COVERAGE-CONSTRAINED NEURAL EXPONENTIAL EMBEDDING: EXACT TEST-TIME PROJECTION FOR FROZEN GENERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time control of a generator is often implemented by prompting, scalar guidance, or reweighting a candidate bank. These mechanisms can improve a preferred attribute while silently concentrating probability on very few outputs. We instead ask for the closest distribution to a frozen generator that matches target statistics to a declared tolerance and retains a minimum effective sample size. The resulting finite-bank problem is a convex, coverage-constrained information projection. We prove existence and uniqueness, derive its low-dimensional dual, and show that an active coverage constraint changes the density ratio from an exponential tilt to a Lambert-W deformation. Thus classical empirical tilting is recovered only when coverage is inactive; rescaling a fitted tilt along a ray is generally not the constrained optimum. A reference-only bootstrap calibrates moment tolerance without evaluation leakage. The fitted density ratio can be used directly for bank resampling or as the terminal potential of learned twisted sequential Monte Carlo. Across frozen image and language generators, the exact projection improves held-out distribution matching over ray tempering, remains competitive with density ratio and stable-balancing baselines, and produces substantially better diffusion targets at matched fit-bank coverage.

1 INTRODUCTION

Modern generative models can produce individually high-quality samples while still producing the *wrong distribution of samples*. For example, a frozen image generator may overproduce bright images, fail to match the mixture of styles represented in a reference collection, or underrepresent a semantic mode. A language model may generate plausible answers on each call yet fail to produce a requested proportion of answer types over many calls. In such settings, the goal is not to find a single output with a high reward. It is to change the **distribution of outputs** produced by an otherwise fixed generator.

This distributional perspective leads to a different test-time control problem from the one addressed by most existing mechanisms. Best-of-N, reward guidance, and related approaches favor samples with large scalar scores, but doing so can concentrate probability on a small set of outputs. Prompting can be highly effective when the desired change can be expressed in words, but many targets are specified only through examples or aggregate statistics, and an instruction need not reproduce a desired population-level proportion (Meister et al., 2025). Fine-tuning can learn a new controller, but it changes model parameters and may alter behavior outside the target task. Classical test-time adaptation likewise modifies predictive parameters or internal statistics in response to test inputs (Wang et al., 2021). Here we consider a different setting: **the generator remains frozen, and only the probability law over its outputs is adapted**.

A simple way to see the problem is to begin with a finite bank of samples $x_{1:n}$ drawn from the frozen generator. Each sample is represented by a statistic $\phi(x)$, which may encode quantities such as brightness, semantic attributes, or neural embedding features. The desired deployment distribution is specified through a target moment $\mu^\star$, or through a reference set from which that moment is estimated. We then seek weights $w_1, \dots, w_n$ that make the weighted candidate bank reproduce the

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.

target statistics. This formulation requires neither access to the generator's parameters nor evaluation of its likelihood; it operates directly on generated samples and their features.

The classical solution to this kind of moment-matching problem is **exponential tilting**. Among all distributions satisfying the desired moment constraint, exponential tilting chooses the one with minimum relative entropy from the original distribution. On a finite candidate bank, this is an information projection, or I-projection, and the resulting weights have an exponential form. The difficulty is that accurate moment matching may require placing almost all probability on a few candidates. The nominal constraint may therefore be satisfied while the reweighted distribution has effectively collapsed.

We make this loss of coverage an explicit part of the optimization problem. We measure concentration using the effective sample size (ESS),

$$\mathrm{ESS}(w) = \frac{1}{\sum_i w_i^2},$$

so that diffuse weights have high ESS and highly concentrated weights have low ESS. Given a required coverage fraction ρ and an allowable moment error ϵ, we solve

$$\min_{w \in \Delta_n} \mathrm{KL}(w\|u) \quad \text{s.t.} \quad \left\| \sum_i w_i \phi_i - \widehat{\mu}^\star \right\|_2 \le \epsilon, \qquad \mathrm{ESS}(w) \ge \rho n, \tag{1}$$

where u is the uniform empirical distribution on the candidate bank. In words, Eq. (1) asks for the **least disruptive reweighting of the frozen generator that matches the target closely enough while remaining sufficiently spread across the available samples**. Because the ESS condition is equivalent to the convex quadratic constraint $\sum_i w_i^2 \le 1/(\rho n)$, this is a convex optimization problem; whenever it is feasible, the optimal distribution is unique.

A widely used practical alternative is to first fit the unconstrained exponential tilt and then weaken it until the ESS requirement is met. We call this procedure **ray tempering**: the natural parameter of the fitted exponential family is rescaled along a single one-dimensional path. Ray tempering is simple and inexpensive, but it solves a restricted problem. It asks how far one may move along a particular exponential path before violating the coverage requirement. Equation (1), by contrast, searches over *all* admissible reweightings and therefore finds the global information projection under the same coverage constraint.

This distinction has an unexpected structural consequence. Without an active coverage constraint, the solution is the familiar exponential density ratio. When the ESS constraint is active, however, the Karush–Kuhn–Tucker equations instead give

$$\log r_i = \theta^\top \phi_i + b - 2\eta r_i, \tag{2}$$

where $r_i = n w_i$ is the density ratio relative to the empirical generator distribution and $\eta \ge 0$ is the multiplier associated with coverage. Solving this equation yields a **Lambert-W deformation of the exponential tilt**. Thus the quadratic coverage constraint does more than change the strength of an exponential controller: it changes the functional form of the optimal density ratio itself. Ordinary empirical tilting is recovered exactly when the coverage constraint is inactive. This explains why post-hoc rescaling of an unconstrained tilt is generally not the constrained optimum.

The finite-bank projection also separates two questions that are often conflated in test-time generation: **what distribution should the generator realize, and how should that distribution be sampled?** Equation (1) answers the first question. Once the optimal density ratio has been obtained, it can be used directly to resample a candidate bank. For generators that produce outputs through a sequence of latent or autoregressive states, the same fitted ratio can instead serve as a terminal potential for twisted sequential Monte Carlo (SMC). Twisted SMC and related sequential controllers provide mechanisms for propagating a terminal preference through the generation process (Wu et al., 2023; Lew et al., 2023); they do not by themselves determine which terminal law is the least disruptive law satisfying a given moment and coverage specification. Our projection provides that law, after which SMC can realize it without modifying the generator parameters.

