Distilling Conditional Image Generators into Spatial Effect Maps

Eun Ryung Lee
Department of Statistics
Sungkyunkwan University
Seoul, Republic of Korea
erlee@skku.edu

Seyoung Park
Department of Applied Statistics
Yonsei University
Seoul, Republic of Korea
ishspsy@yonsei.ac.kr

Abstract

Spatial effect maps summarize associations between image responses and covariates at each location, but these associations remain implicit in conditional image generators. In this article, we propose *generative effect distillation* (GED) to estimate spatial effect maps from a conditional generator. We treat the generator as a teacher and project its information onto a regression student with spatial coefficient functions. The target covariate distribution and the discrepancy used for projection specify the meaning of these maps. We develop mean, effect-gradient, and structured score distillation objectives and establish identities for their population targets. For a fixed teacher, our finite-query risk bounds separate teacher bias, spatial approximation, query design, Monte Carlo, and student optimization error. These bounds quantify the contribution of the query budget to map estimation accuracy. We also study a Gaussian model with a finite spatial basis. A lower bound for jointly estimating the intercept and slope coefficients has the same dependence on query sizes as the squared stochastic terms in the upper bound. Across eleven blocks spanning synthetic, semi-real, pretrained SDXL-Turbo/FFHQ-256, CelebA conditional diffusion, and face-image analyses, GED attains the lowest standardized integrated squared error (SISE) among the direct-regression and augmentation baselines. Paired Student t-tests reject equality with the strongest non-GED baseline at $p < 0.05$ in every block. For a CelebA conditional residual diffusion teacher, GED with single-step Tweedie score access attains SISE 0.009, a reduction by a factor greater than 25 relative to the strongest non-GED baseline.

1 Introduction

In neuroimaging, pathology, medical imaging, and remote sensing, a central statistical problem is to identify image regions associated with clinical or experimental covariates. A classical approach to this problem is image-response regression. For subject i on a spatial image domain $\mathcal{S}$, let $Y_i(s)$ denote the image intensity at spatial location $s \in \mathcal{S}$, and let X_i denote a p-dimensional covariate vector. With an image-specific spatial random effect $\alpha_i(s)$ and a noise process $\epsilon_i(s)$, the spatially varying coefficient model is

$$Y_i(s) = \beta_0(s) + X_i^\top \beta(s) + \alpha_i(s) + \epsilon_i(s), \qquad s \in \mathcal{S}, \tag{1}$$

where $\beta_j(s)$ is an interpretable spatial effect map for the jth covariate. Recent neural regression methods estimate the coefficient functions in (1) using deep networks over spatial coordinates to improve estimation accuracy while preserving the spatially varying coefficient structure (Zhang et al., 2023). However, this regression formulation models only the conditional mean from observed images and does not characterize the full conditional image distribution $P(Y \mid X)$.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In contrast, generative adversarial networks (GANs) (Goodfellow et al., 2014), variational autoencoders (VAEs) (Kingma and Welling, 2014), normalizing flows (Rezende and Mohamed, 2015), and diffusion or score-based models (Ho et al., 2020; Song et al., 2021) model full image distributions. Conditional variants such as conditional GANs (Mirza and Osindero, 2014) and continuous conditional GANs (Ding et al., 2021) learn image distributions indexed by covariates and generate plausible images at fixed covariate values. Related approaches generate counterfactual image pairs (Van Looveren et al., 2021; Augustin et al., 2022; Ha and Bender, 2025), optimize representations by score distillation (Poole et al., 2023), or compress diffusion teachers into faster student generators (Yin et al., 2024; Zhou et al., 2024). However, the covariate effects learned by these models remain implicit in network parameters, latent variables, or score fields. Because a conditional generator may be nonlinear, heteroskedastic, or multimodal, differences between individual generated images depend on the sampling seed and baseline covariates. A population-level spatial effect map requires a target that specifies the covariate population and the teacher information being summarized.

To address this problem, we propose *generative effect distillation* (GED) to extract interpretable spatial effect maps from a conditional image generator. We treat a generator with distribution $P_\theta(\,\mathrm{d}y \mid x)$ as a teacher and project selected teacher information onto a structured regression student. The student's spatial coefficient functions $\beta_j(s)$ summarize covariate effects under a target design ν. This construction defines a statistical target for the maps and separates teacher bias from the error in estimating that target with finitely many queries. The target differs from the direct fit to observed images in image-response regression (Zhang et al., 2023) and from the generated images in conditional image generation (Ding et al., 2021). Appendix D compares these statistical objects.

First, we introduce the teacher-to-map functional $\mathcal{T}_{\nu,\mathcal{D}}(P_\theta)$ to specify the target covariate distribution, the teacher information to be projected, and the discrepancy used in the projection. We develop three objectives for different forms of teacher access: conditional mean distillation, teacher-gradient effect distillation, and structured score distillation based on conditional Fisher divergence. The choice of access and discrepancy determines the population summary represented by the coefficient maps.

Second, we identify the population target of each objective. For a linear student, the mean-distilled map is a covariance-weighted projection of the teacher conditional mean. We establish identities relating this projection to average teacher gradients. Effect distillation targets the teacher gradient averaged over the target covariate population. For score distillation, we consider small admissible changes in the coefficient maps and the resulting changes in the student score. At an interior population optimum, the residual between the teacher and student scores is orthogonal to these changes under the weighting in the objective. These results specify the statistical meaning of the maps obtained from each form of teacher access.

Third, we establish finite-query risk bounds for mean distillation with a fixed teacher and derive sufficient conditions for consistency. The bounds separate teacher bias, spatial basis approximation, query-design error, teacher Monte Carlo error, and optimization error. This separation identifies the terms that can be reduced by increasing the query budget. We also study a Gaussian observation model with a finite spatial basis. In this model, we derive a lower bound for jointly estimating the intercept and slope coefficients. Its dependence on the query sizes matches that of the squared stochastic terms in the upper bound up to constants.

Fourth, we evaluate GED across eleven empirical blocks with small observed sample sizes. These experiments span controlled and learned synthetic teachers, semi-real digits, real handwriting, Olivetti face morphology, FFHQ measured attributes, CIFAR-10 conditional pixel diffusion, and latent diffusion for digits. We also analyze a pretrained SDXL-Turbo/FFHQ-256 generator and a CelebA conditional residual DDPM teacher with mean and single-step Tweedie score access. GED attains the lowest standardized integrated squared error among the direct-regression and augmentation baselines in every block.

The remainder of the article is organized as follows. Sections 2 and 3 introduce the notation and problem setup, Section 4 develops GED, Section 5 establishes the theoretical guarantees, and Section 6 presents the empirical results. Proofs and additional results are given in the Supplement.

2 Notation

Table 2 summarizes the notation used throughout this article. For a scalar function f on $\mathcal{S}$, let $\|f\|^2_{L^2(\mathcal{S})} = \int_{\mathcal{S}} f(s)^2 \, \mathrm{d}s$. For a finite-channel function $f = (f_1, \ldots, f_{C_{\mathrm{img}}})$, use the same notation with $\|f\|^2_{L^2(\mathcal{S})} = \sum_{c=1}^{C_{\mathrm{img}}} \int_{\mathcal{S}} f_c(s)^2 \, \mathrm{d}s$, and for a vector-valued coefficient map $b = (b_1, \ldots, b_p)$, put $\|b\|^2_{L^2(\mathcal{S})} = \sum_{j=1}^{p} \|b_j\|^2_{L^2(\mathcal{S})}$. Let $C > 0$ denote a generic absolute constant independent of the sample size, the number of query points, and the number of teacher samples, where the value of C may change from line to line.

3 Problem setup

Let $X \in \mathcal{X} \subseteq \mathbb{R}^p$ denote covariates and let $Y \in \mathcal{Y} = L^2(\mathcal{S})$ denote an image on a spatial domain $\mathcal{S} \subset \mathbb{R}^d$. We observe training data $\{(X_i, Y_i)\}_{i=1}^n$ and first train, fine-tune, or otherwise obtain a conditional image generator with distribution $P_\theta(\,\mathrm{d}y \mid x)$, henceforth referred to as the teacher. When the conditional expectation exists, write $\mu_\theta(x, s) = \mathbb{E}_\theta\{Y(s) \mid X = x\}$ for the teacher conditional mean. When the noised teacher density at level τ exists, write $u_{\theta,\tau}(y, x) = \nabla_y \log p_{\theta,\tau}(y \mid x)$ for its score. Depending on the architecture, the teacher may provide conditional samples $Y \sim P_\theta(\cdot \mid x)$, a denoised estimate of μ_θ, scores $u_{\theta,\tau}$, and gradients with respect to the conditioning variable x by automatic differentiation. For a C_{img}-channel image, $Y \in L^2(\mathcal{S}; \mathbb{R}^{C_{\mathrm{img}}})$, and we retain the notation $L^2(\mathcal{S})$ for the finite direct-sum Hilbert space. Under the convention in Table 2, all projection identities and norm bounds below apply channel-wise. The location index s has a common scientific meaning only when the analyzed images are expressed in a common spatial coordinate system, making registration, the reference frame, and interpolation part of the definition of the image response. In our face-image analyses, we use aligned images. Because residual registration error can alter the population map targeted at a fixed coordinate, we evaluate sensitivity to spatial misalignment in Appendix F.

To define the population target, we specify a target design distribution ν over the covariates. This distribution may be the empirical distribution of the observed covariates, a balanced design chosen by the analyst, or a scientifically meaningful population distribution. The choice of ν is an essential part of the estimand. In particular, when the teacher is nonlinear in x, a map averaged over younger subjects need not equal a map averaged over older subjects.

3.1 Interpretable student family

The student is a structured image-response model. Let $m_\phi(x, s)$ denote the student conditional mean with parameter ϕ. The simplest version is

$$m_\phi(x, s) = \beta_0(s) + x^\top \beta(s), \qquad \phi = (\beta_0, \beta_1, \ldots, \beta_p), \tag{2}$$

where each $\beta_j : \mathcal{S} \to \mathbb{R}$ is a spatial coefficient map. We also allow structured extensions such as interactions, group-specific effects, or local effect maps. Let $z(x)$ denote a user-chosen basis of covariates and let $B(s)$ denote the corresponding spatial coefficient vector. Then,

$$m_\phi(x, s) = \beta_0(s) + z(x)^\top B(s), \tag{3}$$

where the entries of $B(s)$ are interpreted as spatial summaries of the chosen covariate basis. For the standard basis vector e_j, the linear student satisfies $m_\phi(x + e_j, s) - m_\phi(x, s) = \beta_j(s)$. Thus, $\beta_j(s)$ is the fitted change at location s associated with a one-unit increase in covariate x_j, holding the other included covariates fixed within the student. Under mean access, it is the slope of the best linear projection of the teacher mean in Proposition 2. Under effect access, it equals the design-averaged teacher derivative in Theorem 3, whereas under score access, it satisfies the tangent-space orthogonality condition in Proposition 4. Here, the term "interpretable" refers to this explicit regression estimand indexed by covariates and spatial locations, and not to a causal effect, a scientifically validated mechanism, or a faithful representation of the teacher's internal features. For score distillation, we specify the student as a Gaussian image-response model. Let $P_\phi(\,\mathrm{d}y \mid x)$ denote the student distribution, and let C_ϕ denote a diagonal, low-rank plus diagonal, or spatially smoothed covariance operator. In the Hilbert-space formulation, C_ϕ is a positive self-adjoint and trace-class operator, whereas under finite image discretization, it is the corresponding covariance matrix. We write

$$P_\phi(\,\mathrm{d}y \mid x) = \mathcal{N}\{m_\phi(x), C_\phi\}. \tag{4}$$

In this model, the mean parameters remain the primary target of interpretation. The x-independent covariance is a restriction of the student model used for score projection and not a homoskedasticity assumption on the teacher. Accordingly, a heteroskedastic or multimodal teacher is projected onto this Gaussian student class.

3.2 Distillation as projection, not augmentation

The population target is not a collection of synthetic images, but a projection of teacher information onto the interpretable student class. Let Φ denote the parameter space for the student. Let $\mathcal{A}_\theta(x)$ denote the teacher information available at covariate value x, and let $\mathcal{A}_\phi(x)$ denote the corresponding student quantity. Depending on the access mode, $\mathcal{A}_\theta(x)$ may be the conditional distribution $P_\theta(\cdot \mid x)$, the conditional mean $\mu_\theta(x, \cdot)$, the covariate-gradient field $\nabla_x \mu_\theta(x, \cdot)$, or the noised conditional score $u_{\theta,\tau}(\cdot, x)$. Let $\mathcal{D}$ denote a nonnegative discrepancy between the corresponding teacher and student quantities, where each $\mathcal{D}$ is paired with the declared access mode $\mathcal{A}$. When we emphasize the dependence on the student class and access mode, we write the target functional as $\mathcal{T}^{\Phi,\mathcal{A}}_{\nu,\mathcal{D}}(P_\theta)$.

Definition 1. *Fix ν, a student family $\{P_\phi(\cdot \mid x) : \phi \in \Phi\}$ with mean containing spatial maps, an access mode $\mathcal{A}$, and a discrepancy $\mathcal{D}$ for which $x \mapsto \mathcal{D}\{\mathcal{A}_\theta(x), \mathcal{A}_\phi(x)\}$ is measurable for every $\phi \in \Phi$. Let $\Pi_{\mathrm{map}}(\phi)$ denote the coefficient-map component of ϕ and write $\Phi^\star_{\nu,\mathcal{D}}(P_\theta)$ for the set of $\phi \in \Phi$ at which $\mathbb{E}_{X\sim\nu}\mathcal{D}\{\mathcal{A}_\theta(X), \mathcal{A}_\phi(X)\}$ attains its finite infimum over Φ. The teacher-to-map functional is*

$$\mathcal{T}^{\Phi,\mathcal{A}}_{\nu,\mathcal{D}}(P_\theta) = \left\{\Pi_{\mathrm{map}}(\phi) : \phi \in \Phi^\star_{\nu,\mathcal{D}}(P_\theta)\right\}, \tag{5}$$

which we abbreviate by $\mathcal{T}_{\nu,\mathcal{D}}(P_\theta)$ when Φ and $\mathcal{A}$ are clear from context, and identify with its unique coefficient-map tuple when $\Phi^\star_{\nu,\mathcal{D}}(P_\theta)$ is a singleton. Each element is a distilled spatial effect map.

Different choices of $\mathcal{D}$ define distinct population targets. Specifically, the squared mean discrepancy yields the linear spatial projection of the teacher conditional mean in $L^2(\nu \times \mathcal{S})$, whereas the gradient discrepancy defines an average sensitivity map from $\nabla_x \mu_\theta(X, \cdot)$. Under the conditional Fisher divergence, we obtain the Gaussian student whose score is closest to the teacher score along the student tangent space. When the teacher provides only generated samples, we may also use the maximum mean discrepancy (MMD) (Gretton et al., 2012).

The estimand in (5) distinguishes GED from ordinary synthetic data augmentation. Augmentation enlarges a training set and then fits an existing estimator, whereas GED defines the estimand as a functional of the teacher distribution, the target design, the student class, the access mode, and the discrepancy, and then uses teacher queries to estimate that functional. Hence, when two choices of ν, Φ, $\mathcal{A}$, or $\mathcal{D}$ yield different maps, they are distinct population summaries of the same teacher and not contradictory visualizations.

The following proposition shows that this distinction is not only semantic. To isolate the role of the target design, consider a one-dimensional covariate, let $a, c \in L^2(\mathcal{S})$ with $\|c\|_{L^2(\mathcal{S})} > 0$, fix $\tau^2 > 0$, and for each $m \in \mathbb{R}$ let $\nu_m = N(m, \tau^2)$. Let $\beta^\star_m$ denote the slope component of the mean-distilled linear projection of $\mu_\theta(x, s) = a(s)x + c(s)x^2$ under ν_m.

Proposition 1. *For every $m \in \mathbb{R}$, it holds that $\beta^\star_m(s) = a(s) + 2m\, c(s)$ for almost every $s \in \mathcal{S}$. If $m_1 \neq m_2$, then $\|\beta^\star_{m_1} - \beta^\star_{m_2}\|_{L^2(\mathcal{S})} = 2|m_1 - m_2|\|c\|_{L^2(\mathcal{S})} > 0$.*

Thus, synthetic samples drawn only at an observed covariate design estimate the projection associated with that design, whereas GED can target a scientifically chosen ν and thereby report a different population map for the same teacher distribution.

Remark 1. *The notation $\mathcal{T}_{\nu,\mathcal{D}}(P_\theta)$ records that the interpreted object depends on the teacher distribution, target covariate design, student class, access mode, and discrepancy. In general, $\mathcal{T}_{\nu,\mathcal{D}}(P_\theta)$ is a set; a reported coefficient map is identified only after a uniqueness condition or selection rule is specified. For the Hilbert-space least-squares projections below, the fitted student mean is unique even when coefficient tuples are not, and we report the Moore–Penrose minimum-norm representative when a single tuple is needed. If a population or empirical design matrix is singular, null-direction differences induce the same fitted map under ν or the same fitted values on the query design, respectively, with the same Moore–Penrose reporting convention.*

4 Generative effect distillation

4.1 Mean distillation

Recall that $\mu_\theta(x, s)$ denotes the teacher conditional mean,

$$\mu_\theta(x, s) = \mathbb{E}_\theta\{Y(s) \mid X = x\}. \tag{6}$$

Assume throughout this subsection that $\mathbb{E}_{X\sim\nu}\|\mu_\theta(X, \cdot)\|^2_{L^2(\mathcal{S})} < \infty$. The mean-distilled student solves

$$\phi^\star_{\text{mean}} = \arg\min_{\phi\in\Phi} \mathbb{E}_{X\sim\nu} \int_{\mathcal{S}} [m_\phi(X, s) - \mu_\theta(X, s)]^2 \, \mathrm{d}s. \tag{7}$$

In practice, we estimate μ_θ using generated teacher samples or a diffusion denoiser. The following proposition establishes the Hilbert-space normal equations for the unrestricted linear student. For deterministic covariates, define $z(x) = (1, x^\top)^\top$.

Proposition 2. *Let $Z = (1, X^\top)^\top$, let $\Gamma_X = \mathbb{E}_\nu(ZZ^\top)$, and assume that $\mathbb{E}_\nu\|Z\|_2^2 < \infty$ and $\mu_\theta \in L^2(\nu \times \mathcal{S})$. For $b \in L^2(\mathcal{S}; \mathbb{R}^{p+1})$, write $m_b(x, s) = z(x)^\top b(s)$ and put $h_\theta(s) = \mathbb{E}_\nu\{Z\mu_\theta(X, s)\}$. Then, the minimizer set of (7) over the linear class is nonempty, and $b^\star$ is a minimizer if and only if $\Gamma_X b^\star(s) = h_\theta(s)$ for almost every $s \in \mathcal{S}$. The fitted projection is unique in $L^2(\nu \times \mathcal{S})$. The minimum-norm coefficient representative is $b^\star_{\theta,\min}(s) = \Gamma_X^\dagger h_\theta(s)$, where $\Gamma_X^\dagger$ is the Moore–Penrose inverse. If Γ_X is nonsingular, then the coefficient map is unique and*

$$b^\star_\theta(s) = \Gamma_X^{-1}\mathbb{E}_\nu\{Z\mu_\theta(X, s)\} \quad \textit{for almost every } s \in \mathcal{S}. \tag{8}$$

Let $\bar{x} = \mathbb{E}_\nu X$ and $\Sigma_X = \mathbb{E}_\nu\{(X - \bar{x})(X - \bar{x})^\top\}$. If Σ_X is nonsingular, the slope component satisfies

$$\beta^\star_\theta(s) = \Sigma_X^{-1}\mathbb{E}_\nu\{(X - \bar{x})\mu_\theta(X, s)\} \quad \textit{for almost every } s \in \mathcal{S}. \tag{9}$$

Proposition 5 gives an equivalent centered slope representation in terms of Σ_X. This representation also applies when Σ_X is singular. Under this parameterization, the minimum-norm centered slope map is $\beta^c_{\theta,\min}(s) = \Sigma_X^\dagger h_c(s)$ with $h_c(s) = \mathbb{E}_\nu\{(X - \bar{x})\mu_\theta(X, s)\}$.

4.2 Effect distillation

When the teacher mean is differentiable in the covariates, let $g_{\theta,j}(x, s)$ denote the local covariate effect for covariate j, $g_{\theta,j}(x, s) = \frac{\partial}{\partial x_j}\mu_\theta(x, s)$. When we compute gradients by automatic differentiation through a conditional sampler, Proposition 6 gives the regularity conditions under which $\partial\mu_\theta/\partial x_j$ equals the expectation of the pathwise derivative of the sampler.

A global coefficient map cannot generally equal $g_{\theta,j}(x, s)$ for every x unless the teacher mean is linear in x. We use the gradient in one of two ways.

First, let $\bar{g}_{\theta,j}(s)$ denote the averaged teacher gradient: $\bar{g}_{\theta,j}(s) = \mathbb{E}_{X\sim\nu}\{g_{\theta,j}(X, s)\}$. Second, if nonlinear effects are scientifically important, the student can be expanded to include local or basis-indexed maps, for example $m_\phi(x, s) = \beta_0(s) + z(x)^\top B(s)$, and the derivative $\partial m_\phi(x, s)/\partial x_j$ is matched to $g_{\theta,j}(x, s)$.

The effect-distillation loss is

$$\mathcal{L}_{\text{eff}}(\phi) = \sum_{j=1}^{p} \mathbb{E}_{X\sim\nu} \int_{\mathcal{S}} \Big[\frac{\partial}{\partial x_j} m_\phi(X, s) - g_{\theta,j}(X, s)\Big]^2 \, \mathrm{d}s. \tag{10}$$

Here, $\mathrm{d}s$ belongs to the inner spatial integral, and the outer expectation averages the resulting integrated loss over $X \sim \nu$. For the linear student in (2), we have $\partial m_\phi(x, s)/\partial x_j = \beta_j(s)$, which implies that we minimize (10) by fitting $\beta_j(s)$ to the average teacher gradient.

The following proposition connects mean distillation to average teacher gradients through the covariate design. Following the Stein-kernel formulation in Stein's method and transport inequalities (Ledoux et al., 2015), let $\bar{x} = \mathbb{E}_\nu X$ and let $T_\nu(X)$ be a Stein kernel satisfying $\mathbb{E}_\nu\{(X - \bar{x})f(X)\} = \mathbb{E}_\nu\{T_\nu(X)\nabla_x f(X)\}$ for differentiable functions f in the domain of this Stein identity.

Proposition 3. *Assume that $\mathbb{E}_\nu\|X\|_2^2 < \infty$, $\mu_\theta \in L^2(\nu \times \mathcal{S})$, and $\mu_\theta(\cdot, s)$ belongs to the domain of T_ν for almost every $s \in \mathcal{S}$, with the Stein identity holding for $f = \mu_\theta(\cdot, s)$. Then $\int_{\mathcal{S}} \|\mathbb{E}_\nu\{T_\nu(X)\nabla_x\mu_\theta(X, s)\}\|_2^2 \, \mathrm{d}s < \infty$, and the centered minimum-norm slope representative in Proposition 5 satisfies*

$$\beta^c_{\theta,\min}(s) = \Sigma_X^\dagger \mathbb{E}_\nu\{T_\nu(X)\nabla_x\mu_\theta(X, s)\}. \tag{11}$$

If Σ_X is nonsingular, then the slope map is unique and

$$\beta^\star_\theta(s) = \Sigma_X^{-1}\mathbb{E}_\nu\{T_\nu(X)\nabla_x\mu_\theta(X, s)\} \quad \text{for almost every } s \in \mathcal{S}.$$

In particular, if $X \sim \mathcal{N}(\bar{x}, \Sigma_X)$ with nonsingular Σ_X, we may take the canonical Gaussian Stein kernel $T_\nu(X) = \Sigma_X$, in which case, for almost every $s \in \mathcal{S}$,

$$\beta^\star_{\theta,j}(s) = \mathbb{E}_\nu\Big[\frac{\partial}{\partial x_j}\mu_\theta(X, s)\Big], \qquad j = 1, \dots, p. \tag{12}$$

Proposition 3 identifies the centered minimum-norm slope map as a design-weighted average derivative of the teacher mean. For a nonsingular Gaussian design, the weights reduce to the identity, and the map equals the average local sensitivity of the teacher. For a non-Gaussian design, the Stein identity and its domain condition determine the weighting by the covariate distribution. If either condition fails for the chosen ν, the projection in Proposition 5 remains the mean-distillation target. The map summarizes the teacher over the covariate population and can differ from its derivative at a single point.

