y_i)$ (Douc et al., 2005). The outputs are dependent, and when the same candidate is selected several times, it is counted with its multiplicity in the evaluation.

If the candidates are independent and identically distributed according to the reference distribution and $e^{-\lambda f}$ and $e^{-\lambda f}|r|$ are integrable, then by the law of large numbers

$$\sum_{i=1}^{N} w_i r(Y_i) \;\longrightarrow\; \mathbb{E}_{q_{\lambda f}} r \qquad (N \to \infty),$$

almost surely. For finite candidates this is a self-normalized weighted average, which is different from finite-sample unbiasedness for the population expectation.

Approximation in GUIDED GUIDED uses 320 saved trajectories per prompt (64 of them held out for validation) and learns h_λ with $\exp\{-\lambda(f - a_c)\}$ as the regression target, where a_c is a numerical centering constant for each prompt that does not change the logarithmic gradient. During generation, the mean of the next state at each DDPM step is shifted by $\alpha\sigma_j^2 \nabla \log h_\lambda$, where σ_j^2 is the noise variance of that step; the experiments use $\alpha = 1$.

This correction constructs the shift of the Gaussian mean from the first-order approximation of $\log h$ in the exact discrete h-transform kernel

$$K_j^{\lambda f}(x, \mathrm{d}y) \propto K_j^R(x, \mathrm{d}y) h_{j+1}^{\lambda f}(y).$$

The initial distribution is fixed to the same Gaussian as in the reference process. Therefore, in addition to the learning error of h, the approximation of the transitions and the fixed initial distribution produce the deviation from the exact $P^{\lambda f}$.

E EXPERIMENTAL SETTINGS

E.1 VALIDATION IN THE LOW-DIMENSIONAL SETTING

The reference process is $\mathrm{d}X_t = -\kappa X_t \mathrm{d}t + \sigma \mathrm{d}W_t$ with $\kappa = \sigma = 1$, $X_0 \sim N(0, V_0 I)$, $V_0 = 0.5$, and $T = 1$. The potential is the quadratic function

$$f(y) = a^\top y + \tfrac{1}{2} y^\top Q y, \qquad a = (0.3, -0.2)^\top, \qquad Q = \begin{pmatrix} 0.4 & 0.15 \\ 0.15 & 0.2 \end{pmatrix},$$

and the update directions are the four functions $g \in \{x_1, x_1 x_2, x_1^2 - x_2^2, \sin(x_1 + x_2)\}$.

The process is discretized by the Euler method with $N_{\mathrm{gen}} = 200$ steps, and 128 states are evaluated at each of 21 times. The one-step Gaussian derivative formula uses 4000 antithetic noise pairs per state. ι_t is obtained by averaging over states the product $\widehat{G}_A^\top a \widehat{G}_B$ of two independent estimates. The number of repetitions is 4.

The analytic values are obtained by weighting the conditional Gaussian distribution of the OU process by e^{-f}. Under the reference process, $X_T \mid X_t = x \sim N(\rho_t x, V_t)$ with $\rho_t = e^{-(T-t)}$ and $V_t = \frac{1}{2}(1 - e^{-2(T-t)})I$; for $t < T$ the covariance and the mean after tilting are

$$\Sigma_t = (V_t^{-1} + Q)^{-1}, \qquad \mu_t(x) = \Sigma_t(V_t^{-1}\rho_t x - a).$$

From the moments of this Gaussian and the expectation of the sine function, m, G, and ι_t are computed for the four g. The polynomial f and g fall outside the boundedness required by Condition S, but here, since $Q \succ 0$, the state at each time follows a Gaussian distribution also after tilting, and its second moment is bounded for $t \in [0, T]$. Condition E holds for the four g, and the interchange of differentiation and expectation is justified by Gaussian integration. Moreover, the explicit formulas give $m \in C^{1,2}$ and $|\nabla m_t(x)| \le C(1 + |x|)$, so that $\mathbb{E}_{P^f} \int_0^T |\sigma^\top \nabla m_t(X_t)|^2 \mathrm{d}t < \infty$, and Itô's formula and the isometry apply.

E.2 Reference process and synthetic preferences in the CIFAR-10 experiments

Generator and comparison labels We use the public DDPM (Ho et al., 2020) checkpoint `google/ddpm-cifar10-32`[1] pretrained on CIFAR-10 (Krizhevsky, 2009). The weights of the diffusion model are fixed, and generation uses 100 steps of the stochastic DDPM transitions. The evaluation quantities of the synthetic preferences are the following four.

- COLOR: the difference between the means of the red and blue channels over the whole image.
- CLASS: the probability of the automobile class given by a classifier.
- LOCAL: the negative squared difference between the mean Sobel gradient magnitude and a target value, subsequently standardized.
- MIX: the combination of the standardized CLASS, COLOR, and LOCAL with weights $0.5/0.3/0.2$, standardized again.

The comparison labels are generated from the difference of the evaluation quantities of two images through a logistic function with temperature 0.5, which defines the response probability.

Preference learning The learner is the DA of Appendix D. For MIX, the feature dimension is 64, $\beta = 4$, $\gamma = 1$, the ridge coefficient is 0.03, and one internal update is performed per label. For COLOR, the feature dimension is 32, $\beta = 2$, $\gamma = 1$, the ridge coefficient is 0.1, and three internal updates are performed per label. At each stage a comparison is selected from 24 candidates, and training proceeds up to 120 labels. The reference image set has 8192 images, the image set for evaluating the Fisher information 4096 images, and the test image set 4096 images. Two independent reference image sets, N1 and N2, are used.