A further practical issue is that target moments are often estimated from a small reference set. Enforcing the empirical reference mean exactly can therefore overfit sampling noise. We address this

by calibrating the allowable moment error using a bootstrap performed solely on the target references. The resulting tolerance automatically adapts to the scale and covariance structure of the chosen statistics and uses no evaluation data.

Our contributions are as follows:

1. **Coverage-constrained distributional adaptation.** We formulate test-time adaptation of a frozen generator as a coverage-constrained information projection. The formulation makes the tradeoff between target matching and distributional coverage explicit, and we characterize the exact feasibility frontier and uniqueness of the optimal law.
2. **An exact Lambert-W solution.** We derive the optimal density-ratio form and a dual problem involving only $d + 2$ variables. Classical empirical exponential tilting is recovered when the ESS constraint is inactive; when coverage is active, the optimal link becomes Lambert-W. This also makes precise why ray tempering is generally only an approximation and permits its information-projection gap to be quantified.
3. **Reference-calibrated and sequential realization.** We introduce a scale-aware bootstrap tolerance using target references only. The resulting fitted density ratio is a reusable terminal potential that can be evaluated on a new candidate bank or propagated through a frozen sequential generator using learned twisted SMC.
4. **Evaluation across image and language generators.** We evaluate the method on frozen SDXL, Qwen, and CIFAR diffusion generators. Controlled image and text experiments test recovery of known shifted laws; independent SDXL prompt shifts test transfer to naturally generated reference distributions; and a matched-ESS diffusion study isolates the effect of choosing the exact Lambert terminal law from the effect of the downstream SMC realization.

2 Relation to Existing Methods

Exponential embedding and moment-based reweighting. Exponentially embedded families (EEF) interpolate between probability models through their log densities (Kay, 2005). A related statistical literature uses moment restrictions to reweight observed samples, including empirical likelihood (Owen, 1988), empirical exponential tilting (Kitamura & Stutzer, 1997), and entropy balancing (Hainmueller, 2012). Exponential tilting and entropy balancing provide the closest connection to our objective: they seek weights that match specified feature averages while minimizing relative entropy from a base distribution. Our formulation builds on this information-projection principle but jointly imposes approximate moment matching and an explicit ESS floor. The resulting KKT characterization recovers an exponential density ratio when the coverage multiplier is zero and a Lambert-W link when it is positive. Thus, rescaling a previously fitted exponential tilt is generally not equivalent to solving the coverage-constrained projection..

Density-ratio estimation and stable balancing. Kernel mean matching, KLIEP, and uLSIF estimate importance weights or density ratios to align source and target distributions (Huang et al., 2007; Sugiyama et al., 2008; Kanamori et al., 2009). These methods share our use of target samples to determine how source samples should be reweighted. Stable balancing instead emphasizes low-dispersion weights subject to approximate agreement in selected statistics (Zubizarreta, 2015), within the broader family of minimum-dispersion approximately balancing estimators (Wang & Zubizarreta, 2020). In our stable-balancing comparator, squared weight magnitude is the objective; in Eq. (1), KL divergence is the objective and squared weight magnitude is bounded through the ESS constraint. We compare stable balancing at the same moment tolerance and temper the discriminative density-ratio baseline to the same prescribed ESS floor. Our dual additionally supplies a parameterized density-ratio potential whose fitted-bank weights solve Eq. (1), allowing the fitted function to be evaluated on new candidates. The exact projection guarantee concerns the bank on which the optimization is performed.

Guidance and sequential control. Diffusion guidance modifies reverse-generation trajectories using conditioning information or other preferences (Dhariwal & Nichol, 2021; Ho & Salimans, 2022; Sani et al., 2026). Text-generation controllers similarly use future discriminators, expert ratios, or particle methods (Yang & Klein, 2021; Liu et al., 2021; Lew et al., 2023). Distribution-level guidance is particularly relevant: Sani et al. (2026) steer a frozen diffusion model toward a reference distribution using MMD gradients during sampling. This shares our reference-based adaptation setting, but

approaches distribution matching through changes to the sampling dynamics. Our focus is instead the finite-bank target-selection problem: choosing the minimum-KL law satisfying specified moment and coverage constraints. Once this law is fitted, its density ratio can be used for bank resampling or as a terminal potential, a weighting function of the final output, for twisted SMC (Wu et al., 2023). The two components are complementary: the projection specifies the target potential, and sequential control provides a mechanism for realizing it without updating the generator.

Test-time adaptation. Classical test-time training and adaptation modify predictive parameters or normalization statistics in response to deployment inputs (Sun et al., 2020; Wang et al., 2021; Zhao et al., 2023). Our setting adapts a different object. After target statistics or references become available, we estimate a small density-ratio potential while leaving the generator parameters unchanged. Adaptation therefore changes the relative probabilities assigned to outputs rather than the parameters of the model that produces them.

3 Coverage-Constrained Neural EEF

Our method separates *choosing a distribution* from *sampling from it*. We first reweight a finite bank of outputs from a frozen generator, seeking the smallest relative-entropy change that meets a target-matching tolerance and an effective-sample-size (ESS) requirement. We then use the fitted density ratio either to resample candidates or to guide sequential generation.

3.1 Target Specification on a Finite Bank

A running example. For intuition, consider adjusting the average brightness of images produced by a frozen generator. The method takes three main inputs. The *candidate bank* $x_{1:n}$ contains images sampled from the generator under its original condition; these are the outputs whose sampling probabilities we will change. The *reference set* $y_{1:m}$ contains images representing the desired brightness level. The *statistic* ϕ measures the property to be controlled: in this illustrative example, $\phi(x)$ is the mean brightness of image x. Applying the same statistic to both collections lets us compare the weighted average brightness of the candidates with the average brightness of the references.

The output is a set of sampling weights $w_1, \dots, w_n$ over the candidate images. These weights should bring the selected statistic close to its reference value while remaining as close as possible to uniform weighting and satisfying a minimum ESS requirement. Thus, the finite-bank operation changes *how often an image is selected*, not its pixels or the generator's parameters. Resampling uses the candidate bank; the references provide the target statistics. For more complex properties, we replace scalar brightness with a vector of neural or other features, leaving the roles of the inputs unchanged.

General setup. Let $q_0(x \mid c)$ denote a frozen generator under a fixed condition c, and let $x_{1:n} \sim q_0(\cdot \mid c)$ be a candidate bank. A statistic $\phi : \mathcal{X} \to \mathbb{R}^d$, held fixed during reweighting, represents the properties to be controlled. It may contain scalar attributes, whitened neural coordinates, quadratic features, or random Fourier features. Write $\phi_i = \phi(x_i)$.