4.3 Structured score distillation

Throughout this subsection, score quantities are understood after a fixed finite-dimensional image discretization or in a declared Hilbert score space whenever inverse covariance operators are used.

Diffusion and score-based teachers often provide conditional scores more readily than exact likelihoods (Ho et al., 2020; Song et al., 2021). Let $P_{\theta,\tau}(\cdot \mid x)$ be the teacher distribution after adding Gaussian noise at level τ. The corresponding teacher score is the previously introduced field $u_{\theta,\tau}(y, x)$, that is, $u_{\theta,\tau}(y, x) = \nabla_y \log p_{\theta,\tau}(y \mid x)$. Define the noised Gaussian student distribution analogously by $P_{\phi,\tau}(\cdot \mid x) = \mathcal{N}\{m_\phi(x), C_{\phi,\tau}\}$, where $C_{\phi,\tau}$ is the student covariance after the same noising convention at level τ. We use $C_{\phi,\tau}^{-1}$ only when this inverse is bounded on the chosen score space. For the Gaussian student, let $u_{\phi,\tau}(y, x)$ denote the student score:

$$u_{\phi,\tau}(y, x) = \nabla_y \log p_{\phi,\tau}(y \mid x) = -C_{\phi,\tau}^{-1}\{y - m_\phi(x)\}. \tag{13}$$

Let $R_{\theta,\tau}(\cdot \mid X)$ denote a sampling distribution that is fixed when differentiating the student parameters, such as the noised teacher distribution, a replay distribution, or a stop-gradient noised student distribution, and let W_τ denote a positive semidefinite weighting operator. Write $\|v\|_{W_\tau}^2 = \langle v, W_\tau v\rangle$ for the induced weighted seminorm. We define the structured score-distillation objective as

$$\mathcal{L}_{\mathrm{score}}(\phi) = \mathbb{E}_{X\sim\nu}\mathbb{E}_{\tilde{Y}\sim R_{\theta,\tau}(\cdot|X)}\Big[\Big\|u_{\theta,\tau}(\tilde{Y}, X) - u_{\phi,\tau}(\tilde{Y}, X)\Big\|_{W_\tau}^2\Big]. \tag{14}$$

When $R_{\theta,\tau}(\cdot \mid x) = P_{\theta,\tau}(\cdot \mid x)$, the objective in (14) is a restricted conditional Fisher divergence related to score matching (Hyvärinen, 2005). For other fixed choices of $R_{\theta,\tau}$, we minimize the weighted score residual under that sampling distribution. In both cases, we project the teacher score field onto the score family specified by the regression student. The spatial coefficient maps parameterize this restricted family.

The following proposition gives the first-order condition for the structured score projection. For any ϕ, put $r_{\phi,\tau}(y, x) = u_{\theta,\tau}(y, x) - u_{\phi,\tau}(y, x)$.

Proposition 4. *Write $Z_0 = 1$ and $Z_j = X_j$ for $j = 1, \dots, p$. Suppose that $R_{\theta,\tau}$, $C_{\phi,\tau}$, and W_τ are fixed when differentiating the mean parameters and $m_\phi(x, s) = \sum_{j=0}^p Z_j\beta_j(s)$. For each $j = 0, \dots, p$, let $\mathcal{H}_j$ be the tangent space of admissible perturbations of β_j at ϕ: if $h \in \mathcal{H}_j$, then replacing β_j by $\beta_j + th$ is admissible for all sufficiently small $|t|$. Assume that $\mathcal{L}_{\mathrm{score}}(\phi) < \infty$ and that, for every $h \in \mathcal{H}_j$,*

$$\mathbb{E}\Big[\Big|Z_j\Big\langle C_{\phi,\tau}^{-1}h, W_\tau r_{\phi,\tau}(\tilde{Y}, X)\Big\rangle\Big|\Big] < \infty \quad \text{and} \quad \mathbb{E}\Big[Z_j^2\Big\|C_{\phi,\tau}^{-1}h\Big\|_{W_\tau}^2\Big] < \infty.$$

At any interior population minimizer of (14), the score residual is orthogonal to each intercept and covariate-spatial tangent direction:

$$\mathbb{E}\left[Z_j\left\langle C_{\phi,\tau}^{-1}h, W_\tau r_{\phi,\tau}(\tilde{Y},X)\right\rangle\right]=0, \qquad h\in\mathcal{H}_j, \quad j=0,\dots,p. \tag{15}$$

A vector form of (15), under a full $L^2(\mathcal{S})$ score-space tangent class, is recorded as Proposition 7.

These conditions are the score analogue of least-squares normal equations: they state that no spatial coefficient map direction further reduces the teacher-student score residual, without claiming that the Gaussian student reproduces the full teacher distribution.

4.4 Combined objective

Let $\widehat{\mathcal{L}}(\phi)$ denote the empirical training objective, let $\widehat{\mathcal{L}}_{\mathrm{data}}(\phi)$ denote the empirical data-fitting loss, let $\widehat{\mathcal{L}}_{\mathrm{mean}}(\phi)$, $\widehat{\mathcal{L}}_{\mathrm{eff}}(\phi)$, and $\widehat{\mathcal{L}}_{\mathrm{score}}(\phi)$ denote empirical versions of the distillation losses, let $\Omega(\phi)$ denote a spatial regularizer, and let λ_μ, λ_g, λ_s, and λ_{reg} denote nonnegative tuning parameters. We use the empirical template

$$\widehat{\mathcal{L}}(\phi)=\widehat{\mathcal{L}}_{\mathrm{data}}(\phi)+\lambda_\mu\widehat{\mathcal{L}}_{\mathrm{mean}}(\phi)+\lambda_g\widehat{\mathcal{L}}_{\mathrm{eff}}(\phi)+\lambda_s\widehat{\mathcal{L}}_{\mathrm{score}}(\phi)+\lambda_{\mathrm{reg}}\Omega(\phi). \tag{16}$$

Let P_0 denote the observational distribution of (X,Y) and let $\mathcal{L}_{\mathrm{data}}(\phi;P_0)$ denote the corresponding population data loss. Let

$$\mathcal{L}_{\mathrm{mean}}(\phi)=\mathbb{E}_{X\sim\nu}\int_{\mathcal{S}}\left[m_\phi(X,s)-\mu_\theta(X,s)\right]^2\,\mathrm{d}s$$

denote the population mean-distillation loss in (7). Define

$$Q_{\mathrm{comb}}(\phi)=\mathcal{L}_{\mathrm{data}}(\phi;P_0)+\lambda_\mu\mathcal{L}_{\mathrm{mean}}(\phi)+\lambda_g\mathcal{L}_{\mathrm{eff}}(\phi)+\lambda_s\mathcal{L}_{\mathrm{score}}(\phi)+\lambda_{\mathrm{reg}}\Omega(\phi).$$

When Q_{comb} attains its finite infimum over Φ, put

$$\Phi^\star_{\mathrm{comb}}=\arg\min_{\phi\in\Phi}Q_{\mathrm{comb}}(\phi), \qquad \mathcal{T}_{\mathrm{comb}}=\{\Pi_{\mathrm{map}}(\phi):\phi\in\Phi^\star_{\mathrm{comb}}\}.$$

When $\mathcal{T}_{\mathrm{comb}}$ is not a singleton, reporting a single coefficient-map tuple requires the uniqueness or selection convention in Remark 1. In this objective, we fit the student to observed images and to the chosen teacher information through the data-fitting and distillation losses, respectively. We use the regularizer to impose spatial smoothness or sparsity. In a given application, we include only the distillation terms supported by the available teacher access, and we report the design distribution, access mode, discrepancy, tuning weights, and estimated coefficient maps. Our estimation procedure has five steps. In Step 1, we validate the teacher on held-out observed images. In Step 2, we specify the target design ν. In Step 3, we query the teacher quantity supported by the available access. In Step 4, we fit the structured student. In Step 5, we report the estimated coefficient maps with ν, the discrepancy, M, K, and uncertainty diagnostics. We use mean distillation with sample-only access and may use effect distillation with a teacher that supports differentiation in the covariates. With a calibrated conditional score, we may use score distillation and assess sensitivity to the student covariance. When feasible, we also compare the score-distilled map with the mean-distilled map. Stability across queries assesses fidelity to the teacher. Scientific validity requires external observations, prior scientific structure, or independent replication.

5 Theory, estimation, and uncertainty

Our theoretical analysis identifies the population map targeted by each access mode and decomposes the resulting finite-query estimation error. Specifically, Proposition 3 establishes the Stein gradient identity, Proposition 4 characterizes the normal equation for score projection, and Theorem 3 identifies the population target of effect-gradient distillation. For finite-query estimation, Theorem 1 below separates teacher bias, spatial basis approximation, query-design error, teacher Monte Carlo sampling error, and student optimization error.

5.1 Teacher querying

We query the teacher under three access modes corresponding to our three distillation objectives. Under sample-only access, we draw K teacher samples $Y_{mk} \sim P_\theta(\cdot \mid x_m)$ at each query covariate x_m to estimate $\mu_\theta(x_m)$ by Monte Carlo averaging for the mean-distillation objective in (7). Under differentiable-generator access, we differentiate through the conditional generation or denoising map to estimate $\partial\mu_\theta/\partial x_j$ for the effect-distillation objective in (10). Under score access, we evaluate $u_{\theta,\tau}(y, x)$ at sampled noised images to optimize the score-distillation objective in (14). In our diffusion experiments, we also apply score access through Tweedie's identity (Efron, 2011) to construct single-step denoised queries before linear projection. The empirical score-access map is the projection of the clipped queries specified in Appendix I. A conditional-mean interpretation requires the noising, clipping, and averaging scheme to identify $\mu_\theta(x)$. In all modes, we may choose the query covariates $\{x_m\}_{m=1}^M$ from the empirical covariate distribution, a balanced design, or a prespecified grid over the covariates of interest.

5.2 Finite-query theory

The finite-query result analyzes the pure mean-distillation component after the teacher is fixed; with additional empirical data, effect, score, or regularization terms, the population target becomes the corresponding weighted projection. Let q be a positive integer, let $\psi(s) = (\psi_1(s), \ldots, \psi_q(s))$ be an orthonormal spatial basis in $L^2(\mathcal{S})$, let $a_0, \ldots, a_p \in \mathbb{R}^q$ be basis coefficients, and let $d_{\text{eff}} = (p+1)q$ denote the effective student dimension. Let $\mu_0(x, s) = \mathbb{E}\{Y(s) \mid X = x\}$ be the true conditional mean, and throughout this subsection let coefficient targets be the augmented Moore–Penrose representative defined in Remark 1. Let $\beta^\dagger_{\mu_0}$ and $\beta^\star_\theta$ be the slope components of the linear projections of μ_0 and μ_θ onto the student family; let $\beta^\star_{\theta,q}$, $\beta^\star_{\theta,M}$, and $\widetilde{\beta}_{\theta,MK}$ denote the finite-basis population projection, the empirical projection using exact teacher means, and the Monte Carlo projection, respectively; and let $\widehat{\beta}$ denote the computed estimator. Then,

$$\widehat{\beta} - \beta^\dagger_{\mu_0} = (\beta^\star_\theta - \beta^\dagger_{\mu_0}) + (\beta^\star_{\theta,q} - \beta^\star_\theta) + (\beta^\star_{\theta,M} - \beta^\star_{\theta,q}) + (\widetilde{\beta}_{\theta,MK} - \beta^\star_{\theta,M}) + (\widehat{\beta} - \widetilde{\beta}_{\theta,MK}), \quad (17)$$

corresponding, in order, to teacher bias, spatial basis approximation, finite query-design error, teacher Monte Carlo error, and optimization error.

Define the spatial approximation error by

$$\rho_q = \left\|\beta^\star_{\theta,q} - \beta^\star_\theta\right\|_{L^2(\mathcal{S})}. \quad (18)$$

Assumption 1. *The student uses the spatial basis ψ and satisfies $\int_{\mathcal{S}} \psi(s)\psi(s)^\top \mathrm{d}s = I_q$ and $m_\phi(x, s) = a_0^\top \psi(s) + \sum_{j=1}^p x_j a_j^\top \psi(s)$.*

Under Assumption 1, the finite-basis quantities in (17) are the slope maps of the corresponding Moore–Penrose minimum-norm coefficient matrices, and finite-channel responses are handled channel-wise. For fixed finite C_{img}, the scalar-channel bound applies to each channel and is aggregated in the direct-sum norm.

Let M and K denote the number of query covariates and the number of teacher samples per query, respectively, and let $Z = (1, X^\top)^\top$ and $Z_m = (1, X_m^\top)^\top$. Put $\Gamma_X = \mathbb{E}_\nu(ZZ^\top)$, and let $\kappa > 0$, $B_\theta > 0$, and $\sigma_\theta > 0$ denote a lower eigenvalue bound for Γ_X, the projected residual scale, and the projected teacher sampling scale, respectively. Define $V_\theta(x) = \int_{\mathcal{S}} \mu_\theta(x, s)\psi(s)\,\mathrm{d}s \in \mathbb{R}^q$, and at each query point X_m, define $\widehat{\mu}_\theta(X_m, s) = K^{-1}\sum_{k=1}^K Y_{mk}(s)$, $\widehat{V}_\theta(X_m) = \int_{\mathcal{S}} \widehat{\mu}_\theta(X_m, s)\psi(s)\,\mathrm{d}s$, and $E_m = \widehat{V}_\theta(X_m) - V_\theta(X_m)$, with the teacher samples Y_{mk} drawn as in Assumption 2. Let $A^\star_q = \Gamma^\dagger_X \mathbb{E}_\nu\{ZV_\theta(X)^\top\} \in \mathbb{R}^{(p+1)\times q}$ be the Moore–Penrose minimum-norm solution of the population normal equation $\Gamma_X A^\star_q = \mathbb{E}_\nu\{ZV_\theta(X)^\top\}$, and let $U_\theta(X) = V_\theta(X) - A^{\star\top}_q Z$ denote the projected population regression residual. For a scalar random variable ξ, write $\|\xi\|_{\psi_1} = \inf\{c > 0 : \mathbb{E}\exp(|\xi|/c) \le 2\}$.

Assumption 2. *The query covariates $X_1, \ldots, X_M$ are i.i.d. from ν, Γ_X satisfies $\lambda_{\min}(\Gamma_X) \ge \kappa > 0$, and $\Gamma_X^{-1/2} Z$ is sub-Gaussian with sub-Gaussian norm bounded by an absolute constant. At each query point X_m, we draw K teacher samples $Y_{mk} \sim P_\theta(\cdot \mid X_m)$ independently over k and independently of the teacher-sampling randomness at all other query points and form $\widehat{\mu}_\theta$ as above.*

For every deterministic $H \in \mathbb{R}^{(p+1)\times q}$ *with* $\|H\|_F = 1$, $\left\|B_\theta^{-1}\langle H, ZU_\theta(X)^\top\rangle\right\|_{\psi_1} \le C$, *and, for every* m, $\left\|\sqrt{K}\,\sigma_\theta^{-1}\langle H, Z_m E_m^\top\rangle\right\|_{\psi_1} \le C$.

The sampling protocol in Assumption 2 implies that the pairs (X_m, E_m) are independent across m and that $\mathbb{E}(E_m \mid X_m) = 0$ for every m; both properties are used in the proof of Lemma 2.

Assumption 3. *The teacher and reference conditional means satisfy* $\mu_\theta, \mu_0 \in L^2(\nu \times \mathcal{S})$.

Under Assumption 3 and $\mathbb{E}_\nu\|Z\|_2^2 < \infty$, define the exact projected teacher-bias term $d_\theta = \left\|\beta_\theta^\star - \beta_{\mu_0}^\dagger\right\|_{L^2(\mathcal{S})}$. When an $L^2(\nu \times \mathcal{S})$ teacher-error scale is also reported, we further assume that there exists $\delta_\theta \ge 0$ with $\mathbb{E}_\nu \int_{\mathcal{S}} \{\mu_\theta(X, s) - \mu_0(X, s)\}^2 \,\mathrm{d}s \le \delta_\theta^2$.

The basis approximation error under Assumption 1 is measured by ρ_q in (18), and the sub-exponential tail conditions in Assumption 2 hold under the bounded-design sufficient conditions in Lemma 3. When $\lambda_{\min}(\Gamma_X) \ge \kappa > 0$, Proposition 8 gives $d_\theta \le \kappa^{-1/2}\delta_\theta$ for the $L^2(\nu \times \mathcal{S})$ teacher-error scale δ_θ defined above.

For query covariates $X_1, \ldots, X_M$ with $Z_m = (1, X_m^\top)^\top$, define

$$\Gamma_M = \frac{1}{M}\sum_{m=1}^{M} Z_m Z_m^\top, \qquad R_M = \frac{1}{M}\sum_{m=1}^{M} Z_m U_\theta(X_m)^\top, \qquad N_{MK} = \frac{1}{M}\sum_{m=1}^{M} Z_m E_m^\top,$$

and let $\eta \in (0, 1)$ denote a confidence parameter and $\epsilon_{\mathrm{op}} \ge 0$ an optimization tolerance for the coefficient norm.

Theorem 1. *Suppose that Assumptions 1 and 3 hold and that* $\mathbb{E}_\nu\|Z\|_2^2 < \infty$. *Let* $\widehat{\beta}$ *be any computed estimator satisfying* $\|\widehat{\beta} - \widetilde{\beta}_{\theta,MK}\|_{L^2(\mathcal{S})} \le \epsilon_{\mathrm{op}}$ *for the empirical mean-distillation minimizer* $\widetilde{\beta}_{\theta,MK}$ *based on the* M *query covariates and* K *teacher samples per query. If, for some* $\underline{\kappa}_M > 0$ *and nonnegative numbers* ρ_M *and* ρ_K,

$$\lambda_{\min}(\Gamma_M) \ge \underline{\kappa}_M, \qquad \|R_M\|_F \le \rho_M, \qquad \|N_{MK}\|_F \le \rho_K,$$

then

$$\left\|\widehat{\beta} - \beta_{\mu_0}^\dagger\right\|_{L^2(\mathcal{S})} \le d_\theta + \rho_q + \underline{\kappa}_M^{-1}(\rho_M + \rho_K) + \epsilon_{\mathrm{op}}. \tag{19}$$

If Assumption 2 also holds and $M \ge C\{d_{\mathrm{eff}} + \log(1/\eta)\}$, *then, with probability at least* $1 - \eta$, *it holds that*

$$\left\|\widehat{\beta} - \beta_{\mu_0}^\dagger\right\|_{L^2(\mathcal{S})} \le d_\theta + \rho_q + C\kappa^{-1}\left\{\sigma_\theta\sqrt{\frac{\log(1/\eta) + d_{\mathrm{eff}}}{MK}} + B_\theta\sqrt{\frac{d_{\mathrm{eff}} + \log(1/\eta)}{M}}\right\} + \epsilon_{\mathrm{op}}. \tag{20}$$

Corollary 1 in the Supplement establishes the resulting consistency regime, and Theorem 2 gives a Gaussian minimax lower bound for the finite-basis coefficient-matrix oracle. We also identify the population targets of pure effect distillation and combined mean-effect distillation in Theorem 3 and Proposition 9, with all proofs given in Appendix G.

In conditional diffusion sampling, classifier-free guidance with weight $\omega > 0$ combines conditional and unconditional score estimates (Ho and Salimans, 2022). To quantify the effect of changing the teacher distribution on mean distillation, Proposition 10 in the Supplement analyzes an explicit density tilt of a Gaussian linear model with reference slope matrix B_0. Under that tilted density, the projected slope differs from B_0 whenever $B_0 \ne 0$ and has the relative approximation $(1 + \omega)B_0$ for weak signals. Even with exact scores, the output distribution of a guided diffusion sampler can differ from this tilted density (Bradley and Nakkiran, 2024; Chidambaram et al., 2024). For a fixed queried teacher distribution, the teacher bias d_θ in (20) is unchanged by increasing M or K. We query all teachers without classifier-free guidance in our experiments.

6 Experiments

We evaluate estimation of spatial effect maps and the error sources in (17).

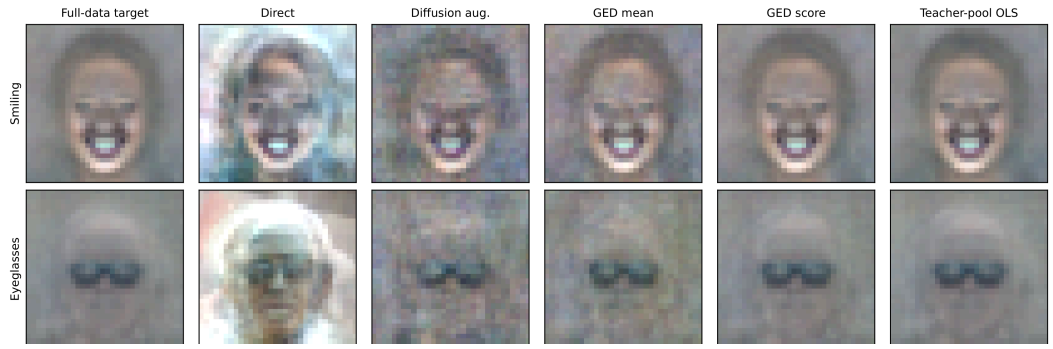

Figure 1: CelebA conditional residual diffusion effect maps for the attributes Smiling and Eyeglasses; the full four-attribute panel is in Figure 5. Columns: full-data target, direct OLS at $n = 80$, diffusion observed-covariate augmentation, mean-access GED via 30-step DDIM (Song, Meng, and Ermon, 2021), score-access GED via the single-step Tweedie denoising procedure described in Appendix I, and teacher-pool OLS. Score-access GED has SISE 0.009, compared with 0.005 for teacher-pool OLS and 0.229 for observed-covariate augmentation.

For a standardized Smiling covariate, the map in Figure 1 gives the fitted channel-wise image change per unit increase, holding the other attributes fixed. We use standardized integrated squared error (SISE), map correlation, and signed-region F1 to assess recovery of this regression target. For a nonlinear analytic teacher with a known projection target, mean, score, and effect GED reduce SISE from 0.012 for teacher augmentation to 0.004, 0.003, and 0.001, respectively. In agreement with the query-budget dependence in (20), SISE decreases from 0.044 to 0.006 as M grows from 50 to 400 at $K = 1$, and from 0.006 to 0.003 as K grows from 1 to 16 at $M = 400$. Across the eleven blocks in Table 1, GED has the lowest SISE among the direct-regression and augmentation baselines, with one-sided paired t-test p-values below 0.05 in every block. The largest reduction, greater than $25\times$, occurs for the CelebA (Liu et al., 2015) conditional residual DDPM with single-step Tweedie score access in Figure 1.

Table 1: Eleven empirical blocks. Best non-GED is the lower-SISE direct or augmentation method, GED is the best access mode, Ratio is computed from rounded SISE values, and Paired p is the one-sided paired t-test. Block parentheses give (n/M); Appendix F reports full rows and uncertainty.

Block	Direct	Best non-GED	GED	Ratio	Paired p
Controlled synthetic $(120/400)$	0.018	0.012	0.001	12.0	1.7×10^{-12}
Learned synthetic teacher $(120/800)$	0.012	0.007	0.002	3.5	1.3×10^{-2}
Semi-real digit teacher $(120/800)$	0.049	0.020	0.016	1.3	2.4×10^{-2}
SDXL-Turbo/FFHQ-256 $(12/512)$	20.018	13.092	0.818	16.0	8.0×10^{-3}
CelebA conditional diffusion $(80/768)$	1.424	0.229	0.009	25.4	1.1×10^{-5}
FFHQ measured attributes $(25/1200)$	0.281	0.079	0.021	3.8	1.4×10^{-5}
Real-image handwriting $(60/800)$	0.093	0.050	0.026	1.9	1.2×10^{-2}
Olivetti face morphology $(80/1200)$	0.039	0.017	0.006	2.8	2.3×10^{-4}
CIFAR-10 pixel diffusion $(160/900)$	0.049	0.014	0.011	1.3	1.0×10^{-3}
Latent-diffusion digit teacher $(120/800)$	0.043	0.034	0.015	2.3	2.0×10^{-5}
CelebA non-diffusion ridge $(80/2000)$	1.657	0.048	0.024	2.0	4.0×10^{-3}

Across these blocks, target-design queries improve map estimation when observed images are scarce and the teacher is sufficiently accurate. Direct regression approaches GED as the observed sample size approaches the teacher training pool size. Increasing M and K decreases the finite-query terms in Theorem 1. Teacher bias, spatial approximation, and changes in the target due to registration error require separate assessment.