Measurement of the local response For each of the 48 states at $j \in \{60, 90\}$, we use three sets of generated samples. The samples used to determine the perturbation direction are called group A, those used to predict the response group B, and those used to measure the response to the perturbation group C; the three groups are generated with mutually independent random numbers. From the estimate of group A we determine the direction $v_A = \widehat{G}_A / \|\widehat{G}_A\|$, and with group B we predict the change per unit perturbation, $-v_A^\top \widehat{G}_B$. In group C we perturb the state to $x \pm \delta v_A$, generate to the end, and measure

$$\frac{\widehat{m}_C(x + \delta v_A) - \widehat{m}_C(x - \delta v_A)}{2\delta}.$$

δ is the length of the perturbation; its conversion to the per-pixel magnitude is given below.

[1] `https://huggingface.co/google/ddpm-cifar10-32`

From the independently constructed state sets C0 and C1, 24 states each are taken, and 48 states are evaluated at each of $j = 60, 90$. Groups A and B each use 256 antithetic noise pairs (512 generations). Group C uses 128 generations for each of the positive and negative perturbations, sharing the random numbers between the two signs.

For the reported results, the perturbation is specified by its root mean square per pixel, δ_{rms}, as

$$\Delta x = \delta_{\mathrm{rms}}\sqrt{D}\, v, \qquad D = 3072, \qquad \delta_{\mathrm{rms}} = 0.02,$$

where v is a unit direction and $\delta = \delta_{\mathrm{rms}}\sqrt{D}$. The random-direction control uses one direction per state and target. The confidence intervals of the regression slopes are computed by a bootstrap with 2000 resamples of the states.

The update directions for MIX are obtained from the training state after 15 labels with reference image set N1, training run 1, and selection of comparisons based on the Fisher information. For the comparison candidate with the largest selection criterion, the updates $g_y = f_{\mathrm{after}}(y) - f_{\mathrm{before}}$ obtained by adding one label with the response $y = 0, 1$ to the same training state are computed, and g_1 and $d = g_1 - g_0$ are used as the terminal quantities. Additive constants, which do not affect the sensitivity or the difference of responses, are removed. At evaluation, these two terminal quantities are fixed, and $f = 0$ and the f after 120 labels of the same training run are used. From each intermediate state, generation is continued with the reference DDPM, and $m_j^f[g]$ is estimated by the sample mean of g in the former case and by the self-normalized weighted mean with the weights e^{-f} in the latter case. The training point from which the update is extracted and the f that determines the distribution under which the response is evaluated are specified separately.

The two-axis analysis and its intervals Each response $y \in \{0, 1\}$ to a candidate z is fed to the learning routine to obtain the update $g_{k,r,z,y}$, where r denotes the training run and $\pi_{k,r,z}$ the predicted probability of response 1. We compare

$$\mathcal{I}_{k,j}^{s,a} = \mathrm{Avg}_{r,z}\Big[(1 - \pi_{k,r,z}) I_j^{\mathrm{ref}}(g_{k,r,z,0}) + \pi_{k,r,z} I_j^{\mathrm{ref}}(g_{k,r,z,1})\Big], \tag{12}$$

in which the information of the responses is averaged with the predicted probabilities and then averaged over runs and candidates. Here s denotes the method used to select comparisons during training, and a the way in which the candidates evaluated at that training state are chosen. Since the averaging weights do not depend on j, the temporal decomposition also holds for $\mathcal{I}_{k,j}$. When the total is positive, the ratio $\mathcal{I}_{k,j}/\mathcal{I}_{k,N_{\mathrm{gen}}}$ is the fraction of the total information that is predictable by time j; it is computed as the ratio of the averaged information.

The shared matrix uses the 512 states of C0+C1 with 32 generations per set and is computed at $j \in \{0, 10, 25, 40, 60, 75, 90, 98, 99, 100\}$. At each saved training state, the updates for the two responses are computed for each candidate. The candidates are the single one with the largest Fisher information or uniformly drawn ones. The numbers of update vectors, with the training runs combined, are 4–5 and 32–40, respectively.

In the bootstrap, the intermediate states are resampled, and the same matrix constructed in each iteration is applied to all saved updates. The intervals therefore preserve the correspondence between curves. These intervals represent the uncertainty of the matrix estimation with the saved training states and candidates fixed; they do not include the variability that would arise from repeating the preference learning.

E.3 Human preference data

Labels and data splits We use pairs of images for the same prompt from ImageRewardDB (Xu et al., 2023). For ALIGNMENT, the label prefers the image with the higher image–text alignment rating, and for OVERALL, the image with the higher overall rank. For ALIGNMENT, tied pairs are excluded. A response preferring image a is $y = 1$.

Following the official train / validation / test split of the dataset, the prompts for training, for configuration selection, and for evaluation are separated. TEST_OLD, TEST_NEW, and TEST_REPLICATION are not names of the official split; they are the names of three prompt-disjoint subsets that this study constructed within the official test set. The images and human ratings are those of the public data, and no new human labels were collected.

From each official split, the candidate prompts are those for which all images are available and uncorrupted and the number of images matches the metadata. When a normalized prompt text or the SHA-256 of an image is duplicated across official splits, it is kept in the order of priority test, validation, train, and prompts with fewer than 2 distinct images are removed. The candidates are sorted in ascending order of the SHA-256 computed from a fixed key (`extended_research_20260919`), the split name, and the prompt ID. Counting the indices in this order from 0, among the 465 valid prompts of the official test set, indices 0–127 form TEST_OLD, 128–383 TEST_NEW, and 384–464 TEST_REPLICATION. The human ratings are not used for the assignment to the splits.

Applying the same ordering rule to the official train split, indices 0–127, 256–383, and 384–511 form the training sets A, B, and C, respectively. The intermediate ranges 128–223 and 224–255 are reserved for feature design and its validation, and the first 64 prompts of the official validation split are used for configuration selection. Training set A contains 1983 pairs for OVERALL and 1422 pairs for ALIGNMENT.