The target is supplied either directly as a moment $\mu_\star$ or through references $y_{1:m}$, from which we compute

$$\widehat{\mu}_\star = \frac{1}{m} \sum_{j=1}^{m} \phi(y_j).$$

For a directly specified moment, we set $\widehat{\mu}_\star = \mu_\star$. The same feature map is applied to candidates and references, but no pairing between individual candidates and references is required. No likelihood for the generator or reference distribution is required.

We assign nonnegative weights w_i summing to one. Relative to the uniform empirical base law $u_i = 1/n$, define density ratios $r_i = n w_i$ and the empirical average $\mathbb{E}_u[f_i] = n^{-1} \sum_i f_i$. Then Eq. (1) becomes

$$\begin{aligned} \min_{r \in \mathbb{R}^n_+} \quad & \mathbb{E}_u[r \log r] \\ \text{s.t.} \quad & \mathbb{E}_u[r] = 1, \quad \left\| \mathbb{E}_u[r\phi] - \widehat{\mu}_\star \right\|_2 \le \epsilon, \quad \mathbb{E}_u[r^2] \le \rho^{-1}, \end{aligned} \tag{3}$$

where $0 < \rho \le 1$, $\epsilon \ge 0$, and $0 \log 0 = 0$. The objective is $\mathrm{KL}(w \| u)$. The moment constraint controls agreement with the target statistics, while the last constraint limits weight concentration,

since

$$\mathrm{ESS}(w) = \frac{1}{\sum_i w_i^2} = \frac{n}{\mathbb{E}_u[r^2]}.$$

For example, uniform weighting over k candidates has ESS k. The same constraint is a Pearson χ^2 ball:

$$\chi^2(w\|u) = \mathbb{E}_u[(r-1)^2] = \mathbb{E}_u[r^2] - 1 \le \rho^{-1} - 1.$$

Here, *coverage* refers to the dispersion of weights on the bank, not a guarantee of semantic diversity or coverage of missing modes. Matching selected moments also need not identify the entire target law.

Feasibility before fitting. A target may be unattainable at the requested coverage, regardless of the optimizer. We therefore first compute the smallest achievable moment error,

$$d_\rho = \min_{\substack{r\ge 0,\ \mathbb{E}_u[r]=1 \\ \mathbb{E}_u[r^2]\le \rho^{-1}}} \left\|\mathbb{E}_u[r\phi] - \widehat{\mu}_\star\right\|_2. \tag{4}$$

This separates an incompatible target–coverage specification from a numerical optimization failure: Eq. (3) is feasible exactly when $\epsilon \ge d_\rho$.

3.2 The Exact Projection and Lambert-W Form

The constraints in Eq. (3) define a convex set, and the relative-entropy objective is strictly convex. The resulting optimum is therefore a unique reweighted distribution, rather than a choice among equally good weighting rules.

Theorem 1 (Coverage-constrained information projection)**.** *For $\rho \in (0,1]$ and $\epsilon \ge d_\rho$, the feasible set of Eq.* (3) *is nonempty and compact, and the problem has a unique minimizer $r^\star$. Suppose additionally that a strictly feasible positive ratio exists: there is an $r > 0$ with $\mathbb{E}_u[r] = 1$ and both inequality constraints strict. Then $r^\star > 0$, and there exist $\theta \in \mathbb{R}^d$, $b \in \mathbb{R}$, and $\eta \ge 0$ such that*

$$\log r_i^\star = \theta^\top \phi_i + b - 2\eta r_i^\star, \qquad \eta\big(\mathbb{E}_u[(r^\star)^2] - \rho^{-1}\big) = 0. \tag{5}$$

Consequently,

$$r_i^\star = \begin{cases} \exp\{\theta^\top \phi_i + b\}, & \eta = 0, \\ \dfrac{W\big(2\eta \exp\{\theta^\top \phi_i + b\}\big)}{2\eta}, & \eta > 0, \end{cases} \tag{6}$$

where W is the principal Lambert function, defined by $W(z)e^{W(z)} = z$ for $z \ge 0$.

The compactness, strict-convexity, and KKT arguments are given in Appendix A.1. Uniqueness concerns the fitted weights; redundant statistics can leave the parameters nonunique. Strict feasibility is required here for the finite-multiplier characterization, not for the existence and uniqueness statement.

An inactive ESS constraint forces $\eta = 0$, recovering ordinary empirical exponential tilting. When $\eta > 0$, the ESS constraint binds and the optimal link is instead Lambert-W. The extra term $-2\eta r_i^\star$ in Eq. (5) penalizes large density ratios within the optimality equations themselves. Thus coverage can change the *form* of the controller, not merely the strength of an already fitted exponential score.

A low-dimensional dual. Although the primal has n weights, its dual has only $d + 2$ variables. With moment multiplier $\lambda \in \mathbb{R}^d$, normalization multiplier $a \in \mathbb{R}$, and coverage multiplier $\eta \ge 0$, the dual is

$$\begin{aligned} \max_{\substack{\lambda\in\mathbb{R}^d,\ a\in\mathbb{R} \\ \eta\ge 0}} \quad & \frac{1}{n}\sum_{i=1}^n \inf_{r\ge 0} \left[r\log r + (\lambda^\top \phi_i + a)r + \eta r^2\right] \\ & - \lambda^\top \widehat{\mu}_\star - \epsilon\|\lambda\|_2 - a - \frac{\eta}{\rho}. \end{aligned} \tag{7}$$

The scalar inner minimizer is given by Eq. (6) with $\theta = -\lambda$ and $b = -1 - a$. Under the strict-feasibility condition above, strong duality holds. We optimize the negative dual using L-BFGS-B, constraining only $\eta \ge 0$, and evaluate the Lambert function in the log domain to avoid overflow. Primal feasibility, projected-gradient norms, and stationarity residuals provide numerical diagnostics; details appear in Appendix A.2.