References

Augustin, M., Boreiko, V., Croce, F., and Hein, M. (2022). Diffusion visual counterfactual explanations. In *Advances in Neural Information Processing Systems*.

Bradley, A. and Nakkiran, P. (2024). Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2408.09000*.

Berman, N., Naiman, I., Eliasof, M., Zisling, H., and Azencot, O. (2025). One-step offline distillation of diffusion-based models via Koopman modeling. In *Advances in Neural Information Processing Systems*.

Bucilă, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chidambaram, M., Gatmiry, K., Chen, S., Lee, H., and Lu, J. (2024). What does guidance do? A fine-grained analysis in a simple setting. *arXiv preprint arXiv:2409.13074*.

Ding, X., Wang, Y., Xu, Z., Welch, W. J., and Wang, Z. J. (2021). CcGAN: Continuous conditional generative adversarial networks for image generation. In *International Conference on Learning Representations*.

De Sousa Ribeiro, F., Xia, T., Monteiro, M., Pawlowski, N., and Glocker, B. (2023). High fidelity image counterfactuals with probabilistic causal models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 7390–7425.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614.

Ha, T. D. and Bender, S. (2025). Diffusion counterfactuals for image regressors. *arXiv preprint arXiv:2503.20595*.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*.

Ho, J. and Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Song, J., Meng, C., and Ermon, S. (2021). Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Ledoux, M., Nourdin, I., and Peccati, G. (2015). Stein's method, logarithmic Sobolev and transport inequalities. *Geometric and Functional Analysis*, 25:256–306.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Van Looveren, A., Klaise, J., Vacanti, G., and Cobb, O. (2021). Conditional generative models for counterfactual explanations. *arXiv preprint arXiv:2101.10123*.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Liu, S., Berman, W., Xu, Y., and Wolf, T. (2022). Diffusers: State-of-the-art diffusion models. GitHub repository. https://github.com/huggingface/diffusers.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. (2023). DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations*.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. (2023). SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Prabhu, V., Yenamandra, S., Chattopadhyay, P., and Hoffman, J. (2023). LANCE: Stress-testing visual models by generating language-guided counterfactual images. In *Advances in Neural Information Processing Systems*.

Samaria, F. S. and Harter, A. C. (1994). Parameterisation of a stochastic model for human face identification. In *Proceedings of the Second IEEE Workshop on Applications of Computer Vision*.

Sauer, A., Lorenz, D., Blattmann, A., and Rombach, R. (2023). Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Turan, E., Siozopoulos, A., Martinez, L., Gaubil, J., Pierson, E., and Ovsjanikov, M. (2026). Unfolding generative flows with Koopman operators: Trajectory-preserving linearization. In *Proceedings of the 43rd International Conference on Machine Learning*.

Wang, J., Wang, K., Yu, Y., et al. (2025). Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nature Medicine*, 31:609–617.

Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W. T., and Park, T. (2024). One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, D., Li, L., Sripada, C., and Kang, J. (2023). Image response regression via deep neural networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1589–1614.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). SmoothGrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Zhou, M., Zheng, H., Wang, Z., Yin, M., and Huang, H. (2024). Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Proceedings of the 41st International Conference on Machine Learning*.

Acknowledgments and Disclosure of Funding

Seyoung Park's work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT) (No. RS-2025-00517793). Eun Ryung Lee's work was supported by the NRF grants funded by the Korean government (MSIT) (Nos. 2022R1A2C1012798 and RS-2025-02216235).

Competing interests. The authors declare no competing interests.

A Notation glossary

Table 2: Notation used in the main manuscript.

Symbol	Meaning
$\mathcal{S}$	Spatial image domain
$X \in \mathbb{R}^p$	Covariate vector
$Y \in L^2(\mathcal{S})$	Image response, with the same notation used for finite-channel images
ν	Target covariate distribution for effect-map projection
$P_\theta(\cdot \mid x)$	Conditional image generator used as teacher
$\mu_\theta(x, s)$	Teacher conditional mean at covariate x and location s
$m_\phi(x, s)$	Interpretable image-response student mean
$\beta_j(s)$	Spatial effect map for covariate j
$\mathcal{A}$	Access operator for distribution, mean, gradient, or score information
$\mathcal{D}$	Discrepancy for teacher-to-student information matching
$\mathcal{T}_{\nu,\mathcal{D}}(P_\theta)$	Teacher-to-map functional returning spatial effect maps
$\beta^\star_\theta(s)$	Slope component of the augmented Moore–Penrose mean-projection representative for μ_θ; equal to (9) when Σ_X is nonsingular
$\beta^c_{\theta,\min}(s)$	Centered Moore–Penrose mean-projection slope representative
$\beta^c_{\text{mean}}(s)$	Alias for $\beta^c_{\theta,\min}(s)$ in the combined mean-effect target
$\beta^\dagger_{\mu_0}(s)$	Slope component of the augmented Moore–Penrose mean-projection representative of the true conditional mean μ_0 used as the finite-query reference target
$\beta^\star_j(s)$	Pure effect-loss target for covariate j
$\bar{g}_{\theta,j}(s)$	Average teacher derivative for covariate j under ν
M	Number of query covariates
K	Number of teacher samples per query covariate
C_{img}	Number of image channels in a finite-channel response
$Z = (1, X^\top)^\top$	Augmented covariate vector
$\bar{x}, \Sigma_X, \Gamma_X$	Design mean, centered covariance, and augmented covariance under ν
Φ, Π_{map}	Student parameter space and coefficient-map projection
T_ν	Stein kernel for ν when the stated Stein identity holds
ψ, q, d_{eff}	Spatial basis, basis size, and finite-query effective dimension
$V_\theta, \widehat{V}_\theta, E_m$	Teacher mean basis vector, empirical teacher mean basis vector, and teacher Monte Carlo error
$A^\star_q, U_\theta$	Population finite-basis coefficient matrix and projected residual
$\rho_q, B_\theta, \sigma_\theta$	Basis approximation error, projected residual scale, and teacher sampling scale
d_θ, δ_θ	Exact projected teacher bias and optional full teacher-mean error scale
Γ_M, R_M, N_{MK}	Empirical design matrix, exact-query residual, and teacher Monte Carlo residual
$\eta, \epsilon_{\text{op}}$	Probability tolerance and optimization tolerance
$\lambda_\mu, \lambda_g, \lambda_s$	Mean, effect, and score distillation weights
$C_{\phi,\tau}, W_\tau, r_{\phi,\tau}$	Score covariance, score weight operator, and teacher–student score residual

B Additional related work and algorithmic specification

Image-response regression and spatial effect maps. Image-response regression treats the image as the response and estimates coefficient functions that can be inspected spatially. Neural image-response regression extends this idea by representing coefficient maps with coordinate networks (Zhang et al., 2023). GED uses the same interpretable student structure, but changes the statistical target: the map is defined as a projection of a conditional image distribution $P_\theta(Y \mid X)$ and not as a direct fit to the observed images only. This distinction is central when the generator contains conditional variation, nonlinear covariate effects, and heteroskedastic image uncertainty that are not represented by a linear student.

Conditional generation and counterfactual images. GANs, conditional GANs, VAEs, normalizing flows, and diffusion models form flexible generative-model families that can be adapted to image generation (Goodfellow et al., 2014; Mirza and Osindero, 2014; Kingma and Welling, 2014; Rezende and Mohamed, 2015; Ho et al., 2020; Song et al., 2021). Continuous conditional GANs specifically address image generation under continuous labels (Ding et al., 2021), and counterfactual generators create image-level changes that alter a target attribute or model output (Van Looveren et al., 2021; Augustin et al., 2022; Ha and Bender, 2025). These methods provide conditional samples or counterfactual image pairs. Recent examples estimate image counterfactuals with deep structural causal models, generate medical images with a multimodal foundation model, or stress-test a classifier with language-guided counterfactual images (De Sousa Ribeiro et al., 2023; Wang et al., 2025; Prabhu et al., 2023). GED defines a design-indexed coefficient map. The interpreted quantity is a population projection and not an individual generated difference.

Why generated differences are not coefficient maps. For a nonlinear conditional generator, the difference between two generated images depends on the sampling seed, the baseline covariate value, the intervention path, and the distribution over nuisance covariates. Averaging such differences yields an estimand only after the target design and averaging rule are specified. GED makes this specification explicit through $\mathcal{T}_{\nu,\mathcal{D}}(P_\theta)$, and the reported map is a projection of teacher information under ν, not a visual contrast chosen after samples are generated.

Knowledge distillation. Classical distillation transfers predictive information from a large model to a smaller model (Bucilă et al., 2006; Hinton et al., 2015). Score distillation uses a pretrained diffusion model as a prior for optimizing a differentiable representation (Poole et al., 2023), and diffusion distillation transfers generative behavior or score information into a faster generator (Yin et al., 2024; Zhou et al., 2024). In contrast, GED uses distillation to define a projection estimand. Because the student is restricted to an interpretable regression class, distillation projects the teacher's distributional, gradient, or score information onto a spatial coefficient map. Related methods such as Koopman Distillation Models and Unfolding Generative Flows learn latent linear operators for one-step sampling and spectral trajectory analysis (Berman et al., 2025; Turan et al., 2026). Although such a generator can supply teacher queries to GED, its latent trajectory operator differs from the spatial coefficient map targeted by GED.

Algorithmic specification. For a chosen target design ν, discrepancy $\mathcal{D}$, and student class Φ, GED consists of the following steps. First, fit or import a conditional teacher $P_\theta(Y \mid X)$ and evaluate its held-out calibration on observed images. Second, draw query covariates $x_1, \ldots, x_M$ from ν or from a balanced design chosen before estimation. Third, obtain teacher information at each query point, namely sample averages for mean distillation, automatic differentiation gradients for effect distillation, or conditional score values for score distillation. Fourth, solve the student optimization problem in (16). Fifth, report $\hat{\beta}_j(s)$, ν, $\mathcal{D}$, teacher held-out error, query sizes M and K, and uncertainty diagnostics.

C Auxiliary methodological results

This appendix collects auxiliary propositions and a corollary referenced in the main manuscript. Their proofs are given in Appendix G.

C.1 Centered slope representation for the mean projection

The following proposition is a centered counterpart to Proposition 2, convenient when Σ_X is singular.

Proposition 5. *Let $\bar{x} = \mathbb{E}_\nu X$, let $\Sigma_X = \mathbb{E}_\nu\{(X-\bar{x})(X-\bar{x})^\top\}$, and assume that $\mathbb{E}_\nu\|X\|_2^2 < \infty$ and $\mu_\theta \in L^2(\nu \times \mathcal{S})$. Consider the centered linear class*

$$m_{c,\beta}(x,s) = c(s) + (x-\bar{x})^\top \beta(s).$$

Put

$$h_c(s) = \mathbb{E}_\nu\{(X-\bar{x})\mu_\theta(X,s)\}.$$

Then, the minimizer set over $(c,\beta) \in L^2(\mathcal{S}) \times L^2(\mathcal{S};\mathbb{R}^p)$ is nonempty. Every minimizer satisfies

$$c(s) = \mathbb{E}_\nu\{\mu_\theta(X,s)\}, \qquad \Sigma_X\beta(s) = h_c(s) \quad \textit{for almost every } s \in \mathcal{S}.$$

The fitted projection is unique in $L^2(\nu \times \mathcal{S})$, and the minimum-norm centered slope representative is

$$\beta^c_{\theta,\min}(s) = \Sigma_X^\dagger h_c(s).$$

If two slope representatives β_1 and β_2 solve the centered normal equations, then $\beta_1(s) - \beta_2(s) \in \ker(\Sigma_X)$ for almost every $s \in \mathcal{S}$, and

$$(X-\bar{x})^\top\{\beta_1(s) - \beta_2(s)\} = 0 \quad \textit{for } \nu \times \mathrm{d}s\textit{-almost every } (X,s).$$

C.2 Pathwise gradient interchange

When gradients are obtained by automatic differentiation through a conditional sampler, the following condition identifies the derivative of the teacher conditional mean and not a sampler-specific pathwise derivative. For $j = 1,\dots,p$, let e_j denote the jth Euclidean basis vector in $\mathbb{R}^p$.

Proposition 6. *Let ζ denote the teacher latent randomness and suppose that, for x in an open set $\mathcal{X}_0 \subseteq \mathcal{X}$,*

$$Y_\theta(x) = G_\theta(x,\zeta) \in L^2(\mathcal{S}), \qquad \mu_\theta(x) = \mathbb{E}\{G_\theta(x,\zeta)\}.$$

Fix $j \in \{1,\dots,p\}$ and $x \in \mathcal{X}_0$. Assume that there exists $\delta > 0$ such that $x + te_j \in \mathcal{X}_0$ for $|t| < \delta$, and, for ζ-almost every ζ, $u \mapsto G_\theta(u,\zeta)$ is absolutely continuous on this line and differentiable at x and almost everywhere in the jth coordinate as an $L^2(\mathcal{S})$-valued map. Assume also that a measurable almost-everywhere version of the coordinate derivative satisfies

$$\mathbb{E}\left[\operatorname*{ess\,sup}_{|t|<\delta}\left\|\frac{\partial}{\partial x_j}G_\theta(x+te_j,\zeta)\right\|_{L^2(\mathcal{S})}\right] < \infty.$$

Then μ_θ is differentiable in the jth coordinate at x as an $L^2(\mathcal{S})$-valued map and

$$\frac{\partial}{\partial x_j}\mu_\theta(x) = \mathbb{E}\left\{\frac{\partial}{\partial x_j}G_\theta(x,\zeta)\right\} \quad \textit{in } L^2(\mathcal{S}).$$

Remark 2. *Without this interchange condition, effect distillation based on backpropagated sampler derivatives targets the declared pathwise access functional and not a distribution-intrinsic derivative of the teacher conditional mean.*

C.3 Vector form of score-residual orthogonality

The next proposition gives a vector form of (15) under a full score-space tangent class.

Proposition 7. *Assume the conditions of Proposition 4, and suppose that (15) holds at ϕ. Fix $j \in \{0,\dots,p\}$ and suppose that $\mathcal{H}_j = L^2(\mathcal{S})$. If, under the score-space convention above, $C_{\phi,\tau}^{-1}$ and W_τ are bounded self-adjoint operators on $L^2(\mathcal{S})$, $\mathbb{E}Z_j^2 < \infty$, and*

$$\mathbb{E}\left[|Z_j|\left\|C_{\phi,\tau}^{-1}W_\tau r_{\phi,\tau}(\tilde{Y},X)\right\|_{L^2(\mathcal{S})}\right] < \infty,$$

then

$$\mathbb{E}\left[Z_j C_{\phi,\tau}^{-1}W_\tau r_{\phi,\tau}(\tilde{Y},X)\right] = 0 \quad \textit{in } L^2(\mathcal{S}). \tag{21}$$

After choosing representatives, (21) holds for almost every $s \in \mathcal{S}$. If $\mathcal{H}_j = L^2(\mathcal{S})$ for all $j = 0,\dots,p$ and the objective is finite for every finite perturbation of the affine mean class, (15) is also sufficient for global minimality over the mean parameters with fixed $R_{\theta,\tau}$, $C_{\phi,\tau}$, and W_τ.

C.4 Stability of the mean teacher-to-map functional

For two square-integrable teacher means μ_a and μ_b, let $m_a^\star$ and $m_b^\star$ denote their fitted linear projections in $L^2(\nu \times \mathcal{S})$, let $b_{a,\min}^\star$ and $b_{b,\min}^\star$ denote the corresponding minimum-norm coefficient representatives, and put

$$\Delta_{ab}^2 = \mathbb{E}_\nu \int_{\mathcal{S}} \{\mu_a(X,s) - \mu_b(X,s)\}^2 \, \mathrm{d}s.$$

Proposition 8. *Assume that $\mathbb{E}_\nu \|Z\|_2^2 < \infty$. Let $\Gamma_X = \mathbb{E}_\nu(ZZ^\top)$ and let κ_+ denote its smallest positive eigenvalue. Then, it holds that*

$$\|m_a^\star - m_b^\star\|_{L^2(\nu\times\mathcal{S})} \le \Delta_{ab}, \qquad \left\|b_{a,\min}^\star - b_{b,\min}^\star\right\|_{L^2(\mathcal{S})} \le \kappa_+^{-1/2}\Delta_{ab}.$$

If $\lambda_{\min}(\Gamma_X) \ge \kappa > 0$, the coefficient representatives are unique. Writing $b_a^\star = b_{a,\min}^\star$, $b_b^\star = b_{b,\min}^\star$, and $\beta_a^\star, \beta_b^\star$ for the corresponding slope components, it holds that

$$\|\beta_a^\star - \beta_b^\star\|_{L^2(\mathcal{S})} \le \|b_a^\star - b_b^\star\|_{L^2(\mathcal{S})} \le \kappa^{-1/2}\Delta_{ab}.$$

Proposition 8 implies that the exact bias term d_θ is bounded by $\kappa^{-1/2}\delta_\theta$ whenever $\lambda_{\min}(\Gamma_X) \ge \kappa > 0$ and the stronger $L^2(\nu \times \mathcal{S})$ teacher-error scale δ_θ is used.

C.5 Consistency corollary

The following corollary records the consistency regime implied by Theorem 1.

Corollary 1. *Suppose that the high-probability conditions of Theorem 1 hold along a sequence of problems with $d_{\mathrm{eff}}/M \to 0$, $d_\theta \to 0$, $\rho_q \to 0$, $\epsilon_{\mathrm{op}} \to 0$,*

$$\kappa^{-1}B_\theta\sqrt{\frac{d_{\mathrm{eff}}}{M}} \to 0, \qquad \kappa^{-1}\sigma_\theta\sqrt{\frac{d_{\mathrm{eff}}}{MK}} \to 0.$$

Then, it holds that

$$\left\|\widehat{\beta} - \beta_{\mu_0}^\dagger\right\|_{L^2(\mathcal{S})} \to 0$$

with probability tending to one. If the teacher bias is controlled through δ_θ instead of d_θ, it is sufficient to replace $d_\theta \to 0$ by $\kappa^{-1/2}\delta_\theta \to 0$.

C.6 Minimax lower bound for finite-query stochastic terms

The following theorem establishes that the query-design and Monte Carlo dependencies in Theorem 1 are unavoidable for the augmented finite-basis coefficient-matrix oracle risk in a Gaussian submodel. Let $r = p + 1$ and recall that $d_{\mathrm{eff}} = rq$. Consider fixed query vectors $Z_1, \dots, Z_M \in \mathbb{R}^r$ and define

$$\Gamma_M = \frac{1}{M}\sum_{m=1}^M Z_m Z_m^\top.$$

Let $\psi(s) = (\psi_1(s), \dots, \psi_q(s))^\top$ be the orthonormal spatial basis in Assumption 1. For any $A' = (a_0', \dots, a_p')^\top \in \mathbb{R}^{r\times q}$, define $b_{A',j}(s) = a_j'^\top \psi(s)$, $j = 0, \dots, p$. Then the orthonormality of ψ gives, for every $\widehat{A} \in \mathbb{R}^{r\times q}$ and $A \in \mathbb{R}^{r\times q}$,

$$\sum_{j=0}^p \|b_{\widehat{A},j} - b_{A,j}\|_{L^2(\mathcal{S})}^2 = \|\widehat{A} - A\|_F^2.$$

Let $B_\theta, \sigma_\theta \ge 0$, and suppose that the projected teacher response at query point m is

$$V_m = A^\top Z_m + \xi_m + e_m, \qquad m = 1, \dots, M, \tag{22}$$

where $\xi_m \sim N(0, B_\theta^2 I_q)$ independently over m, $e_m = K^{-1}\sum_{k=1}^K \varepsilon_{mk}$, $\varepsilon_{mk} \sim N(0, \sigma_\theta^2 I_q)$ independently over (m,k), and all ξ_m and ε_{mk} are mutually independent. Put

$$\tau_K^2 = B_\theta^2 + \frac{\sigma_\theta^2}{K}.$$

Let $\mathbb{P}_A$ and $\mathbb{E}_A$ denote probability and expectation under parameter A, define $\mathbb{B}_F(R) = \{A \in \mathbb{R}^{r\times q} : \|A\|_F \le R\}$, and let $\widehat{A}$ be any measurable function of $Z_1, \dots, Z_M, V_1, \dots, V_M$. Let $\overline{\kappa}$ be a finite positive design constant.

Theorem 2. *Suppose that the Gaussian finite-basis experiment in (22) holds, $\tau_K > 0$, Γ_M is positive definite, and $\lambda_{\max}(\Gamma_M) \le \overline{\kappa}$. If*

$$R^2 \ge \frac{q\ \mathrm{tr}(\Gamma_M^{-1})}{32M}\left(B_\theta^2 + \frac{\sigma_\theta^2}{K}\right),$$

then, it holds that

$$\inf_{\widehat{A}} \sup_{A\in\mathbb{B}_F(R)} \mathbb{E}_A\|\widehat{A} - A\|_F^2 \ge \frac{q\ \mathrm{tr}(\Gamma_M^{-1})}{128M}\left(B_\theta^2 + \frac{\sigma_\theta^2}{K}\right).$$

This implies that

$$\inf_{\widehat{A}} \sup_{A\in\mathbb{B}_F(R)} \mathbb{E}_A\|\widehat{A} - A\|_F^2 \ge \frac{d_{\mathrm{eff}}}{128\,\overline{\kappa}}\left(\frac{B_\theta^2}{M} + \frac{\sigma_\theta^2}{MK}\right).$$

The construction in Theorem 2 has $d_\theta = 0$, $\rho_q = 0$, and $\epsilon_{\mathrm{op}} = 0$, because the reference mean equals the teacher mean, the target coefficient matrix belongs to the finite spatial basis, and the estimator is compared with the oracle finite-dimensional parameter A. Thus, the lower bound applies to the augmented coefficient matrix, including the intercept. It gives a limit on improving the query-design and teacher-sampling dependencies in this oracle coefficient experiment. The slope-only loss in (20) is a different target. The projected residual component ξ_m gives the $B_\theta^2 d_{\mathrm{eff}}/M$ term, whereas the averaged teacher-sampling component e_m gives the $\sigma_\theta^2 d_{\mathrm{eff}}/(MK)$ term. In any design regime with $\overline{\kappa} = O(1)$, the augmented squared-risk dependence on M, K, q, and p matches the corresponding upper-bound stochastic terms up to absolute constants.

C.7 Pure effect-loss target

The following theorem identifies the population target of pure effect-gradient distillation, complementing the projection identity in Proposition 2.

Theorem 3. *Suppose that the linear student in (2) is trained only with the population effect loss (10) over the full coefficient class $\beta_0 \in L^2(\mathcal{S})$ and $\beta_j \in L^2(\mathcal{S})$, $j = 1,\dots,p$, and that $\partial\mu_\theta(X,s)/\partial x_j \in L^2(\nu\times\mathcal{S})$ for $j = 1,\dots,p$. Then every minimizer satisfies, for almost every $s \in \mathcal{S}$,*

$$\beta_j^\star(s) = \mathbb{E}_\nu\{\partial\mu_\theta(X,s)/\partial x_j\}, \qquad j = 1,\dots,p,$$

and any linear student matching these slope equalities is a minimizer.

C.8 Combined mean and effect target

The distinction between mean and effect projections can be made exact for a linear student trained with both losses.