The numbers of prompts and pairs available for evaluation are: for TEST_OLD, 128 prompts and 1800 pairs for OVERALL and 115 prompts and 1011 pairs for ALIGNMENT; for TEST_NEW, 256 prompts and 3532 pairs and 209 prompts and 1895 pairs, respectively; and for TEST_REPLICATION, 81 prompts and 1049 pairs and 66 prompts and 560 pairs. The numbers of prompts differ between the tasks because for some prompts no comparison pair remains after tied pairs are excluded. Training set A is evaluated on TEST_NEW and TEST_OLD, and sets B and C on TEST_NEW and TEST_REPLICATION. The 12 conditions in the main text are these six combinations of training and evaluation sets examined for the two tasks.

The final configuration was selected by the comparison loss on the validation set. TEST_NEW and TEST_REPLICATION were also used to evaluate earlier configurations. TEST_NEW is a name that denotes a set separated from the training prompts. It does not mean a set used for evaluation for the first time.

Reference generator and features We use the public checkpoint of Stable Diffusion v1.5[2]. The parameters of the diffusion model itself are fixed. Generation uses DDPM with 50 steps and a classifier-free guidance scale of 7.5. Samples that pass the official safety checker are used for evaluation. For the features we use the image-text matching (ITM) model of BLIP (base, COCO). We concatenate the representation of the classification token (CLS) of the image-conditioned text encoder (768 dimensions), the ITM logit difference, and the projected image representation (256 dimensions). Principal component analysis is fitted without labels on the reference image set and the images for feature design, and the features are compressed to 128 dimensions for ALIGNMENT and 256 dimensions for OVERALL.

Training and evaluation For ALIGNMENT, $\beta = 1.0$, the ridge coefficient is 0.1, and $\gamma = 2$; for OVERALL, $\beta = 2.0$, the ridge coefficient is 0.03, and $\gamma = 2$. There are 64 candidates at each stage and one internal update per label. The initial 16 labels are acquired with a random sequence common to all methods, and training proceeds up to 256 labels including these. The main experiments in the main text use the training seeds 40–42.

The CE is averaged over the comparison pairs within a prompt and then averaged with equal weights across prompts. The difference between methods is also computed for each prompt, averaged over the training seeds, and then the intervals are formed by a bootstrap with 2000 resamples of prompts. Since the 12 conditions share training runs and test sets, they are not interpreted as 12 independent replications.

D-optimal regularization and the comparison protocol For the public training pool $\mathcal{P}$, with feature dimension p, we set

$$s_{\mathrm{pool}} = \frac{\gamma^2}{4p|\mathcal{P}|} \sum_{z \in \mathcal{P}} \|d_z\|^2, \qquad \lambda_{\mathrm{des}} = 16\alpha s_{\mathrm{pool}}.$$

The scale uses features only, and λ_{des} remains fixed during a run. We select α per task from $\{0.001, 0.01, 0.1, 1, 10\}$ using validation CE, averaged over three training sets and seeds $35, 36$;

[2] `https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5`

both tasks select 10, the upper grid endpoint. All methods use the same learner settings specified above, previously selected using Fisher selection. The comparison therefore isolates the acquisition rule under a fixed learner and does not compare independently tuned learning systems. We use the selection criterion of Kveton et al. (2025) with this learner, rather than their original estimator and its guarantees.

For the four-method comparison, the primary metric is the mean CE on TEST_NEW over 32–256 labels, computed by trapezoidal integration over log label count. For each prompt, this is $(e_{32} + 2e_{64} + 2e_{128} + e_{256})/6$, where e_n is its CE at n labels. We average this quantity over the three training sets and three seeds before taking paired differences and resampling prompts. The validation criterion for α uses the same curve mean. The primary comparisons are Fisher versus D-optimal in the two tasks; 97.5% Bonferroni intervals are also reported. All intervals condition on the fixed training runs and do not include variability from new training seeds. The per-condition Fisher/random results use the same prompt bootstrap with training sets kept separate.

E.4 Selection with Stable Diffusion

Potentials and the selection strength We use potentials learned by Fisher, D-optimal, and random query selection up to 256 labels on ALIGNMENT, training set A. D-optimal uses the same learner, features, candidate-generation procedure, and initial 16 labels as the other methods, with $\alpha = 10$ from the AL validation. No additional images are generated for this comparison. The training seeds are 30 and 31, which are runs different from the main experiments of the human AL. λ is determined on the reference image set used for training (115 prompts, 128 images each). With the effective sample size (ESS) ratio defined as $(\sum_i w_i)^2/(N \sum_i w_i^2)$, we choose the largest $\lambda \leq 4$ that satisfies the following for each Fisher seed.

- The prompt-averaged ESS ratio is at least 0.5.
- At most 5% of the prompts have an ESS ratio below 0.2.
- The lower end of the confidence interval for the prompt-averaged ImageReward improvement is positive. For each prompt, this improvement is the difference between the self-normalized importance-weighted mean and the unweighted mean on the same reference image set.

The search grid is $\{1, 2, 4, 8\}$; since 1, 2, and 4 satisfied the conditions, $\lambda = 4$ was obtained.

Candidate generation and baselines The confirmation set consists of 32 prompts for design validation and 32 prompts for generation evaluation, 64 prompts in total. The design-validation prompts are divided into two groups of 16 according to whether existing trajectories are available, and the three strata DV-old16, DV-new16, and GE-32 are reported. For each prompt, 48 images are generated from new noise, and the 2982 images remaining after removing the 90 flagged by the safety checker are the valid candidates. All 64 prompts are retained, and the same valid candidates are used for every method and every training stage.