Algorithm 1 Coverage-constrained Neural EEF: reference-based fitting

Require: Candidate bank $x_{1:n}$; references $y_{1:m}$; fixed statistic ϕ; coverage $\rho \in (0, 1]$; quantile $\tau \in (0, 1)$; bootstrap count B; slack $\delta > 0$.
Ensure: Weights w, potential parameters (θ, b, η), tolerance ϵ, frontier d_ρ, and numerical diagnostics.
Compute $\phi_i \leftarrow \phi(x_i)$ and $z_j \leftarrow \phi(y_j)$.
Set $\widehat{\mu}_\star \leftarrow m^{-1} \sum_{j=1}^{m} z_j$.
Solve Eq. (4) for d_ρ.
for $t = 1, \ldots, B$ **do**
    Resample m vectors from $z_{1:m}$ with replacement and compute their mean $\widehat{\mu}_\star^{(t)}$.
end for
Compute ϵ_{ref} and $\epsilon \leftarrow \max\{d_\rho + \delta, \epsilon_{\mathrm{ref}}\}$ using Eq. (8).
if $\rho = 1$ **then**
    Set $w_i \leftarrow 1/n$, $\theta \leftarrow \mathbf{0}$, $b \leftarrow 0$, $\eta \leftarrow 0$.
else
    Solve the dual in Eq. (7) for (λ, a, η), with only η constrained nonnegative.
    Set $\theta \leftarrow -\lambda$ and $b \leftarrow -1 - a$.
    Recover r_i from Eq. (6) and set $w_i \leftarrow r_i / \sum_j r_j$.
end if
Compute ESS, achieved moment error, KL divergence, and primal/KKT diagnostics.
return $w, (\theta, b, \eta), \epsilon, d_\rho$, and diagnostics.

3.3 Reference-Calibrated Tolerance and Fitting Algorithm

When only a small reference set is available, its empirical mean is a noisy estimate of the target moment. Enforcing exact agreement can therefore fit reference-sampling noise. We resample the m reference features with replacement to obtain bootstrap means $\widehat{\mu}_\star^{(1)}, \ldots, \widehat{\mu}_\star^{(B)}$, and set

$$\begin{aligned} \epsilon_{\mathrm{ref}} &= Q_\tau\left(\left\{\left\|\widehat{\mu}_\star^{(t)} - \widehat{\mu}_\star\right\|_2\right\}_{t=1}^{B}\right), \\ \epsilon &= \max\{d_\rho + \delta, \epsilon_{\mathrm{ref}}\}, \end{aligned} \tag{8}$$

where Q_τ is the empirical τ-quantile and $\delta > 0$ is a small numerical slack. The two terms play different roles: ϵ_{ref} reflects variability in the reference mean, while $d_\rho + \delta$ ensures compatibility with the candidate bank and coverage requirement. Thus the *bootstrap* uses references only; the final tolerance also accounts for fitted-bank feasibility. Neither term uses evaluation data.

The bootstrap component inherits the units and empirical variability of the chosen statistic. We fix $\tau = 0.5$ before the natural-shift evaluation, targeting average held-out discrepancy rather than claiming a high-confidence uncertainty region. Higher quantiles give larger tolerances and permit more conservative adaptation (Appendix A.5). When the target moment is specified directly rather than estimated from references, the user supplies ϵ, and Eq. (4) checks feasibility.

Algorithm 1 summarizes reference-based fitting. It returns both bank weights and the parameters of the corresponding density-ratio function. The case $\rho = 1$ is handled directly, since only uniform weights satisfy that coverage requirement.

3.4 Ray Approximation and Sequential Realization

Ray tempering as a restricted approximation. An inexpensive alternative first fits an exponential tilt $\widehat{\theta}$ without the ESS constraint and then uses

$$w_i^{\mathrm{ray}}(\alpha) = \frac{\exp\{\alpha \widehat{\theta}^\top \phi_i\}}{\sum_j \exp\{\alpha \widehat{\theta}^\top \phi_j\}}, \qquad \alpha \in [0, 1].$$

ESS is nonincreasing along this ray, so bisection finds the largest α meeting the coverage floor. This search is exact *on the ray*, but the ray need not contain the solution of Eq. (3). Whenever the ray weights also satisfy the same moment tolerance, global optimality gives

$$\mathrm{KL}(w^\star \| u) \le \mathrm{KL}(w^{\mathrm{ray}} \| u).$$

The comparison requires both methods to satisfy the same constraints; an ESS match alone is insufficient. The argument and ESS monotonicity proof are given in Appendix A.3.

Bank resampling and transfer. Sampling an index with probabilities w_i directly realizes the fitted finite-bank law. For new candidates, define the fixed potential $R(x) = r(\phi(x); \theta, b, \eta)$ using Eq. (6). On a new bank $\widetilde{x}_{1:N}$, use

$$\widetilde{w}_j = \frac{R(\widetilde{x}_j)}{\sum_{\ell=1}^{N} R(\widetilde{x}_\ell)}.$$

All fitted parameters remain fixed during evaluation. In particular, b must remain *inside* the Lambert transform: unlike the exponential case, changing this intercept is not equivalent to multiplying all ratios by a common constant. Transfer evaluates the fitted function; it does not solve a new projection. Consequently, the original bank's optimality, moment tolerance, and ESS guarantees do not automatically extend to the new bank.

Learned twisted SMC. For diffusion or autoregressive generation, the same density ratio can serve as a preference over completed outputs. Let $U(x) = \log R(x)$. At an intermediate state $S_k = s$, the exact continuation value is

$$h_k(s) = \mathbb{E}_{q_0}\left[e^{U(X)} \mid S_k = s\right], \qquad q_k^{\text{twist}}(s' \mid s) = q_{0,k}(s' \mid s)\frac{h_{k+1}(s')}{h_k(s)}. \tag{9}$$

Here $q_{0,k}$ is the frozen generator's transition at step k. The continuation value measures how much terminal weight is expected from a partial generation. We learn a positive continuation model from frozen trajectories with a Poisson objective and use incremental continuation ratios in SMC; training and resampling details remain in Appendix B.2. This trains an auxiliary model, not the generator, and leaves the fitted terminal potential unchanged. Learned continuation values and finite particle populations introduce approximation into this realization; they do not extend the exact finite-bank projection guarantee to the sequential sampler.

4 Experiments

We test three claims: (1) exact coverage projection improves held-out target-law recovery over ray tempering; (2) the result transfers from constructed image and text laws to independent target-prompt images; and (3) the Lambert terminal law remains advantageous when realized by twisted SMC. We report only prespecified distributional metrics and keep all target/evaluation splits disjoint.

Metrics and inference. Moment error is the Euclidean distance between weighted candidate and held-out target statistics. Squared maximum mean discrepancy (MMD^2) compares kernel mean embeddings (Gretton et al., 2012); lower is better. Image construction uses CLIP (Radford et al., 2021) and independent evaluation uses DINOv2 (Oquab et al., 2024). Controlled common-support studies additionally retain exact target weights. We use paired bootstrap intervals; natural SDXL intervals resample the six prompt families as independent clusters. Full protocols and all absolute means are in Appendix C.