Proposition 9. *For the linear student in (2), assume that the population objective is minimized over the full coefficient class $\beta_0 \in L^2(\mathcal{S})$ and $\beta \in L^2(\mathcal{S};\mathbb{R}^p)$, assume that $\lambda_\mu, \lambda_g \ge 0$, and assume that they are not both zero. If $\lambda_\mu = 0$ and $\lambda_g > 0$, assume that $\partial\mu_\theta(X,s)/\partial x_j \in L^2(\nu\times\mathcal{S})$ for $j = 1,\dots,p$, and let $\bar{g}_\theta(s) = (\bar{g}_{\theta,1}(s),\dots,\bar{g}_{\theta,p}(s))^\top$ with $\bar{g}_{\theta,j}(s) = \mathbb{E}_\nu\{\partial\mu_\theta(X,s)/\partial x_j\}$. Then the slope component of any population minimizer of $\lambda_g\mathcal{L}_{\mathrm{eff}}(\phi)$ is $\bar{g}_\theta(s)$ for almost every $s \in \mathcal{S}$. If $\lambda_\mu > 0$ and $\lambda_g = 0$, assume that $\mathbb{E}_\nu\|X\|_2^2 < \infty$ and $\mu_\theta \in L^2(\nu\times\mathcal{S})$. Then the mean-loss minimizers have the unique fitted projection characterized in Proposition 5; the reported centered minimum-norm slope representative is $\beta^c_{\mathrm{mean}}(s) = \beta^c_{\theta,\min}(s)$, and it is the unique slope component when Σ_X is nonsingular. If $\lambda_\mu > 0$ and $\lambda_g > 0$, assume that $\mathbb{E}_\nu\|X\|_2^2 < \infty$, $\mu_\theta \in L^2(\nu\times\mathcal{S})$, and $\partial\mu_\theta(X,s)/\partial x_j \in L^2(\nu\times\mathcal{S})$ for $j = 1,\dots,p$, and define β^c_{mean} and $\bar{g}_\theta$ as above. Then $\lambda_\mu\Sigma_X + \lambda_g I_p$ is positive definite, and the slope component of the population minimizer of*

$$\lambda_\mu\mathbb{E}_\nu\int_{\mathcal{S}}\{m_\phi(X,s) - \mu_\theta(X,s)\}^2\,\mathrm{d}s + \lambda_g\mathcal{L}_{\mathrm{eff}}(\phi)$$

is

$$\beta_\lambda(s) = (\lambda_\mu\Sigma_X + \lambda_g I_p)^{-1}\{\lambda_\mu\Sigma_X\beta^c_{\mathrm{mean}}(s) + \lambda_g\bar{g}_\theta(s)\} \quad \text{for almost every } s \in \mathcal{S}.$$

If, in the case $\lambda_\mu > 0$ and $\lambda_g > 0$, the Stein identity in Proposition 3 gives $\beta^c_{\mathrm{mean}}(s) = \bar{g}_\theta(s)$, then $\beta_\lambda(s) = \beta^c_{\mathrm{mean}}(s) = \bar{g}_\theta(s)$ for all positive effect weights.

When Σ_X is nonsingular, β^c_{mean} is the unique mean-projection slope in (9). When Σ_X is singular and $\lambda_g = 0$, the fitted mean projection remains unique and β^c_{mean} is the reporting convention selected by Proposition 5. In the combined case with $\lambda_g > 0$, the matrix $\lambda_\mu \Sigma_X + \lambda_g I_p$ is positive definite even if Σ_X is singular, which identifies the combined target.

Convexity of the fixed-covariance access objective. On a finite spatial discretization, fix the score covariance $C_{\phi,\tau}$, weight W_τ, and replay distribution as in Proposition 4, and omit nonconvex regularization. Writing $Z = (1, X^\top)^\top$, the mean and score losses are convex quadratics in the linear coefficient maps; the score Hessian is $2\mathbb{E}_\nu(ZZ^\top) \otimes C^{-1}_{\phi,\tau} W_\tau C^{-1}_{\phi,\tau} \succeq 0$. The effect loss adds $2\lambda_g I$ on the slope coordinates. Thus, for the weighted access-loss sum, $\lambda_g > 0$ gives a unique global slope minimizer, and gradient descent with step size $0 < \eta < 2/L$ converges to it from every initialization, where L is the largest Hessian eigenvalue. The component gradients may change the weighted estimand but do not create a spurious local minimum in this setting. This conclusion does not cover jointly trained covariance, nonlinear student parameterizations, or nonconvex regularizers.

C.9 Classifier-free guidance and teacher bias

To quantify the mean change associated with the density tilt suggested by classifier-free guidance (Ho and Salimans, 2022), we study a Gaussian linear reference model. For $\omega \geq 0$, we explicitly define the tilted density proportional to $p(y \mid x)^{1+\omega} p(y)^{-\omega}$, where $p(y \mid x)$ and $p(y)$ are the conditional and marginal densities of the same joint model. This density is not in general the output distribution of a guided diffusion sampler, even with exact scores (Bradley and Nakkiran, 2024; Chidambaram et al., 2024).

Proposition 10. *Let $Y \in \mathbb{R}^q$ denote an image on q voxels, identified with a piecewise-constant element of $L^2(\mathcal{S})$. Assume that $X \sim N(0, \Psi)$ and $Y \mid X = x \sim N(B_0 x, \Sigma_\varepsilon)$, where $B_0 \in \mathbb{R}^{q \times p}$, $\Psi \succ 0$, and $\Sigma_\varepsilon \succ 0$, and let $p(y \mid x)$ and $p(y)$ denote the conditional and marginal densities of Y. For $\omega \geq 0$, let $p_\omega(y \mid x) \propto p(y \mid x)^{1+\omega} p(y)^{-\omega}$, and define $G_0 = B_0^\top \Sigma_\varepsilon^{-1} B_0$, $V_0 = (\Psi^{-1} + G_0)^{-1}$, and $H_0 = V_0 G_0$. Then the eigenvalues of H_0 lie in $[0, 1)$, and $p_\omega(\cdot \mid x)$ is a Gaussian density with mean $B_\omega x$, where*

$$B_\omega = (1+\omega) B_0 (I_p + \omega H_0)^{-1}, \qquad B_\omega - B_0 = \omega B_0 (I_p + \omega H_0)^{-1} V_0 \Psi^{-1}. \tag{23}$$

In particular, if $\omega > 0$, then $B_\omega = B_0$ if and only if $B_0 = 0$.

To describe the size of the slope change, let $\|\cdot\|_F$ denote the Frobenius norm and assume $B_0 \neq 0$. For fixed ω, the relative error $\|\{B_\omega - (1+\omega)B_0\}\Psi^{1/2}\|_F / \|B_0 \Psi^{1/2}\|_F$ tends to zero as the largest eigenvalue of $\Psi^{1/2} G_0 \Psi^{1/2}$ tends to zero. If the smallest eigenvalue of this matrix tends to infinity, then $\|(B_\omega - B_0)\Psi^{1/2}\|_F / \|B_0 \Psi^{1/2}\|_F$ tends to zero. The second limit requires strong signal in every covariate direction and concerns relative error; it does not imply that the absolute slope difference tends to zero without further bounds on the coefficients and noise covariance. When H_0 is not diagonal and $\omega > 0$, the matrix $(I_p + \omega H_0)^{-1}$ is not diagonal. Right multiplication by this matrix may combine nonzero reference slope maps, but non-diagonality alone does not imply such mixing for every covariate. When a reference slope map is identically zero, it remains zero under this Gaussian tilt.

To connect the tilted density with Theorem 1, suppose that the queried teacher has conditional density $p_\omega(y \mid x)$ and the reference mean is $\mu_0(x) = B_0 x$. For any design ν with $\mathbb{E}_\nu \|X\|_2^2 < \infty$ and nonsingular Σ_X, the unrestricted affine mean projections have zero intercept maps and slope maps given by the columns of B_0 and B_ω. On the fixed voxel partition with positive voxel volumes, d_θ is the $L^2(\mathcal{S}; \mathbb{R}^p)$ norm of the slope-map difference formed by the columns of $B_\omega - B_0$. It is positive when $\omega > 0$ and $B_0 \neq 0$ and is unchanged by increasing M or K while the teacher, reference, design, and guidance weight remain fixed. These conclusions concern sampling from p_ω. Even in the continuous-time limit with exact scores, guided diffusion samplers need not sample from this density (Bradley and Nakkiran, 2024; Chidambaram et al., 2024), and their projected bias is determined by their own conditional means. For these reasons, we query every teacher without guidance in our experiments. In particular, the SDXL-Turbo experiment sets the guidance scale to 0, which disables classifier-free guidance in the `diffusers` implementation.

D Additional positioning

Table 3: Positioning relative to related directions. GED changes the object of distillation from samples or fast generators to a design-indexed statistical functional of a conditional image distribution.

Direction	Object learned	Limitation for effect-map distillation
Image-response regression	Coefficient maps fitted from observed images under an explicit regression model	Does not use a learned conditional image distribution
Conditional image generation	Conditional distribution $P_\theta(Y \mid X)$ with sample-level covariate changes	Does not define a population coefficient map
Generative counterfactuals	Image pairs that alter a predictor or response	Provides sample-specific explanations, not a design-indexed estimand
Generator distillation	Fast student generator matching a teacher in distribution or samples	Does not restrict the student to an interpretable regression family
GED	$\mathcal{T}_{\nu,\mathcal{D}}(P_\theta)$: projection, average-gradient, or score-orthogonal spatial maps	Separates teacher bias, query-design error, Monte Carlo error, and optimization error

E Additional limitations

Uncertainty and calibration. GED separates teacher uncertainty, query uncertainty, and student uncertainty. These components can be assessed by teacher refitting or ensembles, repeated teacher calls and query-design resampling, and spatial or wild bootstrap methods for the student. The experiments report held-out teacher prediction error and standard errors across independent repetitions. Scientific applications should also report sensitivity to ν and $\mathcal{D}$, region-level stability, and simultaneous spatial uncertainty diagnostics. Absent a causal design, the maps are conditional association summaries and not causal effects.

The controlled calibration study in Table 23 separates these diagnostics. Query agreement measures fidelity to the fixed teacher, held-out prediction error diagnoses conditional-mean disagreement, and proper scores or predictive coverage diagnose scale calibration. None of these internal checks alone proves scientific validity.

Spatial alignment and resolution. A coefficient $\beta_j(s)$ compares the same coordinate s across images and hence presumes a common coordinate system. In the rigid-translation study in Table 26, GED continues to recover the projection of the misaligned image distribution, but that projection becomes an attenuated version of the aligned target. Random image-to-image displacement also increases teacher-sampling variability. Resampling plausible registrations can quantify map or region variation, but averaging across registrations targets a deformation-averaged map. Nonrigid, covariate-dependent, and anatomy-dependent deformations require a separate registration or latent deformation model.

Higher spatial resolution also requires a larger basis dimension q, which increases $d_{\text{eff}} = (p+1)q$ and the query requirement in Theorem 1. Downsampling or a coarse basis removes fine structure before coefficient estimation. The frequency study in Table 24 separates this spatial approximation from linearity in the covariates, but it does not establish native-resolution recovery. SDXL-Turbo outputs are generated at 512×512 and represented by 64×64 coefficient arrays; the CelebA maps use 32×32 arrays. Stable recovery of these projections does not certify fine structural effects at the native image resolution.

GED maps are summaries of a teacher conditional distribution and should not be interpreted as causal effects unless the covariates, sampling design, and identification assumptions support causal interpretation. In addition, teacher misspecification, acquisition artifacts, demographic imbalance, and spurious correlations in P_θ can appear in the distilled maps. The teacher-bias term in (20) separates this source of error from query and optimization error, but additional teacher queries cannot remove bias in P_θ.

A single global $\beta_j(s)$ is an average summary of a nonlinear conditional distribution under the target design ν. When effects vary substantially over covariate space, local designs, interaction bases, or

basis-indexed maps are appropriate. Score distillation also requires calibrated noised teacher scores and a student covariance class containing the intended projection. The score access mode is evaluated against an analytic teacher with a known projection target in the controlled synthetic block, where GED score attains standardized integrated squared error of 0.003 (Table 5), and against the trained CelebA conditional residual DDPM through the single-step Tweedie denoising procedure described in Appendix I, where score-access GED attains SISE 0.009, map correlation 0.996, and signed region F1 0.936 over 15 score-access repetitions, compared with 0.090, 0.955, and 0.769 for the 30-step mean-access GED row of Table 12. The real-image analyses use learned conditional mean teachers and full-data OLS association maps as references. These references are not external ground truth, and the pretrained SDXL-Turbo experiment uses a teacher-defined FFHQ-256 face-image target and not an external causal or scientific target. The empirical conclusion is limited to this regime: GED reduces SISE relative to direct image-response regression when observed images are scarce and a calibrated teacher is available, whereas direct regression matches GED once the observed sample size approaches the teacher-training pool size.

The full-data OLS reference shares estimation noise with the direct-regression baseline, which makes the comparison in the real-image blocks a sample-efficiency comparison at fixed reference and not a recovery comparison against external truth. Sample-size sensitivity in Tables 17–19 isolates the effect of n on direct regression while holding the reference fixed, and the controlled synthetic and finite-query scaling tables explicitly use a teacher with a known projection target. Recovery is measured against an external quantity. The CIFAR-10 conditional pixel-diffusion and latent-diffusion blocks supplement the face-image analyses with a different image domain and a different conditional sampler, and the CelebA non-diffusion ridge block isolates the role of the diffusion sampling mechanism from the design-distribution effect.

Scientific use and external validation. Coefficient maps can locate covariate-associated regions, compare prespecified populations or target designs, and identify teacher–reference discrepancies for independent investigation. Table 27 gives a controlled example in which the difference between a biased-teacher GED map and an independent association map localizes an injected spatial artifact. This use requires an external reference population, known ground truth, or expert assessment. A GED map alone cannot establish that an association is spurious, causal, clinically useful, or faithful to the teacher's internal mechanism, and the present experiments do not test gains in prediction or diagnosis.

Failure modes and diagnostics. The decomposition in (20) identifies three failure modes. First, if P_θ is biased, the distilled map targets the biased teacher projection, and increasing M or K only estimates that biased projection more accurately. Second, if the query design ν places mass on covariate regions where the teacher is poorly calibrated, the map can be stable as a teacher functional while differing from the association summary for the observed population. Third, if the student class is too coarse, the basis approximation term ρ_q can dominate even when the teacher is accurate. We report held-out teacher error, sensitivity to ν, query sizes M and K, and comparison to direct regression whenever an observed-image reference is available.

F Additional empirical results

Empirical protocol. Table 4 summarizes the correspondence between the empirical study and the claims in the main manuscript. We report standardized integrated squared error, map correlation, and signed-region F1 when a reference map is available. For an estimated map $\widehat{\beta}$ and a reference map $\beta^\star$, standardized integrated squared error is

$$\mathrm{SISE}(\widehat{\beta}, \beta^\star) = \frac{\sum_{j=1}^{p} \|\widehat{\beta}_j - \beta_j^\star\|_{L^2(\mathcal{S})}^2}{\sum_{j=1}^{p} \|\beta_j^\star\|_{L^2(\mathcal{S})}^2}.$$

For finite-channel maps, the $L^2(\mathcal{S})$ norm uses the convention in Section 2. Map correlation is the cosine similarity between the vectorized estimated and reference maps over all covariates, pixels, and channels. The high-frequency and localized teacher-bias studies use Pearson correlation after centering these vectors. For signed-region F1, write $\beta^\star_{j,c}(s)$ and $\widehat{\beta}_{j,c}(s)$ for channel $c \in \{1, \dots, C_{\mathrm{img}}\}$ of covariate j, with $C_{\mathrm{img}} = 1$ in scalar responses. For each covariate j, set

$$t_j = 0.35 \max_{u \in \mathcal{S},\ 1 \le c' \le C_{\mathrm{img}}} |\beta^\star_{j,c'}(u)|.$$

The reference and predicted active sets are

$$A_j^\star = \{(s,c) : |\beta_{j,c}^\star(s)| \geq t_j\}, \qquad \widehat{A}_j = \{(s,c) : |\widehat{\beta}_{j,c}(s)| \geq t_j\}.$$

We use the following convention for the reported signed-region F1. A true positive is an entry in $\widehat{A}_j \cap A_j^\star$ with the correct sign. A false positive is an entry in $\widehat{A}_j$ outside $A_j^\star$, and a false negative is an entry in $A_j^\star$ that is not selected with the correct sign. Thus, a wrong-signed selection within the reference active set contributes to the false-negative count. We calculate F1 separately for each covariate and average the scores. In the localized teacher-bias study, precision and recall are also averaged over covariates using these counts. The SDXL-Turbo study uses a pooled version: twice the total number of true positives divided by the sum of the reference and predicted active-set sizes over all covariates, pixels, and channels. In this version, a wrong-signed selection contributes to both the false-positive and false-negative counts. We compare F1 within each study. Because t_j is fixed by the reference map, both versions penalize sign and location errors and shrinkage below t_j. All plus-minus values are standard errors across independent repetitions, computed as the sample standard deviation over repetitions divided by the square root of the number of repetitions. The held-out teacher RMSEs, evaluated outside the query sets used for distillation, are 0.043 ± 0.001 for the learned synthetic teacher, 0.035 ± 0.001 for the semi-real digit teacher, 0.216 ± 0.004 for the FFHQ measured-attribute teacher, 0.148 ± 0.003 for the handwritten morphology teacher, and 0.100 ± 0.001 for the face morphology teacher. For the diffusion-teacher experiments, the residual root mean squared scales around the fitted teacher projection are 0.213 for SDXL-Turbo/FFHQ-256 and 0.191 ± 0.000 for CelebA, because these teachers define stochastic sampling distributions and no external conditional mean is observed. The controlled synthetic and transformed digit images are 32×32, SDXL-Turbo image-to-image outputs are generated at 512×512 from FFHQ-256 base faces and downsampled to 64×64 RGB maps for coefficient estimation, the CelebA and FFHQ measured-attribute images are downsampled to 32×32 RGB maps, the raw scikit-learn digit images are 8×8, and the Olivetti face images are 64×64 in the loader version used here.

For the real-image blocks, the reference target is the OLS association map fitted to all available images in each dataset. We evaluate the accuracy of recovering this reference association map from a small sample of observed images. It does not turn measured image covariates into causal interventions or remove bias present in a learned teacher. Recovery of an external nonlinear projection target is evaluated separately in the synthetic, semi-real, and diffusion-teacher blocks.

Table 4: Empirical claim-to-evidence map. Each block isolates a different term in (20) or a distinct failure mode of ordinary augmentation and direct image-response regression.

Empirical block	Setting	Claim tested
Controlled synthetic teacher	Known nonlinear conditional mean on 32×32 images, with known projection target and 80 repetitions	GED recovers the projection target and separates mean, score, and gradient access
Finite-query scaling	Query sizes $M \in \{50, 100, 200, 400\}$ and teacher samples $K \in \{1, 4, 16\}$	The error decreases with both query design size and teacher Monte Carlo size, as predicted by Theorem 1
Bias and design sensitivity	Biased teacher perturbations $\rho \in \{0, .25, .50, .75\}$ and alternate target designs	GED estimates the teacher-defined and design-indexed projection and not an external target
Learned synthetic teacher	Neural teacher trained on 1000 noisy images, with direct regression restricted to 120 observed images	GED remains effective when teacher information is learned, not analytic
Semi-real digit teacher	Transformed digit-three images with controlled covariates and an oracle mean target	GED improves spatial map recovery under realistic image variation and a learned teacher
Pretrained SDXL-Turbo face teacher	SDXL-Turbo image-to-image teacher applied to 100 FFHQ-256 base faces under prompt-coded scalar covariates, with reference and GED query sizes 512 and direct observed sample size 12	GED improves recovery of the teacher-defined projection over direct regression and observed-covariate augmentation under a pretrained image generator
CelebA semantic diffusion teacher	Conditional residual DDPM trained on 1800 aligned CelebA face images with official attributes Smiling, Eyeglasses, Male, and Blond Hair; full-data association reference over 2000 images; direct observed sample size 80; mean-access and score-access GED rows	GED improves sparse semantic-attribute association-map recovery under an official-attribute conditional diffusion teacher, and score-access GED via single-step Tweedie outperforms mean-access GED at a fraction of the conditional-mean sampling cost
FFHQ measured attributes	FFHQ-256 face images, four image-measured covariates, full-data association reference, and direct observed sample size 25	GED improves measured-attribute association-map recovery without using prompt-coded covariates
Real-image handwriting	Original digit-three images, measured covariates, full-data association reference, and observed-sample scaling	GED reduces SISE for sparse observed-image map estimation and identifies the regime where direct regression attains comparable error
Real-image face morphology	Olivetti face images, three measured covariates, full-data association reference, and observed-sample scaling	GED gives the same sparse observed-image conclusion in a second real-image domain
CIFAR-10 conditional pixel diffusion	Conditional pixel-space DDPM trained on 6000 CIFAR-10 images with two-dimensional Gaussian covariates; direct sample size $n = 160$	GED scales to a natural-image domain outside face morphology
Latent-diffusion digit teacher	Conditional latent DDPM in a 32-dim PCA subspace of transformed digit-three images; direct sample size $n = 120$; mean-access and score-access GED rows	GED tracks the projection target when the conditional sampler operates in a low-dimensional latent space, and the single-step Tweedie denoising procedure generalizes the CelebA score-access advantage to a different conditional sampler
CelebA non-diffusion ridge teacher	Ridge-regression conditional mean teacher on a 1500-image CelebA teacher pool with the same official attributes	GED retains its advantage when the teacher is a smooth deterministic regressor and not a stochastic generator

The cross-regime SISE summary in Table 1 excludes oracle and teacher-pool rows from the baseline comparison; the GED column reports effect-access for the controlled synthetic and learned synthetic teacher blocks, score-access for the CelebA conditional diffusion and latent-diffusion digit teacher blocks, and mean-access for the remaining blocks, with the corresponding block tables in Tables 5,

9, 10, 11, 12, 15, 16, 18, 20, 21, and 22. Five of eleven blocks reject the null at $p < 10^{-3}$, namely the controlled synthetic, CelebA conditional diffusion, FFHQ measured attributes, Olivetti face morphology, and latent-diffusion digit teacher blocks, and four of these five further reject at $p < 10^{-4}$. The SISE ratio is larger than one in every block, with the smallest ratio in the semi-real digit block and the largest ratio in the CelebA conditional diffusion block.

Ablation interpretation. The empirical blocks in Table 4 are designed to isolate different terms in (20). The finite-query scaling table varies M and K and separates query-design error from teacher Monte Carlo error. The teacher-bias table varies the teacher mean while holding the target design fixed, thereby testing the stability term in Proposition 8. The design-sensitivity table changes ν, which changes the estimand itself when $\mu_\theta(x, s)$ is nonlinear in x. The real-image scaling tables vary the number of observed images available to direct regression while keeping the teacher construction fixed, which identifies the sparse observed-image regime where teacher queries are most useful. The pretrained SDXL-Turbo block tests sample-access distillation with a pretrained image-to-image teacher and separates target-design queries from augmentation at the observed covariates. The FFHQ measured-attribute block separates real measured covariates from the prompt-coded scalar covariates used in the pretrained SDXL-Turbo analysis. The CelebA semantic-diffusion block replaces image-derived covariates by official binary face attributes and trains a conditional diffusion teacher, thereby testing whether GED recovers maps for named facial attributes beyond image-measured summaries. Throughout, observed-covariate augmentation is an augmentation baseline tied to the observed design, not an equal-query control. For a linear student, an equal-budget target-design synthetic OLS fit is exactly the mean-access GED estimator, while an equal-budget resampling of observed covariates tests repeated teacher sampling at the observed design. Thus, the augmentation comparisons are interpreted as observed-design augmentation comparisons and not pure query-count ablations.

Controlled simulation design. The controlled simulation uses 32×32 images, $p = 2$ covariates, $n = 120$ observed images for direct regression, $M = 400$ query covariates for the main GED mean experiment, and $K = 8$ teacher samples per query. Direct regression fits OLS to observed images, augmentation queries the teacher at the observed covariates, GED mean queries the target design ν, GED score uses the Gaussian score identity, and GED effect averages analytic teacher gradients. This design separates three effects that are conflated by ordinary synthetic augmentation: the target design distribution, teacher sampling noise, and the choice of teacher information source. Let $\varphi_{a,b,\ell}(s)$ denote the unnormalized Gaussian bump $\exp[-\{(s_1-a)^2+(s_2-b)^2\}/(2\ell^2)]$ on the $[-1,1]^2$ image grid. The fixed nonlinear templates are

$$\gamma_1(s) = 0.62\varphi_{0.00,0.00,0.28}(s) - 0.40\varphi_{0.52,0.18,0.20}(s),$$
$$\gamma_2(s) = 0.55\varphi_{-0.20,0.55,0.18}(s) - 0.45\varphi_{0.58,0.55,0.16}(s),$$
$$\gamma_3(s) = 0.78\varphi_{-0.12,0.58,0.22}(s).$$

The teacher-bias sensitivity experiment uses $b_{\text{bias}}(s) = (0.80\gamma_1(s) - 0.55\gamma_3(s),\ -0.65\gamma_2(s) + 0.45\gamma_1(s))^\top$. The analytic teacher is

$$Y(s) = \beta_0(s) + X^\top\beta(s) + h(X, s) + \xi(s), \qquad X \sim \mathcal{N}(0, I_2),$$

where $\xi(s)$ is smoothed Gaussian image noise and

$$h(X, s) = 0.60(X_1^2 - 1)\gamma_1(s) + 0.45X_1X_2\gamma_2(s) + 0.60\sin(X_2)\gamma_3(s).$$

Because the design is Gaussian with canonical Stein kernel $T_\nu(X) = I_2$, Proposition 3 and Theorem 3 identify the same slope target in this experiment. Under $\nu = \mathcal{N}(0, I_2)$, the projection target satisfies

$$\beta_1^\star(s) = \beta_1(s), \qquad \beta_2^\star(s) = \beta_2(s) + 0.60\exp(-1/2)\gamma_3(s).$$

Table 5: Controlled synthetic experiment. Entries are averages over 80 independent repetitions, with standard errors after the plus-minus sign. The Region F1 column is the signed-region F1 metric defined in Appendix F.