RESAMPLE selects 16 images for each prompt. REF is the average over the valid images among the first 16 candidates. GUIDED generates 16 images with the same initial noise as the first batch of REF. The comparison between RESAMPLE and REF, which differ in the number of candidates, evaluates the effect of selection, and the comparison between Fisher selection and random selection on the same candidate set evaluates the difference between the learned f.

Evaluation and curves by number of labels We use ImageReward-v1.0 and the image–text cosine similarity of Contrastive Language-Image Pre-training (CLIP) L/14 (Radford et al., 2021). The difference of the average score of each prompt is averaged over the two training seeds, and the 95% intervals are computed by a bootstrap with 2000 resamples of prompts. The comparison between GUIDED and REF also preserves the correspondence of the noise of the images.

For the curves by number of labels, the same candidates are evaluated with the f at $0, 16, 32, 64, 128, 256$ labels. The scores of the candidate images are shared, and the weights and the selection results are recomputed with each saved f. At 0 and 16 labels the f of the three methods are identical, and the difference of the weighted expectations is exactly 0. Since the 16 images actually selected differ because of the method-specific random numbers used for drawing, a small difference

remains in the output average. The intervals are pointwise intervals at each number of labels, not simultaneous confidence bands for the whole curve.

E.5 Computational cost

The processing that uses a GPU was measured on a single GPU. After the features and other inputs were prepared, human preference active learning took about 26 seconds per run on a CPU, and the measurement of the local response on images took 15,019 seconds in total for the four processes. Generating the 48 reference candidates $\times$ 8 prompts with Stable Diffusion took 520–915 seconds. For the selection by number of labels, the candidate images were reused; the feature extraction for the additional analysis took about 49 seconds on a GPU and the aggregation about 12 seconds on a CPU.

Once the shared matrix for the two-axis analysis is constructed, it can be reused for all candidates, and the evaluation of each candidate is a single quadratic form.

F Additional Results

F.1 Validation in the low-dimensional setting: four update directions

Figure 1(a) shows agreement between the analytic and estimated vectors G, while Figure 1(b) shows that the estimated information rate follows its analytic profile over time. The discrepancy between $I_0 + \int_0^T \iota_t \, \mathrm{d}t$ and I_T is at most 0.92% with analytic G and 5.8% with estimated G across the four directions (Table 1 and Figure 5, Appendix F.1). The analytic-G comparison checks the decomposition in (5) numerically; using estimated G additionally tests the Gaussian derivative calculation. Agreement at both levels connects the local response vector to the total information of the update.

Table 1 summarizes, for the quadratic f and the four g, the initial term, the integral of the rate of increase of the information, the terminal variance, and the relative error of the decomposition. Figure 5 shows the rate of increase of the information for the three g not shown in Figure 1(b).

F.2 Prediction of local responses in image generation

We test whether the estimated G predicts the change in the expected terminal quantity under a small perturbation of an intermediate state. We use the public denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) checkpoint `google/ddpm-cifar10-32`, pretrained on CIFAR-10, with fixed weights and 100 generation steps. Synthetic preferences evaluate the mean red–blue difference (COLOR), classifier probability of an automobile (CLASS), proximity of the mean Sobel gradient magnitude to a target (LOCAL), or their combination (MIX). Details and the checkpoint source are given in Appendix E.2.

As terminal quantities g, we use the COLOR score and updates from preference learning of MIX (learner in Appendix D). After 15 training labels, we consider one image pair: response $y = 1$ prefers the first image, and $y = 0$ the second. Adding either response as one label to the same training state changes the potential by $g_y = f_{\mathrm{after}}(y) - f_{\mathrm{before}}$. We use g_1 and $d = g_1 - g_0$ as fixed

Table 1: Validation of the low-dimensional temporal decomposition $I_T = I_0 + \int_0^T \iota_t \, \mathrm{d}t$ for the quadratic f. Relative errors, with respect to the terminal variance I_T (Monte Carlo; MC), of the value obtained by integrating the analytic G and of the values reconstructed from the estimated G.

				$(I_0 + \int_0^T \hat{\iota}_t \, dt)/I_T - 1$ [%]	
g	I_0	$\int_0^T \iota_t \, dt$	I_T (MC)	analytic G	Gaussian formula
$g = x_1$	0.0486	0.3698	0.4193	-0.22	-1.60
$g = x_1 x_2$	0.0044	0.1995	0.2058	-0.91	+5.77
$g = x_1^2 - x_2^2$	0.0161	0.7947	0.8059	+0.61	-1.14
$g = \sin(x_1 + x_2)$	0.0411	0.3612	0.4030	-0.18	+4.03

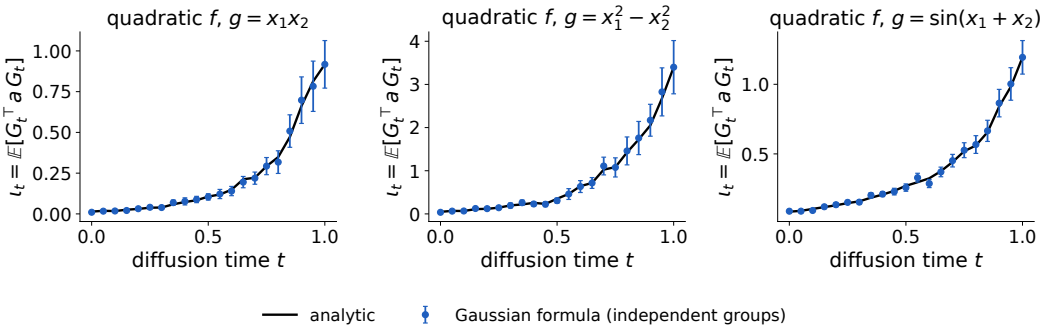


Figure 5: Analytic values and estimates of ι_t for the quadratic f and $g \in \{x_1 x_2,\ x_1^2 - x_2^2,\ \sin(x_1 + x_2)\}$, with the same conventions as in Figure 1(b).