4.1 Controlled image and text laws

Six SDXL families and six Qwen story families are split into design and test banks. On each design bank we construct linear, asymmetric-mixture, and symmetric-mixture target laws, draw 64 target references, and fit kernel statistics without access to test weights. Table 1 compares the exact bootstrap projection with ray EEF at a 60% ESS floor.

Table 1: Held-out MMD^2 under controlled laws. Values are ray $\rightarrow$ exact; parentheses give relative reduction and paired 95% interval. Each cell aggregates 18 family-repeat pairs.

Target law	SDXL image bank	Qwen text bank
Linear	$.00671 \rightarrow .00368$	$.02063 \rightarrow .01233$
	45.2% $[35.5, 53.6]$	40.2% $[31.1, 48.5]$
Asymmetric mixture	$.00810 \rightarrow .00685$	$.01006 \rightarrow .00862$
	15.4% $[8.6, 24.4]$	14.4% $[8.8, 20.6]$
Symmetric mixture	$.00360 \rightarrow .00348$	$.00371 \rightarrow .00385$
	3.4% $[-12.8, 18.1]$	-3.8% $[-9.4, 1.3]$

The linear and asymmetric gains replicate across modalities, with intervals excluding zero. The symmetric result is deliberately retained: when opposing modes cancel in the selected first-order representation, a moment uncertainty ball cannot identify the change. Quadratic or kernel statistics that witness the relevant higher-order structure are required.

4.2 Independent SDXL Prompt Shifts

We next evaluate transfer to reference distributions generated from six frozen SDXL 1.0 prompt pairs (Podell et al., 2024). The base prompts describe a car, cabin, bird, fruit still life, city street, and dog; the corresponding target prompts modify attributes such as color, illumination, weather, environment, or composition. For each of five splits per prompt family, we use 128 base images to fit the controller, 128 unseen base images to evaluate transfer, 64 target-prompt images as references, and 64 additional target-prompt images for evaluation. The fitted density-ratio function is evaluated on the unseen base candidates without refitting. Further generation and feature-construction details are provided in Appendix C.2.

Table 2: Natural SDXL prompt shifts at a 40% required design-bank ESS. Values are means over 30 family–split pairs: six prompt families and five splits per family. Test ESS is reported as a fraction of the test-bank size. Lower is better for neural moment error and DINO MMD^2; bold values identify the lowest point estimates among the reweighting methods. Direct target prompting is an oracle-like reference rather than a reweighting baseline.

Method	Test ESS	Neural moment error	DINO MMD^2
Base	1.000	0.4617	0.1361
Ray EEF	0.647	0.3199	0.1193
Logistic density ratio	0.476	0.3186	**0.1145**
Stable balancing	0.464	0.2780	0.1152
Exact bootstrap	0.443	**0.2770**	0.1147
Direct target prompt	1.000	0.1080	0.0167

Table 2 compares the proposed exact bootstrap projection with the uniform base, ray tempering, discriminative density-ratio fitting, and stable balancing. At a 40% required fitting-bank ESS, exact bootstrap reduces held-out neural moment error from 0.3199 to 0.2770 relative to ray EEF, a reported reduction of 13.4%. DINO MMD^2 decreases from 0.1193 to 0.1147, a reported reduction of 3.8%. The comparisons with the other reweighting baselines are more mixed. Logistic density-ratio fitting achieves a slightly lower DINO MMD^2 point estimate than exact bootstrap (0.1145 versus 0.1147), but has higher neural moment error (0.3186 versus 0.2770). Stable balancing, fitted at the same moment tolerance, obtains similar neural moment error and DINO MMD^2 point estimates (0.2780 and 0.1152). These results support competitive held-out matching rather than uniform superiority across all metrics and baselines.

4.3 Matched-ESS diffusion and twisted SMC

We freeze `google/ddpm-cifar10-32` and fit ray EEF and the Lambert projection at 20% ESS on the same 5,000-candidate bank for a high-brightness CIFAR-10 target. Both use the same scalar

statistic; their continuation models share cached trajectories, architecture, training objective, and particle count. Evaluation uses five paired runs with common random seeds (Appendix C.3).

Table 3: Matched-fit-ESS diffusion comparison. Relative reductions compare the Lambert projection with ray EEF; all five paired splits favor Lambert.

Realization	Metric	Ray	Lambert	Reduction
Terminal bank resampling	Moment error	1.380	.853	38.2%
	MMD^2	.0832	.0377	54.7%
Learned twisted SMC	Moment error	1.397	.746	46.6%
	MMD^2	.0811	.0333	58.9%

Table 3 shows that Lambert reduces MMD^2 relative to ray EEF by 54.7% under terminal resampling and 58.9% under learned twisted SMC, with corresponding improvements in moment error. Learned twisting increases Lambert's unique-output fraction from 23.3% to 73.4%. Its MMD^2 also decreases from 0.0377 to 0.0333. Ray-twisted SMC has the same reported uniqueness (73.4%) but substantially higher MMD^2 (0.0811), so Lambert's matching advantage is not accompanied by greater output duplication. The learned Lambert predictor achieves a held-out log-weight correlation of 0.63. Thus, the projection improves target matching, while the clearest additional benefit of learned realization in this study is reduced duplication. In addition, Figure 1 summarizes the comparisons with ray EEF, showing improvements across image and text generation tasks.

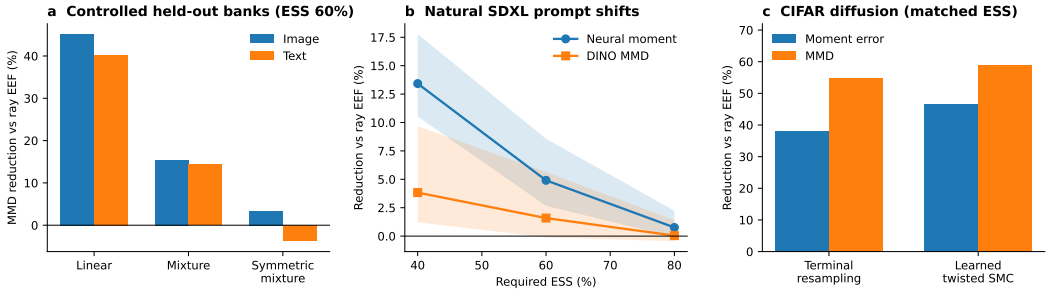

Figure 1: Summary of the comparisons with ray EEF. (a) Exact bootstrap projection improves held-out controlled laws across image and text, except a deliberately unidentified symmetric shift. (b) Natural SDXL gains are largest when the coverage constraint permits meaningful adaptation. Shading is a family-clustered 95% interval. (c) At matched fit-bank ESS, the Lambert target substantially improves both terminal resampling and learned twisted SMC.