Method	Standardized ISE	Region F1
Direct	0.018 ± 0.001	0.957 ± 0.002
Augmentation	0.012 ± 0.001	0.967 ± 0.002
GED mean	0.004 ± 0.000	0.982 ± 0.001
GED score	0.003 ± 0.000	0.984 ± 0.001
GED effect	0.001 ± 0.000	0.992 ± 0.001

Table 6: Finite-query scaling for GED mean in the controlled synthetic experiment. Entries are standardized integrated squared errors averaged over 80 independent repetitions, with standard errors after the plus-minus sign.

M	$K = 1$	$K = 4$	$K = 16$
50	0.044 ± 0.003	0.030 ± 0.003	0.022 ± 0.002
100	0.020 ± 0.002	0.017 ± 0.002	0.013 ± 0.001
200	0.010 ± 0.001	0.008 ± 0.001	0.007 ± 0.001
400	0.006 ± 0.000	0.004 ± 0.000	0.003 ± 0.000

Table 7: Teacher-bias sensitivity. The biased teacher mean is obtained by adding $\rho X^\top b_{\mathrm{bias}}(s)$ to the true teacher mean. The third column measures recovery of the biased teacher projection target, whereas the fourth column measures error relative to the original true target.

ρ	Target shift	SISE to teacher target	SISE to true target
0.00	0.000	0.004 ± 0.000	0.004 ± 0.000
0.25	0.019	0.003 ± 0.000	0.023 ± 0.001
0.50	0.075	0.004 ± 0.000	0.078 ± 0.002
0.75	0.168	0.003 ± 0.000	0.169 ± 0.003

Table 8: Design-distribution sensitivity. The target itself changes with ν because the teacher mean is nonlinear in X. The standard design is $\nu_0 = \mathcal{N}(0, I_2)$, the shifted design is $\nu_{\mathrm{shift}} = \mathcal{N}\{(0, 0.75)^\top, I_2\}$, and the high-variance design is $\nu_{\mathrm{wide}} = \mathcal{N}\{0, \mathrm{diag}(1, 1.50^2)\}$. Recovery SISE is computed relative to the projection target associated with the design in the first column.

Target design ν	Distance from standard target	Recovery SISE
Standard Gaussian	0.000	0.004 ± 0.000
Shifted X_2	0.018	0.004 ± 0.000
High-variance X_2	0.008	0.003 ± 0.000

Simulation analysis. The finite-query table verifies the rate structure in Theorem 1: at fixed $K = 1$, SISE decreases from 0.044 to 0.006 as M increases from 50 to 400, and at fixed $M = 400$, SISE decreases from 0.006 to 0.003 as K increases from 1 to 16. The teacher-bias table shows a different phenomenon: when $\rho = 0.75$, GED estimates the biased teacher target with SISE 0.003 but has SISE 0.169 relative to the original target, matching the teacher-bias term in (20). The design-sensitivity table shows that the target itself changes with ν, and that GED recovers each design-specific target with SISE at most 0.004.

Table 9: Learned neural teacher experiment. A neural conditional mean teacher is trained on 1000 noisy images and then distilled into the same linear image-response student through both mean access and effect access. The mean-access row queries the teacher at $M = 800$ target-design covariates; the effect-access row averages the teacher gradient $\partial\mu_\theta/\partial x_j$ obtained by autodifferentiation through the trained network at the same $M = 800$ query covariates and matches it to the linear student through Theorem 3. The coordinate network row implements an image-response coordinate-MLP baseline (Zhang et al., 2023) fitted to the same $n = 120$ observed image-covariate pairs as the direct-OLS baseline. The Integrated Gradients (Sundararajan et al., 2017) and SmoothGrad (Smilkov et al., 2017) rows compute path-integrated and noise-averaged saliency through the same trained teacher network at the same $n = 120$ observed sample, then set each coefficient map to the sample average of the corresponding attribution maps, equivalently intercept-only least squares across observed instances. Entries are standardized integrated squared errors over 8 independent repetitions.

Method	Standardized ISE
Direct OLS, $n = 120$	0.012 ± 0.002
Coordinate network, $n = 120$	0.009 ± 0.002
Integrated Gradients, $n = 120$	0.007 ± 0.002
SmoothGrad, $n = 120$	0.008 ± 0.001
Teacher augmentation	0.007 ± 0.002
GED learned teacher (mean)	0.003 ± 0.001
GED learned teacher (effect)	0.002 ± 0.000
GED oracle mean	0.002 ± 0.001
Full-data OLS, $n = 1000$	0.002 ± 0.000

Table 10: Semi-real handwritten digit analysis. The teacher is trained on transformed noisy digit-three images and distilled into a linear image-response student. The teacher-training pool has 800 transformed digit samples, direct OLS uses $n = 120$, teacher augmentation repeats the observed covariates eight times, and GED uses $M = 800$ target-design queries. Entries are averages over 8 independent repetitions, with standard errors after the plus-minus sign.

Method	Standardized ISE	Map correlation
Direct OLS, $n = 120$	0.049 ± 0.005	0.976 ± 0.002
Teacher augmentation	0.020 ± 0.002	0.990 ± 0.001
GED learned teacher	0.016 ± 0.001	0.993 ± 0.000
GED oracle mean	0.001 ± 0.000	1.000 ± 0.000

Diffusion-teacher analysis. The pretrained SDXL-Turbo/FFHQ-256 analysis replaces the direct neural mean teacher by a pretrained image-to-image teacher. The teacher uses FFHQ-256 base faces and prompt-coded scalar covariates: x_1 changes eyeglasses and smile content, and x_2 changes background and illumination content. In Table 11, GED pretrained SDXL teacher attains SISE 0.818, compared with 20.018 for direct OLS and 13.092 for observed-covariate augmentation; map correlation improves from 0.240 and 0.215 to 0.650, and signed-region F1 improves from 0.193 and 0.179 to 0.544.

Table 11: Pretrained SDXL-Turbo/FFHQ-256 diffusion-teacher experiment. A pretrained SDXL-Turbo image-to-image teacher is queried on FFHQ-256 base faces under prompt-coded scalar covariates. The reference is an OLS projection fitted from 512 target-design teacher queries; GED uses 512 target-design queries per repetition, direct OLS uses 12 observed teacher samples, and observed-covariate augmentation adds 12 teacher samples at the observed covariates. Entries are averages over 5 independent repetitions, with standard errors after the plus-minus sign.

Method	Standardized ISE	Map correlation	Sign F1
Direct OLS, $n = 12$	20.018 ± 2.912	0.240 ± 0.063	0.193 ± 0.020
SDXL observed-covariate augmentation	13.092 ± 3.032	0.215 ± 0.073	0.179 ± 0.020
GED pretrained SDXL teacher	0.818 ± 0.069	0.650 ± 0.018	0.544 ± 0.026

CelebA conditional diffusion analysis. We analyze aligned CelebA face images with four official binary attributes: Smiling, Eyeglasses, Male, and Blond Hair (Liu et al., 2015). The reference target is the full-data OLS association map over 2000 images after center-cropping and downsampling to 32×32 RGB maps. We train a conditional residual DDPM teacher on 1800 images: a condition-specific mean is estimated from the teacher-training pool, and a conditional DDPM models the residual image distribution around that mean. GED queries this diffusion teacher at 768 empirical-design covariates per repetition, diffusion observed-covariate augmentation queries the same teacher four times at each of the 80 observed covariates, and direct OLS uses 80 observed images. In Table 12, mean-access GED attains SISE 0.090, score-access GED attains SISE 0.009, compared with 1.424 for direct OLS and 0.229 for diffusion observed-covariate augmentation; map correlation improves from 0.603 and 0.890 to 0.955 and 0.996, respectively, and signed-region F1 improves from 0.397 and 0.643 to 0.769 and 0.936. The score-access row uses the single-step Tweedie denoising procedure described in Appendix I. Its three metric values are closer to the teacher-pool OLS reference (0.005, 0.997, 0.932) than the mean-access values, with fewer network evaluations per query. On this conditional residual DDPM, score access reduces SISE by a factor of 10 relative to mean access at the same query budget.

Table 12: CelebA official semantic-attribute conditional diffusion analysis. The covariates are the official CelebA attributes Smiling, Eyeglasses, Male, and Blond Hair, and the reference target is the full-data OLS association map over 2000 aligned face images. The mean-access row queries the teacher through a 30-step DDIM sampler, while the score-access row reads the teacher score off a single-step Tweedie denoiser at noise level $\tau = 20$, as described in Appendix I. The "Teacher-pool OLS" row is the OLS fit on the CelebA image pool used to train the conditional residual DDPM teacher, and is separate from the full-data OLS reference. Entries are averages over 5 independent query and observed-sample repetitions, except for the score-access row, which uses 15 repetitions as indicated in the method label; standard errors are shown after the plus-minus sign. Sign F1 uses the same active-region thresholding rule as the controlled synthetic experiment.

Method	Standardized ISE	Map correlation	Sign F1
Direct OLS, $n = 80$	1.424 ± 0.124	0.603 ± 0.044	0.397 ± 0.024
Diffusion observed-covariate augmentation	0.229 ± 0.010	0.890 ± 0.004	0.643 ± 0.014
GED conditional diffusion teacher (mean)	0.090 ± 0.004	0.955 ± 0.002	0.769 ± 0.005
GED conditional diffusion teacher (score), 15 reps	0.009 ± 0.000	0.996 ± 0.000	0.936 ± 0.001
Teacher-pool OLS	0.005 ± 0.000	0.997 ± 0.000	0.932 ± 0.000

Table 13: Finite-query scaling for score-access GED on the trained CelebA conditional residual DDPM teacher. At each query size M, the single-step Tweedie denoising procedure described in Appendix I is evaluated at noise level $\tau = 20$ and the resulting denoised score-access queries are projected onto the linear student. Entries are SISE averaged over 5 independent repetitions, with standard errors after the plus-minus sign, and show decreasing finite-query error as the query design size increases.

M	SISE
100	0.033 ± 0.001
200	0.018 ± 0.000
400	0.012 ± 0.000
800	0.009 ± 0.000
1600	0.007 ± 0.000

Table 14: Tweedie noise-level sensitivity for score-access GED on the trained CelebA conditional residual DDPM teacher. The score-access estimator at each noise level τ uses a single ϵ-evaluation per query through Tweedie's identity, with $M = 768$ empirical-design queries and 5 repetitions. Lower noise levels reduce denoising bias at the cost of requiring the score to be accurate near the data manifold; score-access GED attains SISE below 0.01 for $\tau \in \{5, 10, 20\}$ and remains below the 0.090 mean-access SISE in Table 12 for every choice in the range.

Tweedie τ	SISE
5	0.006 ± 0.000
10	0.007 ± 0.000
20	0.009 ± 0.000
50	0.013 ± 0.000
100	0.025 ± 0.001

FFHQ measured-attribute analysis. To separate prompt-coded covariates from measured face-image attributes, we also analyze the FFHQ-256 subset using four covariates computed from each image: horizontal and vertical luminance centers, mean luminance, and RGB warmth. The reference target is the full-data OLS association map over 100 FFHQ-256 images after downsampling to 32×32 RGB maps. A quadratic ridge teacher is trained on 80 images, GED queries this teacher at 1200 empirical-design covariates per repetition, and direct OLS uses 25 observed images. In Table 15, GED attains SISE 0.021, compared with 0.281 for direct OLS and 0.079 for teacher augmentation; map correlation improves from 0.890 and 0.961 to 0.990, and signed-region F1 improves from 0.748 and 0.839 to 0.914.

Table 15: FFHQ-256 measured-attribute analysis. The covariates are computed directly from each FFHQ-256 image, and the reference target is the full-data OLS association map over all 100 images. The "Teacher-pool OLS" row is the OLS fit on the FFHQ image pool used to train the teacher, and is separate from the full-data OLS reference. Entries are averages over 8 independent repetitions, with standard errors after the plus-minus sign. Sign F1 uses the same active-region thresholding rule as the controlled synthetic experiment.

Method	Standardized ISE	Map correlation	Sign F1
Direct OLS, $n = 25$	0.281 ± 0.027	0.890 ± 0.006	0.748 ± 0.010
Teacher augmentation	0.079 ± 0.005	0.961 ± 0.003	0.839 ± 0.006
GED FFHQ-attribute teacher	0.021 ± 0.003	0.990 ± 0.001	0.914 ± 0.007
Teacher-pool OLS	0.015 ± 0.001	0.993 ± 0.000	0.927 ± 0.003

Real-data analysis. The semi-real digit analysis tests teacher learning under realistic digit shape variation while keeping the projection target known through a deterministic transformation mean. The real-image morphology analysis removes synthetic covariate injection: covariates are measured from each original digit image, and the full-data OLS map over all 183 images is used only as a reference association map. Table 17 reports the sample-size sensitivity. GED attains SISE 0.026 from a fixed learned teacher and target-design queries; this SISE is below the direct OLS SISE for $n = 30, 60, 90$, while direct OLS reaches SISE 0.021 at $n = 120$. The face analysis repeats this comparison on 400 face images with three measured covariates. GED attains SISE 0.006 at $n = 80$, below direct OLS for $n = 40, 80, 160$, and matches direct regression at $n = 240$. These results show that GED reduces SISE relative to direct regression when observed images are scarce and a calibrated conditional teacher is available, while direct regression matches GED once the observed sample size approaches the teacher-training pool size. In the main handwritten morphology row, the teacher is trained on 145 images, direct OLS uses $n = 60$, teacher augmentation repeats the observed covariates eight times, and GED uses $M = 800$ jittered empirical-design queries. In the main face morphology row, the corresponding values are 320 teacher-training images, $n = 80$, eight augmentation repeats, and $M = 1200$ GED queries.

Table 16: Real-image handwritten morphology analysis. The covariates are measured from original digit-three images, and the reference target is the full-data OLS association map over all 183 real images. The "Teacher-pool OLS" row is the OLS fit on the real-image pool used to train the teacher, and is separate from the full-data OLS reference. Entries are averages over 8 independent repetitions, with standard errors after the plus-minus sign. Sign F1 uses the same active-region thresholding rule as the controlled synthetic experiment.

Method	Standardized ISE	Map correlation	Sign F1
Direct OLS, $n = 60$	0.093 ± 0.011	0.959 ± 0.004	0.849 ± 0.013
Teacher augmentation	0.050 ± 0.009	0.977 ± 0.004	0.886 ± 0.009
GED real-image teacher	0.026 ± 0.003	0.988 ± 0.001	0.916 ± 0.007
Teacher-pool OLS	0.013 ± 0.002	0.994 ± 0.001	0.940 ± 0.004

Table 17: Sample-size sensitivity in the real-image handwritten morphology analysis. The teacher is trained on the same background real-image pool in each split, whereas the number of observed images used to fit the direct or augmented student is varied. Entries are standardized integrated squared errors averaged over 8 independent repetitions, with standard errors after the plus-minus sign.

Observed n	Direct OLS	Teacher augmentation	GED teacher
30	0.269 ± 0.036	0.072 ± 0.012	0.026 ± 0.003
60	0.093 ± 0.011	0.050 ± 0.009	0.026 ± 0.003
90	0.046 ± 0.004	0.028 ± 0.002	0.026 ± 0.003
120	0.021 ± 0.002	0.024 ± 0.002	0.026 ± 0.003

Table 18: Real-image face morphology analysis. The covariates are measured from Olivetti face images, and the reference target is the full-data OLS association map over all 400 images. The "Teacher-pool OLS" row is the OLS fit on the face-image pool used to train the teacher, and is separate from the full-data OLS reference. Entries are averages over 8 independent repetitions, with standard errors after the plus-minus sign. Sign F1 uses the same active-region thresholding rule as the controlled synthetic experiment.

Method	Standardized ISE	Map correlation	Sign F1
Direct OLS, $n = 80$	0.039 ± 0.002	0.981 ± 0.001	0.910 ± 0.005
Teacher augmentation	0.017 ± 0.002	0.992 ± 0.001	0.941 ± 0.004
GED face teacher	0.006 ± 0.001	0.997 ± 0.000	0.965 ± 0.001
Teacher-pool OLS	0.002 ± 0.000	0.999 ± 0.000	0.976 ± 0.001

Table 19: Sample-size sensitivity in the real-image face morphology analysis. The teacher is trained on the same background face-image pool in each split, whereas the number of observed images used to fit the direct or augmented student is varied. Entries are standardized integrated squared errors averaged over 8 independent repetitions, with standard errors after the plus-minus sign.

Observed n	Direct OLS	Teacher augmentation	GED teacher
40	0.111 ± 0.011	0.033 ± 0.003	0.006 ± 0.001
80	0.039 ± 0.002	0.017 ± 0.002	0.006 ± 0.001
160	0.014 ± 0.001	0.009 ± 0.001	0.006 ± 0.001
240	0.006 ± 0.000	0.006 ± 0.001	0.006 ± 0.001

CIFAR-10 conditional pixel-diffusion teacher. To extend the analysis beyond face-image domains, we run a parallel analysis with a conditional pixel-space DDPM trained on CIFAR-10 (Krizhevsky, 2009) at 32×32 RGB resolution. The teacher is trained on 6000 images with two-dimensional Gaussian covariates, and GED queries are drawn from the same Gaussian design with $M = 900$

and $K = 4$. Direct OLS uses $n = 160$ observed images, and the conditional diffusion augmentation baseline matches GED in teacher access but reuses the observed covariates. Held-out teacher RMSE on the residual scale is 0.072 ± 0.000 . In Table 20, GED attains SISE 0.011, compared with 0.049 for direct OLS and 0.014 for diffusion observed-covariate augmentation; map correlation improves from 0.976 and 0.997 to 0.998, and signed-region F1 improves from 0.943 and 0.989 to 0.991. The oracle row reports the SISE for the analytic teacher mean and is included as a reference floor.

Table 20: CIFAR-10 conditional pixel-diffusion teacher experiment. A conditional pixel-diffusion teacher is trained on 6000 CIFAR-10 images with two-dimensional Gaussian covariates and queried under the same design. Entries are averages over 6 independent repetitions, with standard errors after the plus-minus sign.

Method	Standardized ISE	Map correlation	Sign F1
Direct OLS, $n = 160$	0.049 ± 0.004	0.976 ± 0.002	0.943 ± 0.006
CIFAR diffusion augmentation	0.014 ± 0.000	0.997 ± 0.000	0.989 ± 0.001
GED CIFAR diffusion teacher	0.011 ± 0.001	0.998 ± 0.000	0.991 ± 0.001
GED oracle mean	0.001 ± 0.000	1.000 ± 0.000	0.994 ± 0.001

Latent-diffusion teacher with semi-real digit covariates. To separate the conditional sampling mechanism from the pixel space, we also run an analysis in which the teacher is a conditional latent DDPM operating in a 32-dimensional PCA subspace of transformed digit-three images. Training uses 1400 images with two-dimensional Gaussian covariates, GED uses $M = 800$ query covariates and $K = 4$ teacher samples per query, and direct OLS uses $n = 120$ observed images. Held-out teacher RMSE on the residual scale is 0.060 ± 0.000 . In Table 21, mean-access GED attains SISE 0.028 and score-access GED via the same single-step Tweedie denoising procedure used in the CelebA block (applied in the 32-dimensional latent space and decoded back through the PCA codec) attains SISE 0.015, compared with 0.043 for direct OLS and 0.034 for latent-diffusion augmentation; map correlation improves from 0.979 and 0.990 to 0.993 and 0.996, respectively, and signed-region F1 improves from 0.896 and 0.899 to 0.906 and 0.932. Score-access GED on this latent diffusion teacher reduces SISE by a further $1.9\times$ over the mean-access row at the same $M = 800$ query budget ($p \approx 3.2 \times 10^{-5}$, paired one-sided t-test), extending the CelebA score-access SISE reduction to a different conditional sampler.

Table 21: Latent-diffusion teacher experiment with transformed digit-three responses. A conditional latent DDPM is trained in a 32-dim PCA latent space and decoded back to the image space. Entries are averages over 8 independent repetitions, with standard errors after the plus-minus sign.

Method	Standardized ISE	Map correlation	Sign F1
Direct OLS, $n = 120$	0.043 ± 0.002	0.979 ± 0.001	0.896 ± 0.004
Latent diffusion augmentation	0.034 ± 0.003	0.990 ± 0.001	0.899 ± 0.009
GED latent diffusion teacher (mean)	0.028 ± 0.002	0.993 ± 0.001	0.906 ± 0.007
GED latent diffusion teacher (score)	0.015 ± 0.001	0.996 ± 0.000	0.932 ± 0.003
GED oracle mean	0.001 ± 0.000	0.999 ± 0.000	0.987 ± 0.002

CelebA semantic attributes with a non-diffusion teacher. To isolate the role of the diffusion sampling mechanism in the CelebA analysis, we also evaluate GED with a deterministic conditional mean teacher trained by ridge regression on a 1500-image CelebA teacher pool with the same official attributes. Held-out teacher RMSE is 0.271 ± 0.001 . In Table 22, GED attains SISE 0.024 over 8 repetitions, compared with 1.657 for direct OLS and 0.048 for teacher augmentation. The GED SISE reduction relative to direct OLS persists when the teacher is a deterministic regressor and not a stochastic generator. Combined with Table 12, this comparison disentangles the sampling mechanism from the design-distribution effect: under both teacher classes, GED has SISE closer to the teacher-pool OLS reference than direct regression with $n = 80$ observed images.

Table 22: CelebA semantic-attribute analysis with a deterministic ridge teacher, not a conditional diffusion teacher. Entries are averages over 8 independent repetitions, with standard errors after the plus-minus sign.

Method	Standardized ISE	Map correlation	Sign F1
Direct OLS, $n = 80$	1.657 ± 0.152	0.612 ± 0.034	0.394 ± 0.014
Teacher augmentation	0.048 ± 0.006	0.978 ± 0.003	0.827 ± 0.015
GED CelebA-attribute teacher	0.024 ± 0.002	0.989 ± 0.001	0.871 ± 0.012
Teacher-pool OLS	0.025 ± 0.002	0.988 ± 0.001	0.862 ± 0.013

F.1 Additional sensitivity and audit studies

Teacher validity and calibration. We distinguish accurate recovery of the declared teacher projection from agreement between the teacher and an observed-image population. The controlled study in Table 23 uses 400 held-out observations and $M = 400$ independent query covariates in each of 100 repetitions. GED continues to estimate the mean-biased teacher projection accurately, with SISE $0.00232\,(0.00023)$ relative to that projection, while its SISE relative to the true map is $0.17399\,(0.00229)$. Held-out RMSE detects the mean shift, whereas negative log score and predictive coverage detect scale misspecification that leaves the mean map unchanged.

Table 23: Controlled teacher-validity diagnostic over 100 repetitions. Values are means with Monte Carlo standard errors in parentheses. Map SISE is relative to the true projection; NLL is the Gaussian negative log score.

Teacher case	Map SISE	Held-out RMSE	NLL	90% coverage
Calibrated	.00269 (.00027)	.25002 (.00003)	.03274 (.00013)	.89996 (.00005)
Mean biased	.17399 (.00229)	.27467 (.00016)	.13619 (.00072)	.87076 (.00018)
Underdispersed	.00269 (.00027)	.25002 (.00003)	.83987 (.00052)	.58903 (.00009)
Overdispersed	.00269 (.00027)	.25002 (.00003)	.35081 (.00003)	.99899 (.00001)

Spatial frequency and matched query budgets. Table 24 uses the same $M = 400$ query covariates for voxelwise and 8×8 block students under a nonlinear analytic teacher. The voxelwise linear student retains localized high-frequency coefficient maps, whereas the coarse student loses them. This comparison isolates spatial representation from linearity in the covariates and does not establish native-resolution recovery.

Table 24: High-frequency coefficient-map recovery over 100 repetitions. Entries are SISE means with Monte Carlo standard errors in parentheses.