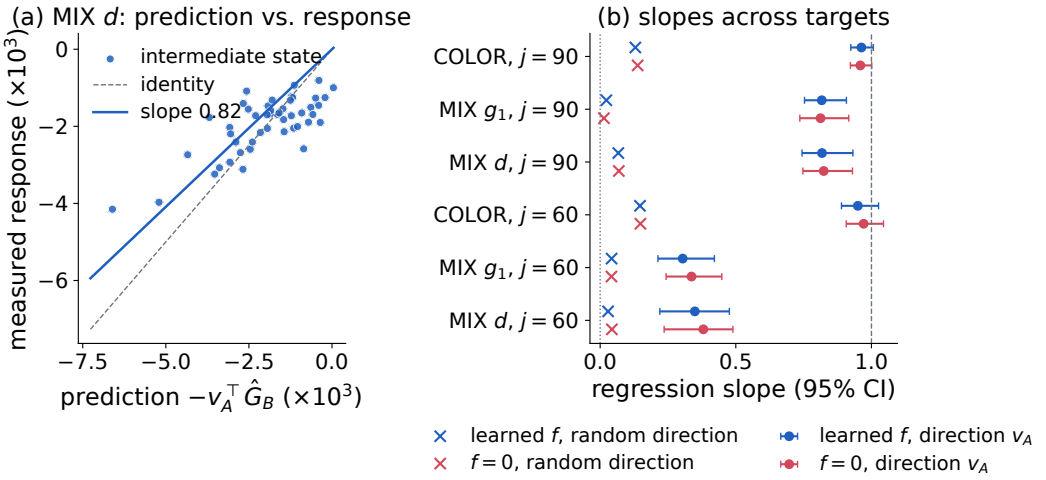


Figure 6: Local response prediction in CIFAR-10 generation ($\delta_{\text{rms}} = 0.02$). (a) Prediction versus measurement for MIX $d = g_1 - g_0$, $j = 90$, and learned f; g_1, g_0 are updates for answers favoring the first or second image. Points are states; prediction and measurement use independent samples. Dashed line: exact agreement. (b) Regression slopes; $\times$: random directions. Error bars: 95% confidence intervals from resampling states. The sample groups and training stages are detailed in Appendices E.2, F.2.

terminal quantities. Their responses are evaluated under both $f = 0$ and the potential f learned from 120 labels of the same run; the latter changes the evaluation weights, not the terminal quantities.

For 48 states at each of $j = 60, 90$, independent sample groups determine the perturbation direction $v = \widehat{G}/\|\widehat{G}\|$, predict the derivative $-v^\top G$, and measure it by central differences of the terminal conditional mean. The two perturbation signs share random numbers. Sample counts, group definitions, and the measurement formula are in Appendix E.2.

Figure 6(a) compares prediction and measurement for MIX at $j = 90$. Figure 6(b) gives regression slopes of 0.80–0.82 for MIX and 0.95–0.97 for COLOR at $j = 90$, but 0.30–0.38 for MIX at $j = 60$. Predicted and measured responses have the same sign in at least 96% of MIX cases at $j = 90$. Results under $f = 0$ and learned f are similar. Thus the predicted directional response is more accurate late in generation.

The independent prediction and measurement show that the estimated G captures how changing an intermediate state affects the expected terminal quantity, including an update obtained from preference learning. The dependence of the slopes on j also shows why this prediction must be assessed

at different generation times. This experiment tests the state derivative $-v^\top G_j$; the temporal information below is computed directly from continuations using §4.

F.3 Two-axis analysis over training stage and generation time: additional table and figure

For the individual attribute scores in Figure 2(a), the fractions predictable by $j = 60$ are 92.1% for COLOR, 58.7% for CLASS, and 41.5% for LOCAL.

Here we consider preference learning of MIX. Figure 7 shows, over all times, the cumulative information $\mathcal{I}_{k,j}/\mathcal{I}_{k,N_{\mathrm{gen}}}$ of the updates arising from the responses to the candidate (with the information of each response averaged). Table 2 shows the fraction predictable by $j = 40$, separated by the method of selecting comparisons during training (Fisher / random) and by the candidates evaluated at the same training state (the one with the largest Fisher information / uniformly drawn).

F.4 Human preference data: details of the comparison loss

Table 3 shows the CE for the 12 evaluation conditions and the paired differences between Fisher selection and random selection. The comparisons of 128 versus 256 labels and of 256 versus 256 labels are both listed. Here new, old, and rep. denote TEST_NEW, TEST_OLD, and TEST_REPLICATION, respectively (Appendix E.3).

Comparison with D-optimal and uncertainty selection Table 4 reports the differences underlying Figure 3(c). Fisher selection has slightly lower mean CE over the evaluated label budgets than D-optimal selection in both tasks, while their intervals include zero. The 97.5% Bonferroni intervals are $[-0.0084, 0.0083]$ for ALIGNMENT and $[-0.0048, 0.0018]$ for OVERALL. The sign of the difference varies across training sets and seeds. At 256 labels, a secondary comparison favors Fisher in OVERALL by -0.0048 $[-0.0087, -0.0008]$; this single-budget result is distinct from the primary curve comparison. D-optimal also improves over random selection, with differences in mean CE of -0.0182 $[-0.0289, -0.0079]$ and -0.0090 $[-0.0133, -0.0047]$, respectively.