5 Conclusion

Coverage-constrained Neural EEF turns ESS from a post-hoc diagnostic into an explicit constraint on test-time distribution control. The resulting finite-bank information projection is unique whenever feasible and has a tractable low-dimensional dual. Under strict feasibility, a positive coverage multiplier yields a Lambert-W density ratio, explaining why rescaling an exponential tilt is generally not optimal. A reference bootstrap calibrates the moment tolerance, while the fitted potential supports bank resampling and learned twisted SMC without updating the generator. Image and language experiments demonstrate the value of jointly controlling target fit and weight concentration, with gains retained under sequential realization in the diffusion study.

AI USE STATEMENT

Generative AI tools polished the paper writing and assisted the coding for algorithm implementation and debugging. The authors inspected the code, verified experimental outputs, and drafted the whole paper.

ETHICS STATEMENT

Distribution control inherits the bias, licensing, factuality, and misuse risks of the frozen generator. Moment matching does not certify fairness or safety. Targets should be justified and audited, and synthetic media should retain provenance. The reported experiments use public pretrained assets, aggregate targets, and synthetic outputs; no private data or human subjects are involved.

REPRODUCIBILITY STATEMENT

All experiments ran on one RTX 4090. Appendix E records banks, splits, optimization tolerances, statistics, seeds, compute, and artifact paths. The implementation includes unit tests for ESS equivalence, feasibility, Lambert reconstruction, bootstrap scaling, and fixed-potential transfer.

REFERENCES

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Sch"olkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012. doi: 10.1093/pan/mpr025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jiayuan Huang, Arthur Gretton, Karsten M. Borgwardt, Bernhard Schölkopf, and Alex J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, volume 19, pp. 601–608, 2007.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

Steven M. Kay. Exponentially embedded families—new approaches to model order estimation. *IEEE Transactions on Aerospace and Electronic Systems*, 41(1):333–345, 2005. doi: 10.1109/TAES.2005.1413765.

Yuichi Kitamura and Michael Stutzer. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4):861–874, 1997. doi: 10.2307/2171942.

Alexander K. Lew, Tan Zhi-Xuan, Gabriel Grand, and Vikash K. Mansinghka. Sequential monte carlo steering of large language models using probabilistic programs. *arXiv preprint arXiv:2306.03081*, 2023.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 6691–6706, 2021. doi: 10.18653/v1/2021.acl-long.522.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 24–49, 2025. doi: 10.18653/v1/2025.naacl-long.2.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Art B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988. doi: 10.1093/biomet/75.2.237.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Matina Mahdizadeh Sani, Nima Jamali, Mohammad Jalali, and Farzan Farnia. Training-free distribution adaptation for diffusion models via maximum mean discrepancy guidance. *arXiv preprint arXiv:2601.08379*, 2026.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von B"unau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Yixin Wang and José R. Zubizarreta. Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations. *Biometrika*, 107(1):93–105, 2020. doi: 10.1093/biomet/asz050. URL https://doi.org/10.1093/biomet/asz050.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535, 2021. doi: 10.18653/v1/2021.naacl-main.276.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42058–42080, 2023.

José R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015. doi: 10.1080/01621459.2015.1023805. URL https://doi.org/10.1080/01621459.2015.1023805.

A THEORY

A.1 PROOF OF THEOREM 1

Let

$$\mathcal{C}_\rho = \left\{ r \in \mathbb{R}_+^n : \frac{1}{n}\sum_i r_i = 1,\ \frac{1}{n}\sum_i r_i^2 \le \frac{1}{\rho} \right\}.$$

The set is closed and bounded and is therefore compact. The continuous moment map sends it to a compact set. By the definition of d_ρ, the intersection with the closed moment ball is nonempty exactly when $\epsilon \ge d_\rho$. The objective $n^{-1}\sum_i r_i \log r_i$ is strictly convex on normalized nonnegative ratios. Hence its minimizer, and therefore its induced law, is unique.

For the positive-component case, introduce multipliers $\lambda \in \mathbb{R}^d$ for the moment relation, $a \in \mathbb{R}$ for normalization, and $\eta \ge 0$ for the quadratic constraint. Minimizing over a moment vector inside $B(\widehat{\mu}_\star, \epsilon)$ contributes $-\lambda^\top \widehat{\mu}_\star - \epsilon\|\lambda\|_2$. Pointwise stationarity in r_i gives

$$\log r_i + 1 + \lambda^\top \phi_i + a + 2\eta r_i = 0.$$

With $\theta = -\lambda$ and $b = -1 - a$, this is Eq. (2). If $\eta > 0$, multiplying by $2\eta e^{2\eta r_i}$ and applying the principal Lambert function gives Eq. (6). If $\eta = 0$, stationarity gives the exponential ratio. Complementary slackness yields $\eta(\mathbb{E}_u r^2 - 1/\rho) = 0$. Slater's condition holds whenever a strictly feasible positive ratio exists; boundary cases follow by continuity.

A.2 DUAL OBJECTIVE AND DERIVATIVES

Define

$$\psi_i(r; \lambda, a, \eta) = r \log r + (\lambda^\top \phi_i + a) r + \eta r^2.$$

For fixed dual variables its minimizing ratio is

$$r_i(\lambda, a, \eta) = \begin{cases} \exp\{-1 - \lambda^\top \phi_i - a\}, & \eta = 0, \\ \dfrac{W(2\eta \exp\{-1 - \lambda^\top \phi_i - a\})}{2\eta}, & \eta > 0. \end{cases}$$

Substitution gives Eq. (7); the inner infimum equals $-r_i - \eta r_i^2$. Envelope derivatives are the moment, normalization, and quadratic residuals. We optimize the negative concave dual with L-BFGS-B, constraining only $\eta \ge 0$. A log-domain Lambert evaluation avoids overflow.