Frequency	GED mean, voxelwise	GED effect, voxelwise	GED mean, 8×8 blocks
1	.000206 (.000023)	.000062 (.000008)	.08340 (.00001)
4	.000210 (.000021)	.000075 (.000009)	.76532 (.00000)
8	.000217 (.000019)	.000078 (.000008)	.98530 (.00000)

The matched-budget study in Table 25 conditions on one fixed learned teacher with held-out RMSE 0.0447 and compares every estimator with that teacher's $50{,}000$-query population projection. At $M = n = 120$, mean-access GED and the observed-design teacher fit are statistically indistinguishable, while effect access reduces SISE. Increasing M to 800 reduces both GED errors. These comparisons quantify query and design randomness conditional on the teacher and do not include teacher-training variability.

Table 25: Matched-query comparison over 100 repetitions. SISE is relative to the fixed teacher projection and is reported as mean with Monte Carlo standard error in parentheses. Paired p compares each GED row with the observed-design teacher fit using a one-sided paired t-test.

Method	SISE to teacher projection	Paired p
Observed-design teacher, $n = 120$	.011972 (.000978)	—
GED mean, $M = 120$	.012266 (.001111)	.577
GED effect, $M = 120$	.005549 (.000471)	1.20×10^{-8}
GED mean, $M = 800$	.001911 (.000183)	7.09×10^{-17}
GED effect, $M = 800$	.000844 (.000066)	1.01×10^{-19}

Registration sensitivity. For each image in the controlled 32×32 setting, we apply an independent translation distributed as $N(0, \sigma^2 I_2)$ pixels using bilinear interpolation and nearest-edge padding. Table 26 shows that GED remains accurate for the projection defined by the misaligned image distribution, but the projection itself moves away from the aligned target. The target change dominates the aligned-target error at one to two pixels.

Table 26: Rigid-translation sensitivity of mean-access GED over 100 repetitions with $M = 400$ and $K = 8$. GED entries are SISE means with Monte Carlo standard errors in parentheses; target shift is the SISE of the misaligned population projection relative to the aligned target.

σ (pixels)	Target shift	GED to aligned target	GED to misaligned projection
0.0	.00000	.00393 (.00042)	.00393 (.00042)
0.5	.00095	.00495 (.00041)	.00423 (.00042)
1.0	.00660	.00969 (.00025)	.00350 (.00028)
2.0	.05642	.05989 (.00035)	.00451 (.00034)

Localized teacher-bias audit. To test a concrete map use, we distill the biased analytic teacher at $\rho = 0.75$ with $M = 400$ and $K = 8$ and subtract an association map estimated from 400 independent observations drawn from the unbiased distribution: $\widehat{A}_j(s) = \widehat{\beta}_{\mathrm{teacher},j}(s) - \widehat{\beta}_{\mathrm{ref},j}(s)$. To define the active artifact region, we apply the same sign-sensitive thresholding rule used in the other localization studies. Over 100 repetitions, the resulting audit map attains correlation $0.9795\,(0.0016)$, signed-region F1 $0.9021\,(0.0046)$, precision $0.9182\,(0.0041)$, and recall $0.8972\,(0.0072)$. When the teacher is unbiased, applying the same magnitude threshold flags $0.0015\,(0.0005)$ of the pixels. This controlled diagnostic requires an independent reference sample. Clinical validity and automatic detection of spurious associations without an external reference require separate evaluation.

Table 27: Controlled localization of an injected teacher artifact over 100 repetitions. Numbers in parentheses are Monte Carlo standard errors.

Map correlation	Signed F1	Precision	Recall	Null flagged fraction
.9795 (.0016)	.9021 (.0046)	.9182 (.0041)	.8972 (.0072)	.0015 (.0005)

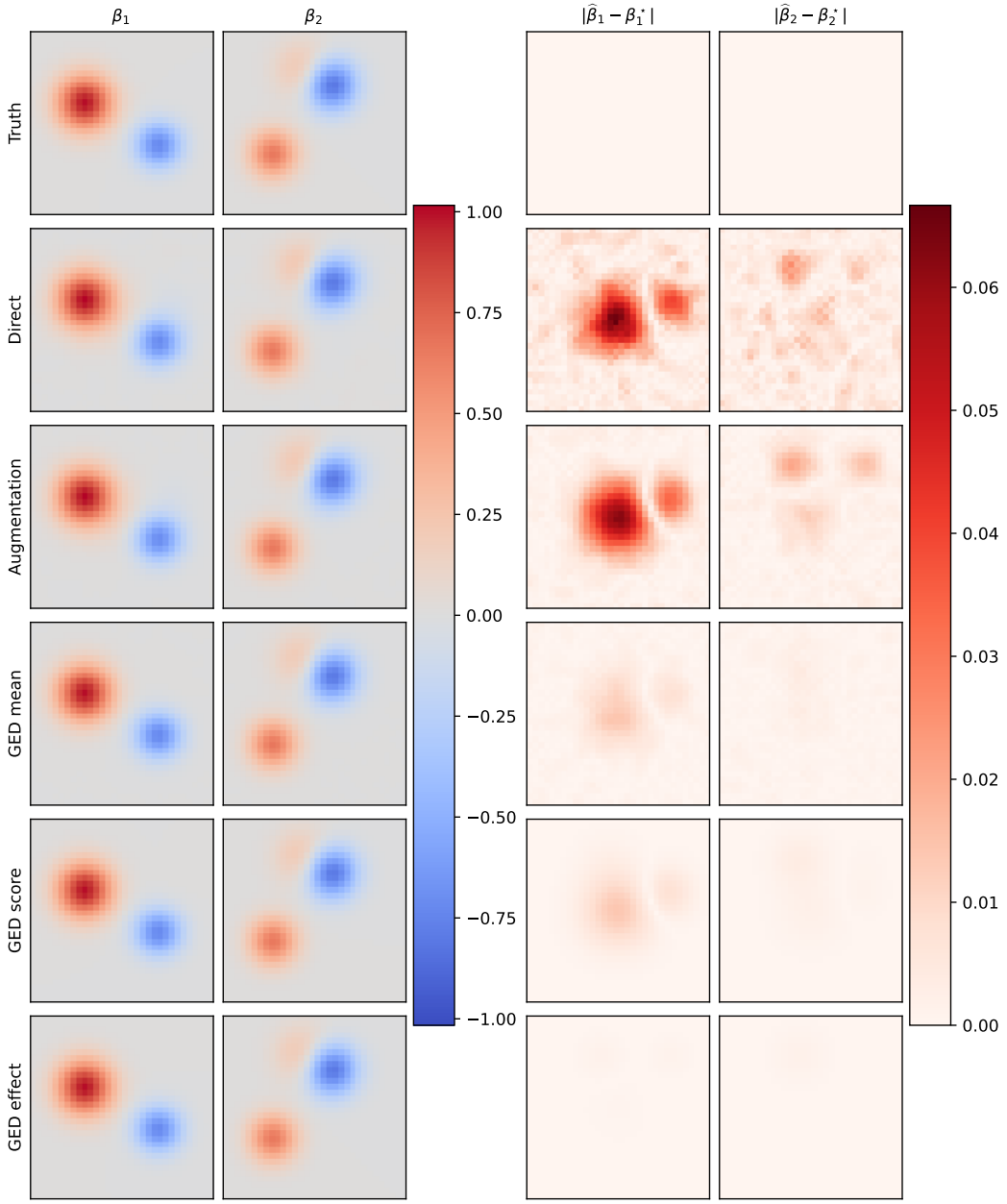

Figure 2: Estimated spatial effect maps in the controlled synthetic experiment. Left panel: β_1 and β_2 for Truth, Direct, Augmentation, GED mean, GED score, and GED effect. Right panel: per-pixel absolute error $|\widehat{\beta}_j - \beta_j^\star|$ on a common scale. GED rows have smaller per-pixel absolute error than Direct and Augmentation for both covariates, and GED effect has the smallest per-pixel absolute error, consistent with the $12\times$ SISE ratio in Table 5.

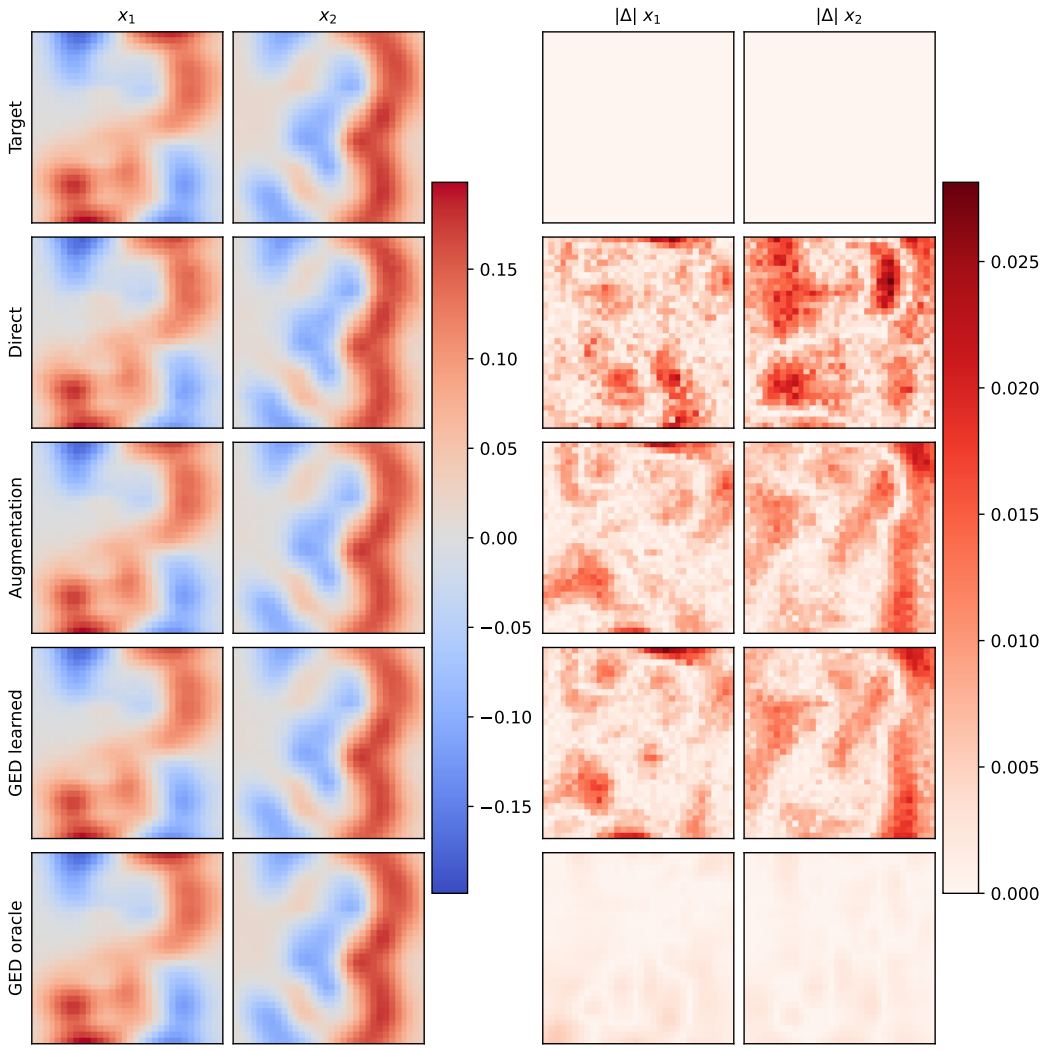

Figure 3: Semi-real handwritten digit effect maps. Left panel: x_1 (rotation) and x_2 (translation) coefficient maps for Target, Direct, Augmentation, GED learned, and GED oracle. Right panel: per-pixel absolute error $|\widehat{\beta}_j - \beta_j^\star|$ on a common scale. GED learned has smaller per-pixel absolute error than Direct under the same sparse observed-image regime, and GED oracle has the smallest per-pixel absolute error.

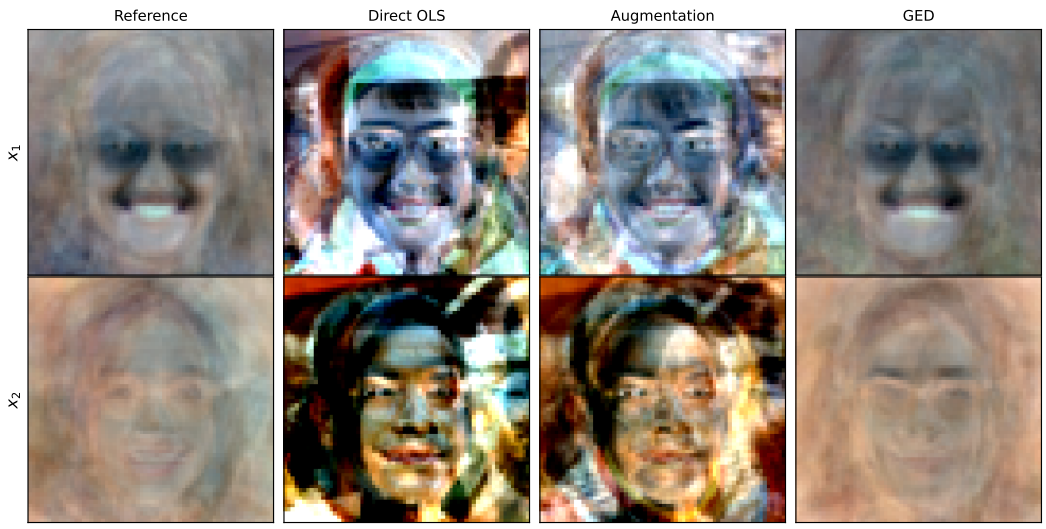

Figure 4: Pretrained SDXL-Turbo/FFHQ-256 teacher effect maps. GED uses SDXL-Turbo teacher images at target-design query covariates and is compared with direct regression, observed-covariate augmentation, and the high-query reference projection. For visualization, RGB coefficient maps use a common scale set by the reference and GED maps; the direct and augmentation maps can saturate under the same scale.

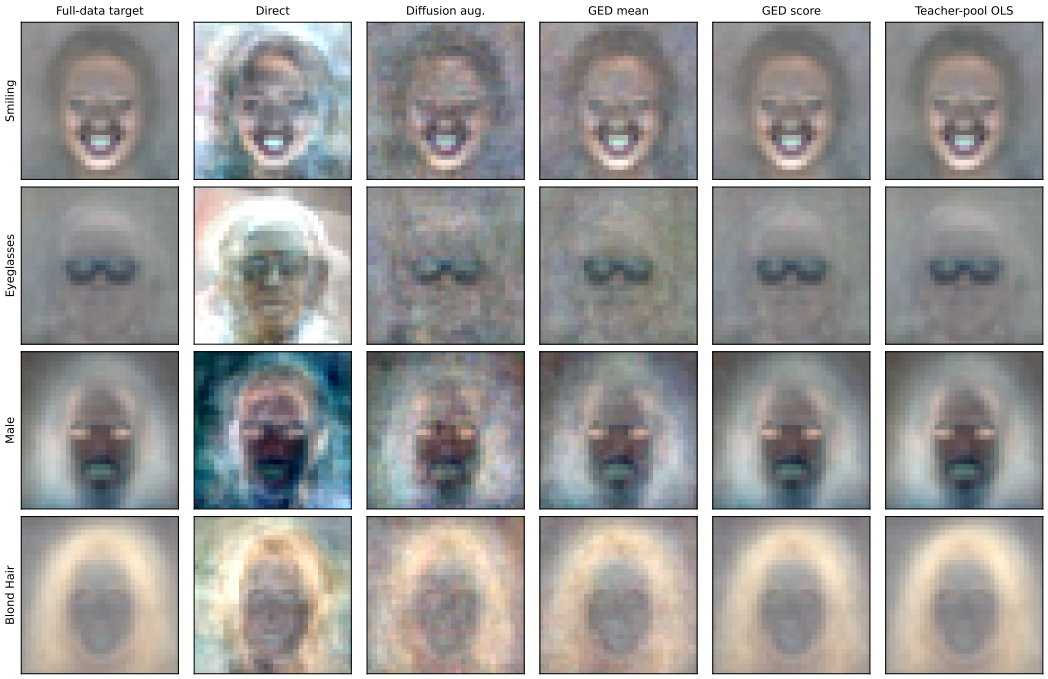

Figure 5: CelebA official semantic-attribute conditional diffusion effect maps for the four attributes Smiling, Eyeglasses, Male, and Blond Hair. The columns are the full-data target, direct OLS at $n = 80$, diffusion observed-covariate augmentation, mean-access GED via 30-step DDIM sampling, score-access GED via the single-step Tweedie denoising procedure described in Appendix I, and the teacher-pool OLS reference. Score-access GED has SISE 0.009, compared with 0.005 for teacher-pool OLS and 0.229 for observed-covariate augmentation in Table 12. "Teacher-pool OLS" uses the OLS fit on the CelebA image pool used to train the diffusion teacher, not the full-data reference.

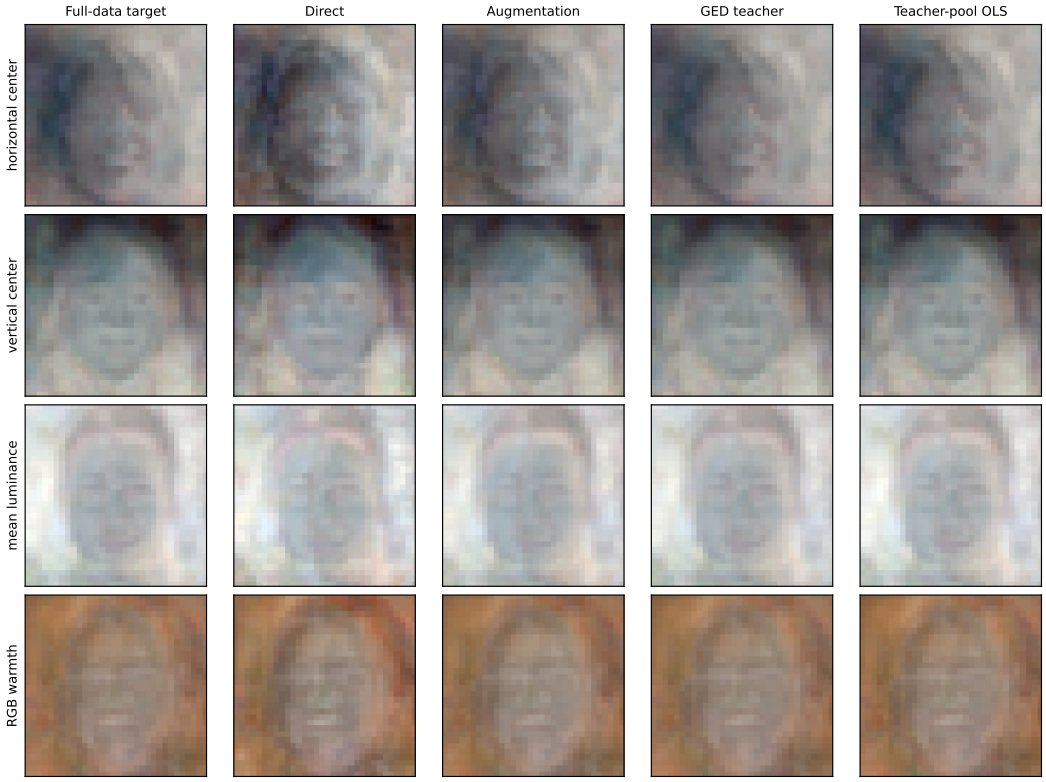

Figure 6: FFHQ-256 measured-attribute effect maps. GED maps have lower SISE relative to the full-data association target than direct regression and teacher augmentation under the same sparse observed-image regime, as reported in Table 15. The column labelled "Teacher-pool OLS" uses the OLS fit on the FFHQ image pool used to train the teacher, not the full-data reference.

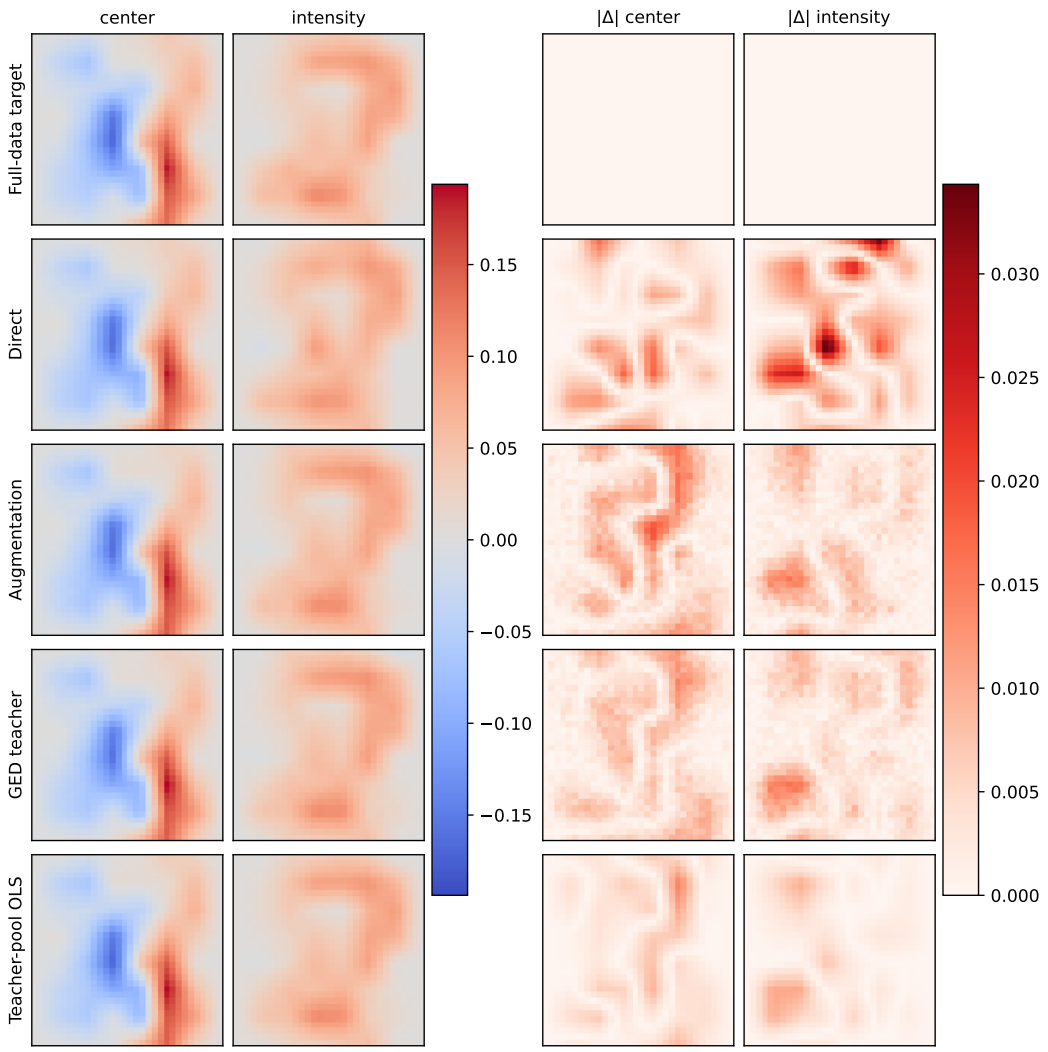

Figure 7: Real-image handwritten morphology effect maps. Left panel: center and intensity coefficient maps for Full-data target, Direct, Augmentation, GED teacher, and Teacher-pool OLS (rendered from 4 repetitions). Right panel: per-pixel absolute error $|\widehat{\beta} - \beta^{\star}|$ on a common scale. The Direct row has the largest per-pixel absolute error; GED teacher and Teacher-pool OLS have the smallest. The "Teacher-pool OLS" row uses the OLS fit on the real-image pool used to train the teacher.

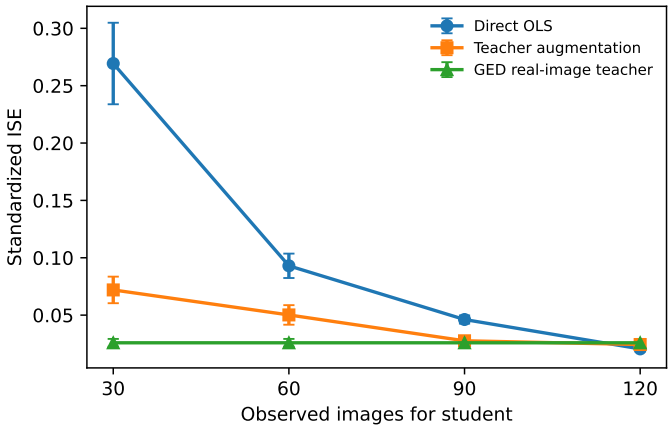

Figure 8: Sample-size sensitivity in the real-image handwritten morphology analysis. GED has the smallest standardized error in the sparse observed-image regimes, while direct OLS attains comparable error once the number of observed images approaches the teacher-training pool size.