Pairs selected by uncertainty sampling Table 5 summarizes a post-hoc analysis of acquired pairs after the initial 16 labels. We average over acquisitions 17–256 within each run and then equally over the nine runs per task. Uncertainty sampling selects pairs with smaller squared feature differences than random selection in all nine matched runs of each task. Its mean $|\pi - 1/2|$ is below 0.001 in both tasks. For the response model in Appendix D.3, ambiguity alone does not determine parameter information: $\operatorname{tr} F_{\mathrm{label}} = \gamma^2\pi(1-\pi)\|d_z\|^2$ in the fixed feature coordinates. In particular, $d_z = 0$ yields maximal ambiguity but no parameter information. The observed smaller feature differences suggest that uncertainty sampling may spend labels on weakly informative comparisons.

This interpretation concerns the chosen feature representation, not human-perceived image similarity. The runs have different learning histories, so the diagnostic does not establish the cause of the CE gap or measure the size of the actual DA updates. A post-hoc paired comparison gives uncertainty-minus-random differences in mean CE over 32–256 labels of 0.0118 $[0.0074, 0.0162]$ for ALIGN-

Table 2: Fraction $\mathcal{I}_{k,40}/\mathcal{I}_{k,N_{\mathrm{gen}}}$ predictable by $j = 40$ for MIX (the intervals are 95% confidence intervals obtained by resampling the intermediate states with the saved updates and candidates fixed). The selection method during training (Fisher / random) and the candidates evaluated at the same training state (the one with the largest Fisher information / uniformly drawn) are shown separately.

task	image set	training query selection	candidate	$k = 0$	$k = 15$	$k = 60$
MIX	N1	fisher	max. Fisher	0.030 [0.021, 0.041]	0.116 [0.088, 0.147]	0.130 [0.089, 0.178]
MIX	N1	fisher	uniform	0.088 [0.077, 0.102]	0.120 [0.095, 0.145]	0.117 [0.080, 0.162]
MIX	N1	random	max. Fisher	0.030 [0.021, 0.041]	0.069 [0.059, 0.079]	0.069 [0.059, 0.079]
MIX	N1	random	uniform	0.088 [0.077, 0.102]	0.091 [0.080, 0.102]	0.073 [0.062, 0.082]
MIX	N2	fisher	max. Fisher	0.103 [0.060, 0.160]	0.127 [0.095, 0.169]	0.130 [0.088, 0.200]
MIX	N2	fisher	uniform	0.077 [0.068, 0.086]	0.113 [0.086, 0.148]	0.125 [0.084, 0.193]
MIX	N2	random	max. Fisher	0.103 [0.060, 0.160]	0.075 [0.062, 0.092]	0.077 [0.068, 0.087]
MIX	N2	random	uniform	0.077 [0.068, 0.086]	0.077 [0.068, 0.085]	0.076 [0.068, 0.085]

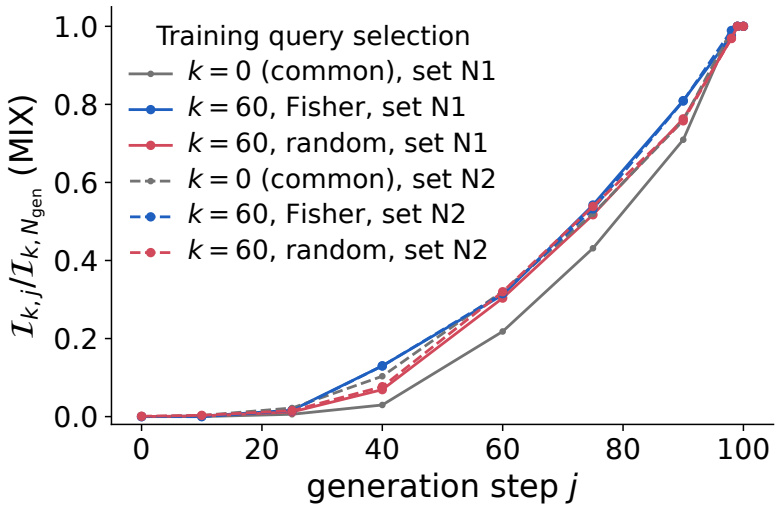


Figure 7: Cumulative information $\mathcal{I}_{k,j}/\mathcal{I}_{k,N_{\text{gen}}}$ of the updates arising from the responses to the candidate for MIX (the information of each response averaged, Eq. (12); the candidate is the one with the largest Fisher information; reference image sets N1/N2). At $k = 0$ the state is common to Fisher and random selection. At $k = 60$, the fraction predictable by $j = 40$ is larger for Fisher selection than for random selection in both reference image sets.

Table 3: Active learning with human preferences (12 conditions). The CE and the differences between methods within the same prompts are shown. F/R denote Fisher selection / random selection, and the subscript is the number of acquired labels. The intervals are 95% confidence intervals with prompts as the resampling unit.

			CE		CE difference [95% CI]	
task	train	test	F_{128}	R_{256}	$F_{128} - R_{256}$	$F_{256} - R_{256}$
ALIGNMENT	A	new	0.6341	0.6565	−0.0224 [-0.0349, -0.0100]	−0.0220 [-0.0322, -0.0109]
ALIGNMENT	A	old	0.6138	0.6444	−0.0306 [-0.0430, -0.0170]	−0.0292 [-0.0406, -0.0177]
ALIGNMENT	B	new	0.6506	0.6592	−0.0085 [-0.0247, +0.0087]	−0.0181 [-0.0305, -0.0053]
ALIGNMENT	B	rep.	0.6103	0.6464	−0.0361 [-0.0615, -0.0110]	−0.0379 [-0.0562, -0.0199]
ALIGNMENT	C	new	0.6444	0.6575	−0.0132 [-0.0356, +0.0096]	−0.0165 [-0.0313, -0.0018]
ALIGNMENT	C	rep.	0.5841	0.6382	−0.0541 [-0.0826, -0.0264]	−0.0431 [-0.0631, -0.0233]
OVERALL	A	new	0.6620	0.6773	−0.0153 [-0.0215, -0.0088]	−0.0152 [-0.0220, -0.0088]
OVERALL	A	old	0.6520	0.6727	−0.0208 [-0.0301, -0.0114]	−0.0197 [-0.0292, -0.0100]
OVERALL	B	new	0.6679	0.6749	−0.0070 [-0.0136, -0.0005]	−0.0094 [-0.0151, -0.0039]
OVERALL	B	rep.	0.6602	0.6721	−0.0119 [-0.0218, -0.0024]	−0.0127 [-0.0213, -0.0041]
OVERALL	C	new	0.6686	0.6756	−0.0071 [-0.0170, +0.0040]	−0.0106 [-0.0188, -0.0028]
OVERALL	C	rep.	0.6465	0.6700	−0.0235 [-0.0387, -0.0094]	−0.0217 [-0.0339, -0.0104]