The implementation reports primal feasibility, dual projected-gradient norm, and the root-mean-square stationarity residual

$$\left[\frac{1}{n} \sum_i (\log r_i - \theta^\top \phi_i - b + 2\eta r_i)^2 \right]^{1/2}.$$

A.3 RAY TEMPERING IS A RESTRICTED PROBLEM

Let $\widehat{\theta}$ be an unconstrained empirical-tilt solution and define $\mathcal{R} = \{w(\alpha) : \alpha \in [0, 1]\}$, where $w_i(\alpha) \propto \exp(\alpha \widehat{\theta}^\top \phi_i)$. Ray tempering selects a feasible point from $\mathcal{R}$. Since $\mathcal{R} \cap \mathcal{F}_{\rho,\epsilon} \subseteq \mathcal{F}_{\rho,\epsilon}$, the exact projection satisfies

$$\mathrm{KL}(w^{\mathrm{exact}} \| u) \le \mathrm{KL}(w^{\mathrm{ray}} \| u)$$

whenever the two are compared at the same moment tolerance. This is a direct optimality statement over all finite-bank laws, not only a monotonicity result along $\mathcal{R}$.

For completeness, with scores s_i and $Z(\alpha) = \sum_i e^{\alpha s_i}$,

$$\log \mathrm{ESS}(\alpha) = 2 \log Z(\alpha) - \log Z(2\alpha), \qquad \frac{d}{d\alpha} \log \mathrm{ESS}(\alpha) = 2\mathbb{E}_{w_\alpha} s - 2\mathbb{E}_{w_{2\alpha}} s \le 0.$$

Thus the ray baseline is solved exactly by bisection even though the ray itself need not contain the global constrained optimum.

A.4 Analytic scalar feasibility frontier

For a scalar statistic z_i, maximizing $\mathbb{E}_u[rz]$ over $\mathcal{C}_\rho$ has KKT form

$$r_i = c(z_i - t)_+.$$

The threshold t is the unique root, outside degenerate ties, of

$$\frac{\mathbb{E}_u[(z-t)_+^2]}{\mathbb{E}_u[(z-t)_+]^2} = \frac{1}{\rho},$$

and $c = 1/\mathbb{E}_u[(z-t)_+]$. The minimum follows by replacing z with $-z$. If the target lies between the uniform and extreme moments, a convex mixture of the uniform and extreme ratios matches it exactly while respecting the quadratic ball. This reduces the 5,000-candidate CIFAR feasibility stage to a one-dimensional root search.

A.5 Bootstrap tolerance

The bootstrap in Eq. (8) resamples only target-reference features. It is exactly equivariant to a global feature rescaling: replacing ϕ by $c\phi$ multiplies every bootstrap displacement and the selected radius by $|c|$. This property is absent from a fixed chi-square radius applied after arbitrary neural feature maps.

The median radius targets average held-out discrepancy rather than a 90% confidence set. In pilot sensitivity tests, the 90% radius was stable but too conservative for mixture targets. We fixed the median before the natural SDXL evaluation. A deployment requiring conservative uncertainty coverage may set a higher quantile, accepting weaker adaptation.

B Algorithms

B.1 Finite-Bank Implementation Details

Algorithm 1 gives the complete reference-based fitting procedure. Here we specify numerical checks and the quantities needed to reuse the fitted controller. The dual derivation and stationarity diagnostic are given in Appendix A.2.

Normalization and primal feasibility. Let $\widetilde{r}_i$ denote the ratios reconstructed from the numerical dual solution using Eq. (6), before the final normalization in Algorithm 1. The normalization error is

$$e_{\mathrm{norm}} = \left| \frac{1}{n} \sum_{i=1}^{n} \widetilde{r}_i - 1 \right|.$$

For the returned weights $w_i = \widetilde{r}_i / \sum_j \widetilde{r}_j$, the moment and coverage violations can be measured as

$$\begin{aligned} v_{\mathrm{mom}} &= \left[\left\| \sum_{i=1}^{n} w_i \phi_i - \widehat{\mu}_\star \right\|_2 - \epsilon \right]_+, \\ v_{\mathrm{cov}} &= \left[n \sum_{i=1}^{n} w_i^2 - \rho^{-1} \right]_+, \end{aligned} \tag{10}$$

where $[a]_+ = \max\{a, 0\}$. Both violations vanish for a feasible returned weight vector. Evaluating them after normalization checks the probabilities actually used for resampling, whereas e_{norm} checks the normalization equation of the dual solution. Final normalization alone does not ensure moment or coverage feasibility.

Optimization diagnostics. Primal feasibility should be checked alongside the dual projected-gradient norm and the root-mean-square stationarity residual defined in Appendix A.2. The latter checks Eq. (5) using the reconstructed ratios $\widetilde{r}_i$ and the corresponding fitted parameters (θ, b, η), before the final resampling normalization. A small stationarity residual alone does not establish that the moment and coverage constraints have been met; these are separate checks on the numerical solution.

Table 4: Controlled results at 60% ESS. Means over 18 paired cases.

Domain	Target	Method	Moment error	MMD^2	TV
Image	Linear	Ray EEF	.1184	.00671	.2186
		Exact bootstrap	.0887	.00368	.1797
Image	Asymmetric	Ray EEF	.1167	.00810	.2124
		Exact bootstrap	.1040	.00685	.1968
Image	Symmetric	Ray EEF	.0738	.00360	.1559
		Exact bootstrap	.0732	.00348	.1572
Text	Linear	Ray EEF	.1317	.02063	.2628
		Exact bootstrap	.1015	.01233	.2251
Text	Asymmetric	Ray EEF	.1063	.01006	.2273
		Exact bootstrap	.0978	.00862	.2144
Text	Symmetric	Ray EEF	.0609	.00371	.1784
		Exact bootstrap	.0622	.00385	.1813

Retaining a reusable controller. A reusable controller is specified by the fixed feature map ϕ and the fitted parameters (θ, b, η). Any fitted preprocessing used to define ϕ must therefore remain fixed when evaluating new candidates. Retaining only the weight vector w is sufficient for resampling the fitting bank, but not for evaluating the potential on new outputs. For transfer, use the fixed-potential convention in Section 3.4: evaluate the fitted ratio first and normalize the resulting values afterward. In particular, normalization must not overwrite or shift b inside the Lambert transform. The same fixed terminal potential is supplied to the learned twisted-SMC procedure in Appendix B.2.

B.2 Learned twisted SMC

For frozen base trajectories $(S_0, \dots, S_K, X)$, compute the terminal density ratio $R(X) = e^{U(X)}$ and normalize it over training trajectories. A positive continuation network $\widehat{h}_\omega(S_k, k)$ is trained by Poisson regression,

$$\min_{\omega} \mathbb{E}\left[\widehat{h}_\omega(S_k, k) - R(X) \log \widehat{h}_\omega(S_k, k)\right].$$

At selected reverse-diffusion steps, incremental potentials are ratios of successive continuation estimates. Particles are systematically resampled when ESS falls below half the population and at the terminal step. In the reported CIFAR experiment, stochastic DDIM transitions allow duplicated particles to branch after resampling. Generator parameters are never updated.