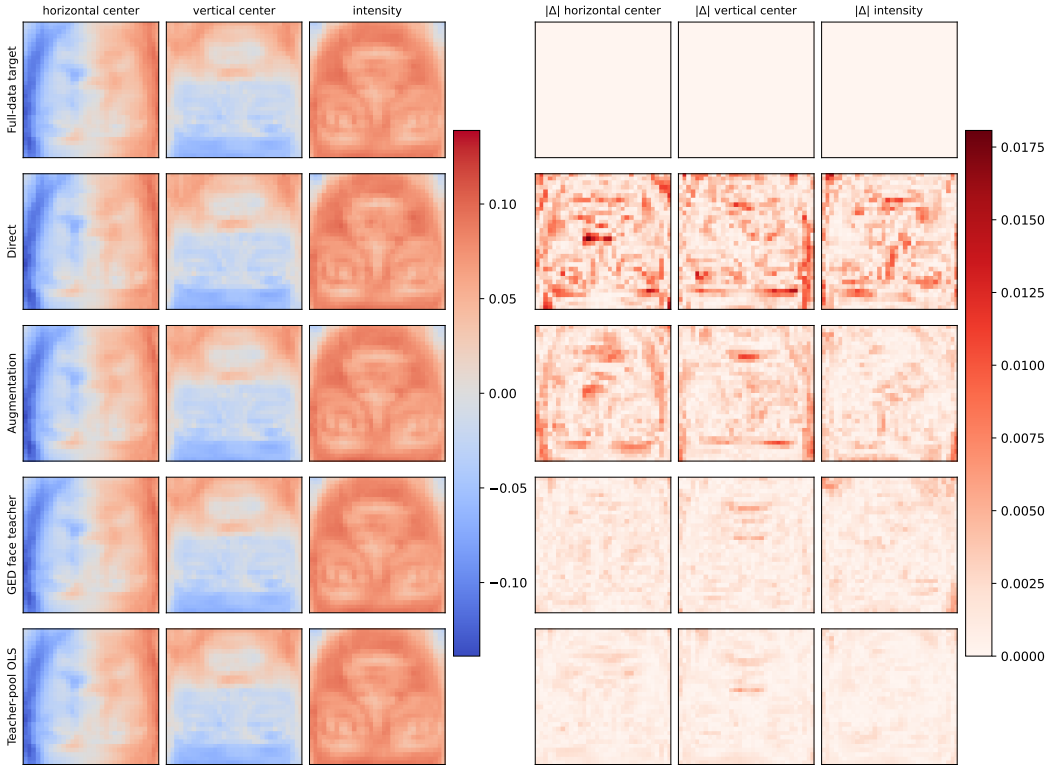

Figure 9: Real-image face morphology effect maps. Left panel: horizontal center, vertical center, and intensity coefficient maps for Full-data target, Direct, Augmentation, GED face teacher, and Teacher-pool OLS (rendered from 4 repetitions). Right panel: per-pixel absolute error $|\widehat{\beta} - \beta^{\star}|$ on a common scale. GED face teacher has lower SISE than Direct and Augmentation, consistent with the $6\times$ SISE ratio over direct OLS in Table 18. The "Teacher-pool OLS" row uses the OLS fit on the face-image pool used to train the teacher.

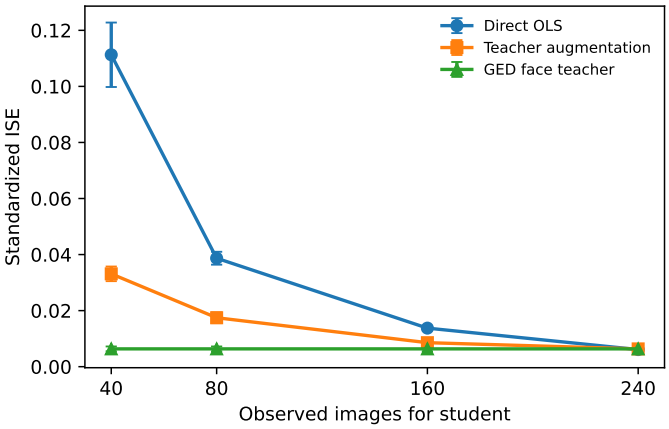

Figure 10: Sample-size sensitivity in the real-image face morphology analysis. GED has the smallest standardized error in the sparse observed-image regimes, while direct OLS attains comparable error once the number of observed images approaches the teacher-training pool size.

G Proofs

G.1 Proof of Proposition 1

Let $X = m + T$, where $T \sim N(0, \tau^2)$. For each fixed s outside a null set, the intercept and slope of the scalar linear projection satisfy the normal equations. Because $\mathrm{Var}_{\nu_m}(X) = \tau^2 > 0$, eliminating the intercept gives

$$\beta_m^\star(s) = \frac{\mathrm{Cov}_{\nu_m}\{X, \mu_\theta(X, s)\}}{\mathrm{Var}_{\nu_m}(X)}$$

for almost every $s \in \mathcal{S}$. The integrability needed for these equations follows from $a, c \in L^2(\mathcal{S})$ and finite Gaussian moments. Because $\mathrm{Var}_{\nu_m}(X) = \tau^2$ and

$$\begin{aligned}\mathrm{Cov}_{\nu_m}\{X, \mu_\theta(X, s)\} &= \mathrm{Cov}_{\nu_m}\{X, a(s)X + c(s)X^2\} \\ &= a(s)\,\mathrm{Var}_{\nu_m}(X) + c(s)\,\mathrm{Cov}_{\nu_m}(X, X^2),\end{aligned}$$

it suffices to compute $\mathrm{Cov}_{\nu_m}(X, X^2)$. With $X = m + T$,

$$\mathrm{Cov}_{\nu_m}(X, X^2) = \mathbb{E}\{(X - m)X^2\} = \mathbb{E}\{T(m + T)^2\} = 2m\tau^2,$$

because $\mathbb{E}T = 0$ and $\mathbb{E}T^3 = 0$. Hence

$$\beta_m^\star(s) = a(s) + 2m\,c(s)$$

for almost every $s \in \mathcal{S}$. The final identity follows by subtracting the two slopes and integrating:

$$\|\beta_{m_1}^\star - \beta_{m_2}^\star\|_{L^2(\mathcal{S})} = 2|m_1 - m_2|\|c\|_{L^2(\mathcal{S})}.$$

Because $m_1 \neq m_2$ and $\|c\|_{L^2(\mathcal{S})} > 0$, the norm is positive. This completes the proof.

G.2 Proof of Proposition 2

We first note that $h_\theta \in L^2(\mathcal{S}; \mathbb{R}^{p+1})$. Indeed, for each coordinate ℓ of Z, Cauchy–Schwarz gives

$$|h_{\theta,\ell}(s)|^2 = |\mathbb{E}_\nu\{Z_\ell \mu_\theta(X, s)\}|^2 \leq \mathbb{E}_\nu(Z_\ell^2)\mathbb{E}_\nu\{\mu_\theta(X, s)^2\}$$

for almost every s, and summing over ℓ and integrating over $\mathcal{S}$ gives the claim. If $v \in \ker(\Gamma_X)$, then $v^\top \Gamma_X v = \mathbb{E}_\nu\{(v^\top Z)^2\} = 0$, and hence $v^\top Z = 0$ ν-almost surely. This gives $v^\top h_\theta(s) = \mathbb{E}_\nu\{v^\top Z \mu_\theta(X, s)\} = 0$ for almost every s. Because Γ_X is symmetric and nonnegative definite, $h_\theta(s) \in \mathrm{range}(\Gamma_X)$ for almost every s.

Put $b_0(s) = \Gamma_X^\dagger h_\theta(s)$. Then $b_0 \in L^2(\mathcal{S}; \mathbb{R}^{p+1})$ and $\Gamma_X b_0(s) = h_\theta(s)$ almost everywhere. For any $b \in L^2(\mathcal{S}; \mathbb{R}^{p+1})$, Tonelli's theorem and the preceding integrability bounds give

$$\begin{aligned}Q_\theta(b) &:= \mathbb{E}_\nu \int_{\mathcal{S}} \{Z^\top b(s) - \mu_\theta(X, s)\}^2 \,\mathrm{d}s \\ &= \int_{\mathcal{S}} \left[b(s)^\top \Gamma_X b(s) - 2b(s)^\top h_\theta(s) + \mathbb{E}_\nu\{\mu_\theta(X, s)^2\}\right] \mathrm{d}s.\end{aligned}$$

Because $\Gamma_X b_0(s) = h_\theta(s)$, it holds that

$$Q_\theta(b) - Q_\theta(b_0) = \int_{\mathcal{S}} \{b(s) - b_0(s)\}^\top \Gamma_X \{b(s) - b_0(s)\} \,\mathrm{d}s \geq 0.$$

Equality holds if and only if $b(s) - b_0(s) \in \ker(\Gamma_X)$ for almost every s, which is equivalent to $\Gamma_X b(s) = h_\theta(s)$ almost everywhere. Thus the minimizer set is exactly the solution set of the normal equations. If b_1 and b_2 are two minimizers, then

$$\mathbb{E}_\nu \int_{\mathcal{S}} \{Z^\top b_1(s) - Z^\top b_2(s)\}^2 \,\mathrm{d}s = \int_{\mathcal{S}} \{b_1(s) - b_2(s)\}^\top \Gamma_X \{b_1(s) - b_2(s)\} \,\mathrm{d}s = 0,$$

which makes the fitted projection unique in $L^2(\nu \times \mathcal{S})$. The Moore–Penrose solution is the unique coefficient representative orthogonal to $\ker(\Gamma_X)$ and hence has minimum $L^2(\mathcal{S}; \mathbb{R}^{p+1})$ norm among all solutions.

If Γ_X is nonsingular, the normal equation gives (8). To derive the slope formula, write $b(s) = (a(s), \beta(s)^\top)^\top$. The normal equations imply

$$\mathbb{E}_\nu\{\mu_\theta(X,s) - a(s) - X^\top\beta(s)\} = 0$$

and

$$\mathbb{E}_\nu\{X[\mu_\theta(X,s) - a(s) - X^\top\beta(s)]\} = 0.$$

The first equation gives $a(s) = \mathbb{E}_\nu\{\mu_\theta(X,s)\} - \bar{x}^\top\beta(s)$. Substituting this into the second equation yields $\Sigma_X\beta(s) = \mathbb{E}_\nu\{(X-\bar{x})\mu_\theta(X,s)\}$. Because Σ_X is nonsingular, this gives (9). This completes the proof.

G.3 Proof of Proposition 5

We first show that $h_c \in L^2(\mathcal{S};\mathbb{R}^p)$. By Cauchy–Schwarz,

$$\|h_c(s)\|_2^2 \le \mathbb{E}_\nu\|X-\bar{x}\|_2^2\,\mathbb{E}_\nu\{\mu_\theta(X,s)^2\}$$

for almost every s, and integration over $\mathcal{S}$ gives the claim. If $v \in \ker(\Sigma_X)$, then $v^\top\Sigma_X v = \mathbb{E}_\nu\big[\{v^\top(X-\bar{x})\}^2\big] = 0$, and hence $v^\top(X-\bar{x}) = 0$ ν-almost surely. This gives $v^\top h_c(s) = 0$ for almost every s, which implies $h_c(s) \in \mathrm{range}(\Sigma_X)$ for almost every s.

For fixed s, the risk in the centered parametrization is

$$\mathbb{E}_\nu\{c(s) + (X-\bar{x})^\top\beta(s) - \mu_\theta(X,s)\}^2.$$

Because $\mathbb{E}_\nu(X-\bar{x}) = 0$, expanding the square and integrating over $\mathcal{S}$ give

$$Q_c(c,\beta) = \int_{\mathcal{S}} \big[\{c(s) - \mathbb{E}_\nu\mu_\theta(X,s)\}^2 + \beta(s)^\top\Sigma_X\beta(s) - 2\beta(s)^\top h_c(s) + r_\theta(s)\big]\,\mathrm{d}s,$$

where

$$r_\theta(s) = \mathbb{E}_\nu\{\mu_\theta(X,s)^2\} - \{\mathbb{E}_\nu\mu_\theta(X,s)\}^2$$

does not depend on (c,β). The first term is minimized by $c(s) = \mathbb{E}_\nu\{\mu_\theta(X,s)\}$. Because $h_c(s) \in \mathrm{range}(\Sigma_X)$ almost everywhere, the slope part is minimized exactly by the solutions of

$$\Sigma_X\beta(s) = h_c(s) \quad \text{for almost every } s.$$

The Moore–Penrose solution $\Sigma_X^\dagger h_c(s)$ is the unique solution orthogonal to $\ker(\Sigma_X)$ and hence has minimum $L^2(\mathcal{S})$ norm among all centered slope representatives.

If β_1 and β_2 solve the centered normal equations, then $\beta_1(s) - \beta_2(s) \in \ker(\Sigma_X)$ for almost every s. Hence

$$\mathbb{E}_\nu\left[\{(X-\bar{x})^\top(\beta_1(s)-\beta_2(s))\}^2\right] = \{\beta_1(s)-\beta_2(s)\}^\top\Sigma_X\{\beta_1(s)-\beta_2(s)\} = 0$$

for almost every s. Integrating over $\mathcal{S}$ gives equality of the fitted projection in $L^2(\nu\times\mathcal{S})$. This completes the proof.

G.4 Proof of Proposition 6

For $0 < |t| < \delta$, the absolute-continuity assumption gives

$$\frac{G_\theta(x+te_j,\zeta) - G_\theta(x,\zeta)}{t} = \int_0^1 \frac{\partial}{\partial x_j}G_\theta(x+rte_j,\zeta)\,\mathrm{d}r \quad \text{in } L^2(\mathcal{S}).$$

The norm of this difference quotient is bounded by the integrable random variable

$$\operatorname*{ess\,sup}_{|u|<\delta}\left\|\frac{\partial}{\partial x_j}G_\theta(x+ue_j,\zeta)\right\|_{L^2(\mathcal{S})}.$$

Thus, the difference quotient is Bochner integrable, and taking expectations gives

$$\frac{\mu_\theta(x+te_j) - \mu_\theta(x)}{t} = \mathbb{E}\left[\int_0^1 \frac{\partial}{\partial x_j}G_\theta(x+rte_j,\zeta)\,\mathrm{d}r\right].$$

As $t \to 0$, differentiability at x gives convergence of the difference quotient in $L^2(\mathcal{S})$ to $\partial G_\theta(x,\zeta)/\partial x_j$ for almost every ζ. Dominated convergence for Bochner integrals then gives

$$\left\|\frac{\mu_\theta(x+te_j)-\mu_\theta(x)}{t} - \mathbb{E}\left\{\frac{\partial}{\partial x_j}G_\theta(x,\zeta)\right\}\right\|_{L^2(\mathcal{S})} \to 0.$$

This proves the derivative identity in $L^2(\mathcal{S})$ and completes the proof.

G.5 Proof of Proposition 8

Put $\delta(X,s) = \mu_a(X,s) - \mu_b(X,s)$ and $h_\delta(s) = \mathbb{E}_\nu\{Z\delta(X,s)\}$. By Proposition 2, the minimum-norm coefficient representative of the difference between the two fitted projections is $\Gamma_X^\dagger h_\delta(s)$. For almost every s, the vector $h_\delta(s)$ belongs to $\mathrm{range}(\Gamma_X)$. For any $v \in \mathbb{R}^{p+1}$ with $v^\top \Gamma_X v \le 1$, Cauchy–Schwarz gives

$$\{v^\top h_\delta(s)\}^2 = \{\mathbb{E}_\nu(v^\top Z\,\delta(X,s))\}^2 \le v^\top \Gamma_X v\, \mathbb{E}_\nu\{\delta(X,s)^2\} \le \mathbb{E}_\nu\{\delta(X,s)^2\}.$$

Because $h_\delta(s) \in \mathrm{range}(\Gamma_X)$,

$$h_\delta(s)^\top \Gamma_X^\dagger h_\delta(s) = \sup_{v:\, v^\top \Gamma_X v \le 1} \{v^\top h_\delta(s)\}^2 \le \mathbb{E}_\nu\{\delta(X,s)^2\}.$$

Thus,

$$\begin{aligned}\|m_a^\star - m_b^\star\|_{L^2(\nu\times\mathcal{S})}^2 &= \int_{\mathcal{S}} h_\delta(s)^\top \Gamma_X^\dagger h_\delta(s)\,\mathrm{d}s \\ &\le \int_{\mathcal{S}} \mathbb{E}_\nu\{\delta(X,s)^2\}\,\mathrm{d}s = \Delta_{ab}^2.\end{aligned}$$

For the coefficient bound, use the spectral decomposition of Γ_X on its range. Because $\Gamma_X^{\dagger 2} \preceq \kappa_+^{-1}\Gamma_X^\dagger$ on $\mathrm{range}(\Gamma_X)$,

$$\begin{aligned}\left\|b_{a,\min}^\star - b_{b,\min}^\star\right\|_{L^2(\mathcal{S})}^2 &= \int_{\mathcal{S}} h_\delta(s)^\top \Gamma_X^{\dagger 2} h_\delta(s)\,\mathrm{d}s \\ &\le \kappa_+^{-1} \int_{\mathcal{S}} h_\delta(s)^\top \Gamma_X^\dagger h_\delta(s)\,\mathrm{d}s \le \kappa_+^{-1}\Delta_{ab}^2.\end{aligned}$$

If $\lambda_{\min}(\Gamma_X) \ge \kappa > 0$, then $\kappa_+ \ge \kappa$ and the coefficient representative is unique. The slope map is obtained by deleting the intercept component, and hence its $L^2(\mathcal{S})$ norm is no larger than the augmented coefficient norm. This completes the proof.

G.6 Proof of Proposition 3

Put $W = X - \bar{x}$ and

$$h_c(s) = \mathbb{E}_\nu\{W\mu_\theta(X,s)\}.$$

By Cauchy–Schwarz,

$$\|h_c(s)\|_2^2 \le \mathbb{E}_\nu\|W\|_2^2\, \mathbb{E}_\nu\{\mu_\theta(X,s)^2\}$$

for almost every s. Because $\mathbb{E}_\nu\|W\|_2^2 < \infty$ and $\mu_\theta \in L^2(\nu\times\mathcal{S})$, it follows that $\int_{\mathcal{S}} \|h_c(s)\|_2^2\,\mathrm{d}s < \infty$. The Stein-domain condition yields

$$h_c(s) = \mathbb{E}_\nu\{T_\nu(X)\nabla_x\mu_\theta(X,s)\}$$

for almost every s, which establishes the square integrability of the expectation on the right-hand side.

Let $\mathcal{S}_0 \subseteq \mathcal{S}$ be a full-measure set on which the centered normal equations and the Stein identity hold. Fix $s \in \mathcal{S}_0$. By Proposition 5,

$$\beta_{\theta,\min}^c(s) = \Sigma_X^\dagger h_c(s).$$

Substituting the Stein identity for $h_c(s)$ proves (11). If Σ_X is nonsingular, then $\Sigma_X^\dagger = \Sigma_X^{-1}$ and the centered representative is the unique slope map. If $X \sim \mathcal{N}(\bar{x}, \Sigma_X)$ with nonsingular Σ_X, choose the canonical Gaussian Stein kernel $T_\nu(X) = \Sigma_X$. Substituting this choice into (11) gives $\beta_\theta^\star(s) = \mathbb{E}_\nu\{\nabla_x\mu_\theta(X,s)\}$. Taking components gives (12). Because $s \in \mathcal{S}_0$ was arbitrary, the identities hold for almost every $s \in \mathcal{S}$. This completes the proof.

G.7 Proof of Proposition 4

We first note that $R_{\theta,\tau}$ is fixed when differentiating the mean parameters. Fix $j \in \{0,\dots,p\}$ and let $h \in \mathcal{H}_j$. Consider the path ϕ_t that replaces β_j by $\beta_j + th$ and leaves all other parameters fixed. By the definition of $\mathcal{H}_j$, this path is admissible for all sufficiently small $|t|$. Along this path,

$$m_{\phi_t}(x) = m_\phi(x) + tZ_j h, \qquad u_{\phi_t,\tau}(y,x) = -C_{\phi,\tau}^{-1}\{y - m_\phi(x) - tZ_j h\}.$$

Thus,

$$r_{\phi_t,\tau}(y,x) = r_{\phi,\tau}(y,x) - tZ_j C_{\phi,\tau}^{-1} h.$$

The integrability assumptions and $\mathcal{L}_{\text{score}}(\phi) < \infty$ permit expansion of the quadratic objective along this one-dimensional path:

$$\begin{aligned}\mathcal{L}_{\text{score}}(\phi_t) &= \mathbb{E}\left[\left\| r_{\phi,\tau}(\tilde{Y},X) - tZ_j C_{\phi,\tau}^{-1} h \right\|_{W_\tau}^2\right] \\ &= \mathcal{L}_{\text{score}}(\phi) - 2t\mathbb{E}\left[Z_j \left\langle C_{\phi,\tau}^{-1} h, W_\tau r_{\phi,\tau}(\tilde{Y},X)\right\rangle\right] + t^2\mathbb{E}\left[Z_j^2 \left\| C_{\phi,\tau}^{-1} h\right\|_{W_\tau}^2\right].\end{aligned}$$

The scalar quadratic polynomial in t has a local minimum at zero, and hence its linear coefficient equals zero. This proves (15). This completes the proof.

G.8 Proof of Proposition 7

Because (15) holds at ϕ and $\mathcal{H}_j = L^2(\mathcal{S})$, any $L^2(\mathcal{S})$ representer G_j of the tangent functional satisfies $\langle h, G_j\rangle = 0$ for every $h \in L^2(\mathcal{S})$. Taking $h = G_j$ gives $G_j = 0$ in $L^2(\mathcal{S})$.

It remains to verify the stated sufficient conditions for the vector representer. If, under the score-space convention in the statement, $C_{\phi,\tau}^{-1}$ and W_τ are bounded self-adjoint operators and

$$\mathbb{E}\left[|Z_j| \left\| C_{\phi,\tau}^{-1} W_\tau r_{\phi,\tau}(\tilde{Y},X)\right\|_{L^2(\mathcal{S})}\right] < \infty,$$

then the Bochner expectation

$$G_j = \mathbb{E}\left[Z_j C_{\phi,\tau}^{-1} W_\tau r_{\phi,\tau}(\tilde{Y},X)\right]$$

exists in $L^2(\mathcal{S})$. For any $h \in L^2(\mathcal{S})$, self-adjointness gives

$$\begin{aligned}\mathbb{E}\left[Z_j \left\langle C_{\phi,\tau}^{-1} h, W_\tau r_{\phi,\tau}(\tilde{Y},X)\right\rangle\right] &= \mathbb{E}\left[Z_j \left\langle h, C_{\phi,\tau}^{-1} W_\tau r_{\phi,\tau}(\tilde{Y},X)\right\rangle\right] \\ &= \langle h, G_j\rangle.\end{aligned}$$

For every $h \in L^2(\mathcal{S})$,

$$\mathbb{E}\left[Z_j^2 \left\| C_{\phi,\tau}^{-1} h\right\|_{W_\tau}^2\right] \le \mathbb{E}Z_j^2 \left\|W_\tau\right\|_{\text{op}} \left\|C_{\phi,\tau}^{-1}\right\|_{\text{op}}^2 \|h\|_{L^2(\mathcal{S})}^2 < \infty.$$

Hence G_j represents the tangent functional in $L^2(\mathcal{S})$, and the preceding paragraph gives $G_j = 0$. This is exactly the vector equation (21). Equality in $L^2(\mathcal{S})$ implies equality for almost every $s \in \mathcal{S}$ after choosing representatives.