Table 4: Differences in mean CE over 32–256 labels on TEST_NEW, averaged over three training sets and three seeds. Negative values favor Fisher selection. Intervals use 2000 paired prompt bootstrap resamples.

task	comparison	mean CE difference	95% CI
ALIGNMENT	Fisher − D-optimal	−0.0003	[-0.0076, +0.0071]
ALIGNMENT	Fisher − random	−0.0185	[-0.0351, -0.0014]
ALIGNMENT	Fisher − uncertainty	−0.0303	[-0.0507, -0.0088]
OVERALL	Fisher − D-optimal	−0.0017	[-0.0046, +0.0013]
OVERALL	Fisher − random	−0.0107	[-0.0180, -0.0033]
OVERALL	Fisher − uncertainty	−0.0159	[-0.0249, -0.0074]

MENT and 0.0052 $[0.0034, 0.0070]$ for OVERALL, using the same averaging and prompt bootstrap as above.

Table 5: Post-hoc statistics of pairs selected at acquisitions 17–256, averaged over three training sets and three seeds per task. π is the predicted response probability immediately before acquisition; feature distances use the learner's fixed representation.

task	selection rule	mean $\lvert\pi - 1/2\rvert$	mean $\lVert\phi_a - \phi_b\rVert^2$
ALIGNMENT	random	0.04283	0.551
ALIGNMENT	uncertainty	0.00085	0.276
OVERALL	random	0.03262	0.558
OVERALL	uncertainty	0.00055	0.313

F.5 Selection with Stable Diffusion: additional results

We apply the same selection rule to potentials learned on ALIGNMENT, set A, with seeds $30, 31$. D-optimal uses $\alpha = 10$ selected in the preceding AL validation without retuning for image selection. Table 6 gives paired comparisons at 256 labels and comparisons of 128 against 256 labels, and Table 7 gives the differences by prompt stratum. At 256 labels, D-optimal exceeds random by 0.099 $[0.060, 0.140]$, while Fisher minus D-optimal is 0.003 $[-0.026, 0.035]$. At 128 labels, D-optimal also exceeds random at 256 by 0.091 $[0.055, 0.127]$. Figure 8 shows gains over REF across label counts. All RESAMPLE intervals use the same prompt-bootstrap samples across comparisons. GUIDED results retain their original analysis and are listed separately in Table 6.

F.6 Selection of comparisons with weighted generation times

As an example of using the temporal decomposition in the acquisition function, we emphasize the information that becomes predictable in a particular range of generation times. For early, middle, and late weighting, the generation times are divided into three intervals by the boundaries $(j_0, j_1, j_2, j_3) = (0, 40, 75, 100)$; the information of interval b is $J_b^{\mathrm{ref}}(g) = I_{j_{b+1}}^{\mathrm{ref}}(g) - I_{j_b}^{\mathrm{ref}}(g)$, and w_b are nonnegative weights. For the MIX setting emphasizing $j = 25$–60, we instead use boundaries $(0, 25, 60, 100)$ and weights $(w_0, w_1, w_2) = (1, 4, 1)$. On the fixed reference process $f = 0$, we select the comparison that maximizes

$$A_w^{\mathrm{ref}}(z) = \sum_{y \in \{0,1\}} p_f(y \mid z) \left[I_0^{\mathrm{ref}}(g_{z,y}) + \sum_{b=0}^{2} w_b J_b^{\mathrm{ref}}(g_{z,y}) \right].$$

The coefficient of the initial term is 1. By Proposition 1, if all weights are 1, this reduces to the terminal Fisher information of the reference process averaged over the responses. Unlike (7), which evaluates information under the current q_f, A_w^{ref} uses the fixed reference process. We therefore evaluate the effect of the weights relative to A_1^{ref}, which uses the same reference process. The matrices are estimated with the two independent sets of samples of §4, and no computation of G is required.

Table 6: Paired image-selection comparisons over 64 prompts. F/D/R denote Fisher/D-optimal/random query selection. Potentials use 256 labels unless a subscript specifies a different budget. Intervals are 95% confidence intervals from prompt resampling.

comparison	ImageReward Δ	95% CI	seed 30 / 31	prompts improved	CLIP Δ
RESAMPLE: F − REF	+0.2105	[+0.1380, +0.2882]	+0.229 / +0.192	0.81	+0.0083
RESAMPLE: D − REF	+0.2073	[+0.1384, +0.2859]	+0.213 / +0.201	0.78	+0.0072
RESAMPLE: R − REF	+0.1079	[+0.0581, +0.1609]	+0.142 / +0.074	0.75	+0.0045
RESAMPLE: F − R	+0.1025	[+0.0599, +0.1486]	+0.087 / +0.118	0.67	+0.0038
RESAMPLE: D − R	+0.0994	[+0.0598, +0.1403]	+0.072 / +0.127	0.72	+0.0028
RESAMPLE: F − D	+0.0031	[-0.0261, +0.0346]	+0.016 / −0.010	0.55	+0.0010
RESAMPLE: $F_{128} - R_{256}$	+0.0936	[+0.0526, +0.1397]	+0.072 / +0.115	0.67	+0.0035
RESAMPLE: $D_{128} - R_{256}$	+0.0913	[+0.0549, +0.1268]	+0.076 / +0.107	0.73	+0.0015
GUIDED (F) − REF	−0.0003	[-0.0106, +0.0117]	−0.002 / +0.001	0.47	+0.0002
GUIDED (R) − REF	+0.0013	[-0.0053, +0.0084]	+0.001 / +0.002	0.56	+0.0003