C Experimental Details and Additional Results

C.1 Controlled image and text banks

Image features come from six frozen SDXL prompt families. Each family contains 256 base images; 128 form the design bank and 128 form the test bank. Text features come from 3,072 one-paragraph Qwen2.5-1.5B-Instruct generations across six semantic contexts (Qwen Team, 2024); each repeat samples 512 items per context and splits them evenly. In both modalities, an eight-dimensional PCA is fitted on design features and mapped to 32 random Fourier statistics.

For each regime, a known oracle law is constructed on the design and test banks. Its strength is calibrated to 40% design ESS. We draw 64 references with replacement from the design oracle and discard the oracle weights during fitting. Evaluation compares transferred weights with exact test-bank oracle weights. Reported MMD uses an RBF kernel with median-distance bandwidth in the independent evaluator representation. The main table contains 18 paired cases per regime and modality: six families times three predetermined seeds.

We additionally compare the exact frontier with stable balancing, defined as minimum $\mathbb{E}_u r^2$ subject to the same moment ball and normalization. Stable balancing is strong on the design bank but its affine nonnegative extrapolation is not an exact density-ratio family on unseen banks. The KL projection is the appropriate comparator when the scientific claim is least information change relative to the generator.

Table 5: Frozen-Qwen repeated categorical control over 39 targets. Lower is better. This study uses the unconstrained special case, not the main Lambert comparison.

Method	Expected MAE	MC-100 MAE	Invalid rate
Zero-shot	.1928	.1949	.0422
Explicit target prompt	.1580	.1620	.4115
Cross-prompt EEF	.1163	.1212	.0487
Finite-bank EEF, 32	.0543	.0710	.0058
Finite-bank EEF, 64	**.0395**	**.0601**	**.0055**

C.2 Natural SDXL prompt shifts

The six base prompts describe a car, cabin, bird, fruit still life, city street, and dog. Target prompts add combinations of color, illumination, weather, environment, or composition. SDXL 1.0 is frozen at 512^2, 30 denoising steps, and classifier-free guidance 5. Each bank contains 256 base images and 128 independently generated target-prompt images.

For each of five splits per family, base images are divided into 128 design and 128 test candidates; targets are divided into 64 references and 64 evaluation images. PCA and random Fourier features are fitted on base design data only. DINO kernel bandwidth is the median-distance value from base design features, not target evaluation data. Table 2 in Section 4.2 reports the results at a 40% required design-bank ESS.

Direct target prompt compares two halves of the target generator distribution and is shown only as an oracle-like reference. It changes the conditioning call and therefore does not solve reference-defined adaptation of the base prompt.

C.3 CIFAR diffusion and matched SMC

We use the frozen `google/ddpm-cifar10-32` checkpoint with 50 stochastic DDIM steps. The target is the top 30% of CIFAR-10 test images by mean RGB brightness. Interpretable image features are standardized by the generated bank. The ray score from the original finite EEF is used as a scalar sufficient statistic for both methods. Ray tempering and the exact scalar projection are both fitted at 20% ESS on all 5,000 generated candidates.

The coverage solution has active multiplier $\eta = 2.304$ and fit-bank ESS .200000. Seven intermediate states per trajectory from 5,000 frozen trajectories train the continuation CNN; trajectory identifiers, rather than individual states, define the 80/20 split. The exact and ray networks use the same architecture, AdamW optimizer, batch size 512, and Poisson objective. The Lambert network's best validation correlation between predicted and terminal log ratios is .626.

Evaluation uses 512 particles, five common random seeds, terminal resampling at every run, and 400 samples for MMD. MMD bandwidth is chosen from pooled base and target feature distances within each split. The main table reports paired t-tests and paired bootstrap intervals. For terminal MMD, the Lambert reduction interval is $[47.3, 60.2]\%$; for learned-SMC MMD it is $[42.6, 69.7]\%$.

Within the Lambert arm, learned SMC changes MMD from .0377 to .0333, an 11.6% mean reduction, but the five-split paired test is not significant ($p = .590$). We therefore do not claim that learned SMC universally beats terminal resampling. Its clear benefit in this pilot is diversity: unique output fraction rises from .233 to .734 while preserving the stronger Lambert target.

C.4 Practical repeated-output language control

The exact coverage experiments above use open-ended Qwen representations. A separate categorical study verifies that the broader Neural EEF framework can control actual repeated LLM outputs. Thirty-nine occupation targets and eight prompt paraphrases request a one-token gender label. The generator is frozen; invalid completions remain in the candidate bank. A two-dimensional statistic contains female and invalid indicators.

MC-100 error decreases monotonically with 4, 8, 16, 32, and 64 candidates. Residual error is concentrated in one-sided banks, making the support limit observable. We do not mix this experiment into the matched-ESS claims because its target is categorical and often exactly attainable without an active ESS constraint.

D Comparator Definitions

Ray EEF fits the regularized empirical exponential-family dual and shrinks its natural parameter to the ESS floor. **Exact frontier** sets $\epsilon = d_\rho + \delta$. **Exact bootstrap** uses Eq. (8). **Stable balancing** minimizes $\frac{1}{2}\mathbb{E}_u r^2$ at the same moment tolerance, then fits a nonnegative affine ratio for unseen-bank evaluation. **Density ratio** is a class-balanced logistic regression distinguishing design candidates from target references; its logit is tempered to the same ESS floor. **Base** is uniform on the candidate bank. **Direct target prompt** samples the generator from the verbalized target condition and is not available for general reference-only targets.

Scalar EEF uses one standardized target score. Neural EEF uses whitened frozen embeddings. Quadratic EEF appends unique second-order products. Kernel EEF uses D random Fourier features $\sqrt{2/D}\cos(\omega_j^\top z + b_j)$ with frequencies fixed from design data. MMD and all evaluation encoders are excluded from fitting unless explicitly named as the sufficient statistic.

E Reproducibility

All experiments ran on one NVIDIA RTX 4090 with 24 GB memory. Large generators were loaded from frozen local checkpoints; SDXL generation used roughly 10.5 GiB peak allocated memory. Projection and reporting are CPU operations. After caching the feasibility frontier, the exact KL stage takes roughly 1–4 ms on the reported neural-feature banks.