Table 7: Paired RESAMPLE comparisons by prompt stratum at 256 labels. F/D/R denote Fisher/D-optimal/random query selection. Intervals are 95% confidence intervals from prompt resampling.

stratum	comparison	prompts	ImageReward Δ	95% CI
DV-old16	F − R	16	+0.1088	[+0.0010, +0.2295]
DV-old16	D − R	16	+0.1118	[+0.0201, +0.2061]
DV-old16	F − D	16	−0.0030	[-0.0687, +0.0684]
DV-new16	F − R	16	+0.1346	[+0.0632, +0.2131]
DV-new16	D − R	16	+0.0894	[+0.0184, +0.1701]
DV-new16	F − D	16	+0.0452	[-0.0085, +0.1022]
GE-32	F − R	32	+0.0833	[+0.0278, +0.1429]
GE-32	D − R	32	+0.0982	[+0.0395, +0.1609]
GE-32	F − D	32	−0.0149	[-0.0600, +0.0256]

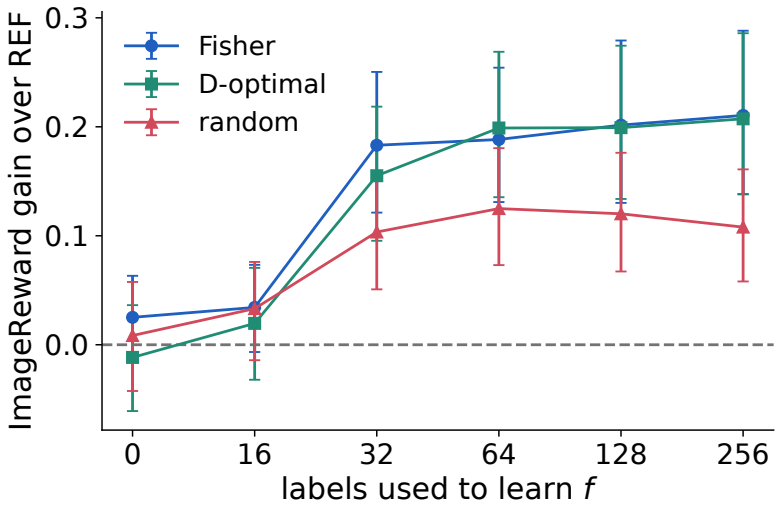


Figure 8: Difference in ImageReward between the 16 images selected with the f at each number of training labels and REF (the first 16 images), computed within the same prompts (95% intervals). From 32 labels on, all three query selection rules yield potentials that improve over REF.

Changing weights at a fixed learning state To isolate the effect of time weighting on query ranking, we evaluate the same 200 candidate pairs at each saved learning state under uniform weights and under a fourfold weight on each of the early, middle, or late intervals. We use N1/N2, the Fisher and random learning histories, and $k \in \{0, 15, 60\}$; the shared initial state is counted once per reference set, giving ten combinations of learning state and reference image set per task. Table 8 reports the mean absolute difference in standardized color scores among the ten highest-ranked pairs, averaged equally over these states. Emphasizing the early interval increases this mean in nine of ten states for both tasks. The highest-ranked pair changes in three of ten states for COLOR and four of ten for MIX. These are rankings of existing candidates, not ten newly acquired labels; the states are not independent training replications. This descriptive comparison uses the saved information estimates and does not quantify uncertainty in the estimated rankings.

Table 8: Mean color difference among the ten highest-ranked candidate pairs with the learning state and candidates held fixed. The three generation intervals are 0–40, 40–75, and 75–100. The initial information term always has weight 1. Values average these ten configurations per task.

task	uniform $(1, 1, 1)$	early $\times 4$ $(4, 1, 1)$	middle $\times 4$ $(1, 4, 1)$	late $\times 4$ $(1, 1, 4)$
COLOR	1.598	1.833	1.582	1.501
MIX	1.272	1.450	1.293	1.218

Selected pairs and learning performance For the task-specific weighting settings above, we also examine comparisons acquired during learning. Over the two reference image sets and three runs each, we averaged the absolute difference of the standardized color score of the image pairs selected as the 17th to 60th labels. For COLOR, quadrupling the weight of the early generation interval $j = 0$–40 increased it from 1.730, with all weights equal to 1, to 1.855, and the difference was positive in all six runs. For MIX, quadrupling only the interval $j = 25$–60 increased it from 1.484 to 1.681, and the difference was positive in five of the six runs. These are consistent with the analysis in §5.2 that color information appears at relatively early generation times.

In the confirmatory active learning, we compared A_w^{ref} with A_1^{ref} using the same learner and the same candidate-generation procedure. The difference in the test comparison loss averaged over the range of 16–60 labels was within ± 0.003 for each setting when all comparison pairs were considered. For the color-based preference COLOR with the weight of the middle interval quadrupled, the difference was -0.00167 with a 95% confidence interval of $[-0.00327, -0.00033]$ with runs as the unit. For the other weights considered, the interval of the difference in the same metric included 0. Therefore, although the temporal distribution can be used to change the selection of comparisons, the improvement in learning performance itself obtained in this setting is small.