Finally, assume that $\mathcal{H}_j = L^2(\mathcal{S})$ for all $j = 0, \ldots, p$, that the objective is finite for every finite perturbation, and that (15) holds at ϕ. Let

$$\Delta m(X) = \sum_{j=0}^{p} Z_j h_j$$

be an arbitrary finite perturbation of the mean parameters. The residual under the perturbed mean is $r_{\phi,\tau}(\tilde{Y},X) - C_{\phi,\tau}^{-1}\Delta m(X)$, which gives

$$\begin{aligned}&\mathcal{L}_{\text{score}}(\phi + \Delta) - \mathcal{L}_{\text{score}}(\phi) \\ &\quad = -2\mathbb{E}\left[\left\langle C_{\phi,\tau}^{-1}\Delta m(X), W_\tau r_{\phi,\tau}(\tilde{Y},X)\right\rangle\right] + \mathbb{E}\left[\left\| C_{\phi,\tau}^{-1}\Delta m(X)\right\|_{W_\tau}^2\right].\end{aligned}$$

By linearity and (15), the first term on the right-hand side is zero. The second term is nonnegative because W_τ is positive semidefinite. Thus, no finite perturbation can decrease the objective, which proves global minimality over the affine mean class. This completes the proof.

G.9 Auxiliary lemmas for Theorem 1

Recall that Z, Γ_X, $V_\theta(X)$, E_m, $A_q^\star$, and $U_\theta(X)$ are defined before Assumption 2. For query covariates $X_1, \ldots, X_M$, define

$$\Gamma_M = \frac{1}{M}\sum_{m=1}^{M} Z_m Z_m^\top, \qquad R_M = \frac{1}{M}\sum_{m=1}^{M} Z_m U_\theta(X_m)^\top,$$

where $Z_m = (1, X_m^\top)^\top$. Also define

$$N_{MK} = \frac{1}{M}\sum_{m=1}^{M} Z_m E_m^\top.$$

When Γ_M is nonsingular, let $A_M^\star$ and $\widetilde{A}_{MK}$ denote the exact-query and Monte Carlo coefficient matrices,

$$A_M^\star = \Gamma_M^{-1}\left\{\frac{1}{M}\sum_{m=1}^{M} Z_m V_\theta(X_m)^\top\right\}, \quad \widetilde{A}_{MK} = \Gamma_M^{-1}\left\{\frac{1}{M}\sum_{m=1}^{M} Z_m \widehat{V}_\theta(X_m)^\top\right\}.$$

For nonnegative numbers r_M and r_K, define the event $\mathcal{E}(r_M, r_K)$ by

$$\lambda_{\min}(\Gamma_M) \geq \kappa/2, \qquad \|R_M\|_F \leq r_M, \qquad \|N_{MK}\|_F \leq r_K.$$

The high-probability part of Theorem 1 uses Lemma 1, Lemma 2, and the deterministic decomposition in (17).

Lemma 1. *Suppose that $\mathcal{E}(r_M, r_K)$ holds. Then, it holds that*

$$\left\|A_M^\star - A_q^\star\right\|_F \leq 2\kappa^{-1} r_M, \qquad \left\|\widetilde{A}_{MK} - A_M^\star\right\|_F \leq 2\kappa^{-1} r_K.$$

Proof. We first note that the population normal equation gives $\Gamma_X A_q^\star = \mathbb{E}_\nu\{Z V_\theta(X)^\top\}$, and the definition of U_θ gives $V_\theta(X_m) = A_q^{\star\top} Z_m + U_\theta(X_m)$. Hence,

$$\Gamma_M(A_M^\star - A_q^\star) = \frac{1}{M}\sum_{m=1}^{M} Z_m\{V_\theta(X_m) - A_q^{\star\top} Z_m\}^\top = R_M.$$

On $\mathcal{E}(r_M, r_K)$, $\|\Gamma_M^{-1}\|_{\mathrm{op}} \leq 2\kappa^{-1}$, and hence $\|A_M^\star - A_q^\star\|_F \leq 2\kappa^{-1} r_M$. Similarly,

$$\Gamma_M(\widetilde{A}_{MK} - A_M^\star) = \frac{1}{M}\sum_{m=1}^{M} Z_m\{\widehat{V}_\theta(X_m) - V_\theta(X_m)\}^\top = N_{MK},$$

which gives the second inequality. This completes the proof. □

For $\eta \in (0,1)$, let

$$r_M(\eta) = C B_\theta \sqrt{\frac{d_{\mathrm{eff}} + \log(1/\eta)}{M}}, \qquad r_K(\eta) = C\sigma_\theta \sqrt{\frac{d_{\mathrm{eff}} + \log(1/\eta)}{MK}}.$$

Lemma 2. *Suppose that Assumption 2 holds. If $M \geq C\{d_{\mathrm{eff}} + \log(1/\eta)\}$, then $\mathcal{E}\{r_M(\eta), r_K(\eta)\}$ holds with probability at least $1 - \eta$.*

Proof. We first control the empirical design matrix. Put $W_m = \Gamma_X^{-1/2} Z_m$. Assumption 2 implies $\mathbb{E}(W_m W_m^\top) = I_{p+1}$ and gives a uniform sub-Gaussian norm bound for W_m. Let $\mathcal{N}_1$ be a $1/4$-net of the unit sphere in $\mathbb{R}^{p+1}$ with $|\mathcal{N}_1| \leq 9^{p+1}$. For any fixed $u \in \mathcal{N}_1$, the random variables $u^\top(W_m W_m^\top - I_{p+1})u$ are centered sub-exponential variables with an absolute sub-exponential norm bound. Bernstein's inequality and the union bound over $\mathcal{N}_1$ imply that, with probability at least $1 - \eta/3$,

$$\sup_{u \in \mathcal{N}_1}\left|u^\top\left\{\frac{1}{M}\sum_{m=1}^{M}(W_m W_m^\top - I_{p+1})\right\}u\right| \leq \frac{1}{8},$$

provided $M \geq C\{p + \log(1/\eta)\}$. Put $A = M^{-1}\sum_{m=1}^{M}(W_m W_m^\top - I_{p+1})$. For any unit vector v, choose $u \in \mathcal{N}_1$ with $\|u - v\|_2 \leq 1/4$. Because A is self-adjoint, it holds that

$$|v^\top A v| \leq |u^\top A u| + 2\|A\|_{\mathrm{op}}\|u - v\|_2 + \|A\|_{\mathrm{op}}\|u - v\|_2^2 \leq |u^\top A u| + \frac{9}{16}\|A\|_{\mathrm{op}}.$$

Taking the supremum over unit vectors gives $\|A\|_{\mathrm{op}} \leq (16/7)\sup_{u\in\mathcal{N}_1}|u^\top A u| \leq 2/7 < 1/2$. Hence,

$$\left\|\Gamma_X^{-1/2}\Gamma_M\Gamma_X^{-1/2} - I_{p+1}\right\|_{\mathrm{op}} \leq \frac{1}{2}$$

on the same event. Hence, $\lambda_{\min}(\Gamma_M) \geq \kappa/2$.

We next control the exact-query residual term. For each unit vector $u \in \mathbb{R}^{d_{\mathrm{eff}}}$, let $H_u \in \mathbb{R}^{(p+1)\times q}$ be its matrix form with $\|H_u\|_F = 1$, for which $u^\top \mathrm{vec}\{ZU_\theta(X)^\top\} = \langle H_u, ZU_\theta(X)^\top\rangle$. The definition of $A_q^\star$ gives

$$\mathbb{E}_\nu\{ZU_\theta(X)^\top\} = \mathbb{E}_\nu\{ZV_\theta(X)^\top\} - \Gamma_X A_q^\star = 0,$$

The equality above shows that $u^\top \mathrm{vec}\{ZU_\theta(X)^\top\}$ is centered. Assumption 2 gives the sub-exponential bound

$$\|u^\top \mathrm{vec}\{ZU_\theta(X)^\top\}\|_{\psi_1} \leq CB_\theta.$$

Let $\mathcal{N}_2$ be a $1/4$-net of the unit sphere in $\mathbb{R}^{d_{\mathrm{eff}}}$ with $|\mathcal{N}_2| \leq 9^{d_{\mathrm{eff}}}$. Applying Bernstein's inequality to each $u \in \mathcal{N}_2$ and then using the union bound gives, with probability at least $1 - \eta/3$,

$$\sup_{u\in\mathcal{N}_2}\left|u^\top \mathrm{vec}(R_M)\right| \leq CB_\theta\sqrt{\frac{d_{\mathrm{eff}} + \log(1/\eta)}{M}}.$$

For any unit vector $v \in \mathbb{R}^{d_{\mathrm{eff}}}$, choose $u \in \mathcal{N}_2$ with $\|u - v\|_2 \leq 1/4$. Then, $|v^\top \mathrm{vec}(R_M)| \leq |u^\top \mathrm{vec}(R_M)| + \|\mathrm{vec}(R_M)\|_2/4$. Taking the supremum over v gives

$$\|R_M\|_F = \left\|\frac{1}{M}\sum_{m=1}^{M}\mathrm{vec}\{Z_m U_\theta(X_m)^\top\}\right\|_2 \leq CB_\theta\sqrt{\frac{d_{\mathrm{eff}} + \log(1/\eta)}{M}}$$

where the linear term in Bernstein's inequality is dominated by the quadratic term under the condition $M \geq C\{d_{\mathrm{eff}} + \log(1/\eta)\}$.

It remains to control the Monte Carlo teacher-query term. For the same matrix form H_u, the sampling protocol in Assumption 2 yields $\mathbb{E}(E_m \mid X_m) = 0$, which gives

$$\mathbb{E}\{\langle H_u, Z_m E_m^\top\rangle \mid X_m\} = \langle H_u, Z_m\mathbb{E}(E_m \mid X_m)^\top\rangle = 0.$$

Thus $u^\top \mathrm{vec}(Z_m E_m^\top) = \langle H_u, Z_m E_m^\top\rangle$ is centered. The same remark gives independence across m for these centered variables. Assumption 2 also gives $\|u^\top \mathrm{vec}(Z_m E_m^\top)\|_{\psi_1} \leq C\sigma_\theta/\sqrt{K}$. Applying Bernstein's inequality to this scalar for each $u \in \mathcal{N}_2$ and taking the union bound over $\mathcal{N}_2$ yield

$$\sup_{u\in\mathcal{N}_2}\left|u^\top \mathrm{vec}(N_{MK})\right| \leq C\sigma_\theta\sqrt{\frac{d_{\mathrm{eff}} + \log(1/\eta)}{MK}}.$$

For any unit $v \in \mathbb{R}^{d_{\mathrm{eff}}}$, choose $u \in \mathcal{N}_2$ with $\|u - v\|_2 \leq 1/4$. Then,

$$|v^\top \mathrm{vec}(N_{MK})| \leq |u^\top \mathrm{vec}(N_{MK})| + \frac{1}{4}\|\mathrm{vec}(N_{MK})\|_2.$$

Taking the supremum over v gives

$$\|N_{MK}\|_F \leq C\sigma_\theta\sqrt{\frac{d_{\mathrm{eff}} + \log(1/\eta)}{MK}}$$

with probability at least $1 - \eta/3$. Combining the three events by the union bound proves the lemma. This completes the proof. □

Lemma 3. *Suppose that the definitions before Assumption 2 hold. Let $L_Z, L_U, L_E > 0$ satisfy $L_Z L_U \leq CB_\theta$ and $L_Z L_E \leq C\sigma_\theta$. If $\|Z\|_2 \leq L_Z$ almost surely and $\|Z_m\|_2 \leq L_Z$ almost surely for every query point m,*

$$\left\|L_U^{-1}\|U_\theta(X)\|_2\right\|_{\psi_1} \leq C, \qquad \left\|\sqrt{K}\,L_E^{-1}\|E_m\|_2\right\|_{\psi_1} \leq C$$

for every m, then the two projected tail inequalities in Assumption 2 hold.

Proof. Let $H \in \mathbb{R}^{(p+1)\times q}$ be deterministic with $\|H\|_F = 1$. By the Cauchy–Schwarz inequality,

$$|\langle H, ZU_\theta(X)^\top\rangle| \le \|Z\|_2 \|U_\theta(X)\|_2 \le L_Z \|U_\theta(X)\|_2.$$

The homogeneity of the Orlicz norm and $L_Z L_U \le C B_\theta$ imply

$$\left\| B_\theta^{-1} \langle H, ZU_\theta(X)^\top\rangle \right\|_{\psi_1} \le C.$$

The same argument gives, for each m,

$$|\langle H, Z_m E_m^\top\rangle| \le L_Z \|E_m\|_2,$$

and hence

$$\left\| \sqrt{K}\, \sigma_\theta^{-1} \langle H, Z_m E_m^\top\rangle \right\|_{\psi_1} \le C.$$

This completes the proof. □

G.10 Proof of Theorem 1

We first prove the deterministic statement. Assumption 1 identifies each finite-dimensional coefficient map with a coefficient matrix in $\mathbb{R}^{(p+1)\times q}$, and the orthonormality of ψ makes the $L^2(\mathcal{S})$ norm of an augmented map equal to the Frobenius norm of its coefficient matrix. The slope-map norm is no larger than the corresponding augmented Frobenius norm, because the slope matrix is obtained by deleting the intercept row.

On the event

$$\lambda_{\min}(\Gamma_M) \ge \underline{\kappa}_M, \qquad \|R_M\|_F \le \rho_M, \qquad \|N_{MK}\|_F \le \rho_K,$$

we have $\|\Gamma_M^{-1}\|_{\mathrm{op}} \le \underline{\kappa}_M^{-1}$. The population finite-basis normal equation and the definition of U_θ give

$$\Gamma_M(A_M^\star - A_q^\star) = \frac{1}{M}\sum_{m=1}^M Z_m\{V_\theta(X_m) - A_q^{\star\top} Z_m\}^\top = R_M.$$

Thus

$$\|A_M^\star - A_q^\star\|_F \le \underline{\kappa}_M^{-1} \rho_M.$$

Similarly,

$$\Gamma_M(\widetilde{A}_{MK} - A_M^\star) = \frac{1}{M}\sum_{m=1}^M Z_m\{\widehat{V}_\theta(X_m) - V_\theta(X_m)\}^\top = N_{MK},$$

and hence

$$\|\widetilde{A}_{MK} - A_M^\star\|_F \le \underline{\kappa}_M^{-1} \rho_K.$$

Passing from augmented coefficient matrices to slope maps can only decrease the norm, which gives

$$\left\|\beta_{\theta,M}^\star - \beta_{\theta,q}^\star\right\|_{L^2(\mathcal{S})} \le \underline{\kappa}_M^{-1}\rho_M, \qquad \left\|\widetilde{\beta}_{\theta,MK} - \beta_{\theta,M}^\star\right\|_{L^2(\mathcal{S})} \le \underline{\kappa}_M^{-1}\rho_K.$$

For the population bias term, the definition of d_θ gives exactly

$$\left\|\beta_\theta^\star - \beta_{\mu_0}^\dagger\right\|_{L^2(\mathcal{S})} = d_\theta.$$

The basis-approximation term is exactly ρ_q by (18), and the optimization condition gives $\|\widehat{\beta} - \widetilde{\beta}_{\theta,MK}\|_{L^2(\mathcal{S})} \le \epsilon_{\mathrm{op}}$. Substituting these bounds into (17) proves (19).

It remains to prove the high-probability consequence. Under Assumption 2, Lemma 2 gives, with probability at least $1-\eta$,

$$\lambda_{\min}(\Gamma_M) \ge \kappa/2, \qquad \|R_M\|_F \le C B_\theta \sqrt{\frac{d_{\mathrm{eff}} + \log(1/\eta)}{M}},$$

and

$$\|N_{MK}\|_F \le C\sigma_\theta \sqrt{\frac{d_{\mathrm{eff}} + \log(1/\eta)}{MK}}.$$

Applying (19) with $\underline{\kappa}_M = \kappa/2$ and absorbing numerical constants into C gives (20). If the optional $L^2(\nu \times \mathcal{S})$ teacher-error scale is used, Proposition 8 and $\lambda_{\min}(\Gamma_X) \ge \kappa$ in Assumption 2 gives $d_\theta \le \kappa^{-1/2}\delta_\theta$, which proves the weaker δ_θ version on the same event. This completes the proof.

G.11 Proof of Corollary 1

Fix $\eta \in (0,1)$. For all sufficiently large problem sizes, the sample-size condition $M \geq C\{d_{\text{eff}} + \log(1/\eta)\}$ holds because $d_{\text{eff}}/M \to 0$. Theorem 1 applies with probability at least $1-\eta$. On this event, the deterministic terms d_θ, ρ_q, and ϵ_{op} converge to zero by assumption. Because $d_{\text{eff}} \geq 1$ and η is fixed, the product-rate conditions imply

$$\kappa^{-1} B_\theta \sqrt{\frac{d_{\text{eff}} + \log(1/\eta)}{M}} \to 0, \qquad \kappa^{-1} \sigma_\theta \sqrt{\frac{d_{\text{eff}} + \log(1/\eta)}{MK}} \to 0.$$

Thus the right-hand side of (20) converges to zero on an event whose probability is at least $1-\eta$. For every $\varepsilon > 0$,

$$\limsup \mathbb{P}\left(\left\| \widehat{\beta} - \beta^\dagger_{\mu_0} \right\|_{L^2(\mathcal{S})} > \varepsilon \right) \leq \eta.$$

Because η is arbitrary, the convergence holds in probability. If the teacher bias is controlled by δ_θ, the result follows from $d_\theta \leq \kappa^{-1/2} \delta_\theta$. This completes the proof.

G.12 Proof of Theorem 2

We first note that the joint distribution in (22) is a Gaussian matrix regression experiment. Let $\Gamma_M = Q \Lambda Q^\top$ be an eigendecomposition, where Q is orthogonal and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_r)$ with $\lambda_a > 0$ for $a = 1, \dots, r$. Put

$$\delta_a^2 = \frac{\tau_K^2}{32 M \lambda_a}, \qquad a = 1, \dots, r,$$

and let $D = \operatorname{diag}(\delta_1, \dots, \delta_r)$. For each sign matrix $\Xi \in \{-1,1\}^{r \times q}$, define

$$A^\Xi = Q D \Xi.$$

The radius condition gives

$$\|A^\Xi\|_F^2 = \|D\Xi\|_F^2 = q \sum_{a=1}^r \delta_a^2 = \frac{q \operatorname{tr}(\Gamma_M^{-1})}{32M} \tau_K^2 \leq R^2,$$

and hence all matrices A^Ξ belong to $\mathbb{B}_F(R)$.

Let $\mathbb{P}_\Xi = \mathbb{P}_{A^\Xi}$, and let $\mathbb{E}_\Xi$ denote the corresponding expectation. For a coordinate $(a,b) \in \{1, \dots, r\} \times \{1, \dots, q\}$, let Ξ^{ab} be obtained from Ξ by changing only the sign of the (a,b) entry. Because the observations are independent Gaussian vectors with covariance $\tau_K^2 I_q$, the Kullback–Leibler divergence between neighboring vertices satisfies

$$\begin{aligned} D_{\text{KL}}(\mathbb{P}_\Xi \,\|\, \mathbb{P}_{\Xi^{ab}}) &= \frac{1}{2\tau_K^2} \sum_{m=1}^M \|(A^\Xi - A^{\Xi^{ab}})^\top Z_m\|_2^2 \\ &= \frac{2\delta_a^2}{\tau_K^2} \sum_{m=1}^M (e_a^\top Q^\top Z_m)^2 \\ &= \frac{2M\lambda_a \delta_a^2}{\tau_K^2} = \frac{1}{16}, \end{aligned}$$

where e_a is the ath Euclidean basis vector in $\mathbb{R}^r$. Pinsker's inequality gives

$$\|\mathbb{P}_\Xi - \mathbb{P}_{\Xi^{ab}}\|_{\text{TV}} \leq \left\{ \frac{1}{2} D_{\text{KL}}(\mathbb{P}_\Xi \,\|\, \mathbb{P}_{\Xi^{ab}}) \right\}^{1/2} \leq \frac{1}{4}.$$

Fix an arbitrary estimator $\widehat{A}$ and put $\widetilde{A} = Q^\top \widehat{A}$. Define the sign estimator $\widehat{\Xi}_{ab} = 1$ if $\widetilde{A}_{ab} \geq 0$ and $\widehat{\Xi}_{ab} = -1$ if $\widetilde{A}_{ab} < 0$. If $\widehat{\Xi}_{ab} \neq \Xi_{ab}$, then $(\widetilde{A}_{ab} - \delta_a \Xi_{ab})^2 \geq \delta_a^2$. Hence,

$$\begin{aligned} \sup_{A \in \mathbb{B}_F(R)} \mathbb{E}_A \|\widehat{A} - A\|_F^2 &\geq 2^{-d_{\text{eff}}} \sum_{\Xi \in \{-1,1\}^{r \times q}} \mathbb{E}_\Xi \|\widehat{A} - A^\Xi\|_F^2 \\ &= 2^{-d_{\text{eff}}} \sum_{\Xi} \mathbb{E}_\Xi \|\widetilde{A} - D\Xi\|_F^2 \\ &\geq \sum_{a=1}^r \sum_{b=1}^q \delta_a^2 2^{-d_{\text{eff}}} \sum_{\Xi} \mathbb{P}_\Xi(\widehat{\Xi}_{ab} \neq \Xi_{ab}). \end{aligned}$$

For each fixed (a, b), grouping sign matrices into pairs Ξ, Ξ^{ab} with $\Xi_{ab} = 1$ gives

$$2^{-d_{\text{eff}}} \sum_{\Xi} \mathbb{P}_{\Xi}(\widehat{\Xi}_{ab} \neq \Xi_{ab}) \geq \frac{1}{2} \inf_{\Xi:\Xi_{ab}=1} \{1 - \|\mathbb{P}_{\Xi} - \mathbb{P}_{\Xi^{ab}}\|_{\text{TV}}\} \geq \frac{3}{8}.$$

Combining the preceding inequalities, we obtain

$$\sup_{A \in \mathbb{B}_F(R)} \mathbb{E}_A \|\widehat{A} - A\|_F^2 \geq \frac{3}{8} \sum_{a=1}^{r} \sum_{b=1}^{q} \delta_a^2 \geq \frac{q}{4} \sum_{a=1}^{r} \delta_a^2 = \frac{q \operatorname{tr}(\Gamma_M^{-1})}{128M} \tau_K^2.$$

Because the lower bound holds for every estimator $\widehat{A}$, taking the infimum over $\widehat{A}$ proves the first assertion. Finally, $\operatorname{tr}(\Gamma_M^{-1}) \geq r/\lambda_{\max}(\Gamma_M) \geq r/\overline{\kappa}$, and $rq = d_{\text{eff}}$. Substituting $\tau_K^2 = B_\theta^2 + \sigma_\theta^2/K$ gives the second assertion. This completes the proof.

G.13 Proof of Theorem 3

We first note that, for the linear student, $\partial m_\phi(x, s)/\partial x_j = \beta_j(s)$. The population effect loss separates over j and can be written as

$$\mathcal{L}_{\text{eff}}(\beta) = \sum_{j=1}^{p} \int_{\mathcal{S}} \mathbb{E}_\nu \left[\beta_j(s) - \frac{\partial}{\partial x_j} \mu_\theta(X, s) \right]^2 \mathrm{d}s.$$

Let $\bar{g}_{\theta,j}(s) = \mathbb{E}_\nu\{\partial \mu_\theta(X, s)/\partial x_j\}$, which belongs to $L^2(\mathcal{S})$ by Jensen's inequality and the assumed square integrability. Hence the choice $\beta_j = \bar{g}_{\theta,j}$ for $j = 1, \ldots, p$ and any $\beta_0 \in L^2(\mathcal{S})$ is admissible and gives a finite value of the loss. Thus the minimizer set is nonempty. Expanding the square and using Fubini's theorem give

$$\mathcal{L}_{\text{eff}}(\beta) = \sum_{j=1}^{p} \int_{\mathcal{S}} \left[\{\beta_j(s) - \bar{g}_{\theta,j}(s)\}^2 + \operatorname{Var}_\nu \left\{ \frac{\partial}{\partial x_j} \mu_\theta(X, s) \right\} \right] \mathrm{d}s.$$

The second term does not depend on β_j, and the first term is minimized uniquely in $L^2(\mathcal{S})$ by $\beta_j(s) = \bar{g}_{\theta,j}(s)$ for almost every s. The loss contains no term involving β_0; every intercept function gives the same minimum. This proves both necessity and sufficiency for the stated slope equalities. This completes the proof.

G.14 Proof of Proposition 9

If $\lambda_\mu = 0$, then $\lambda_g > 0$ and the objective reduces to $\lambda_g \mathcal{L}_{\text{eff}}(\phi)$. Theorem 3 gives the slope $\bar{g}_\theta(s)$ for almost every $s \in \mathcal{S}$. It remains to consider the case $\lambda_\mu > 0$.

We first handle the intercept in the mean loss. For a fixed slope $\beta(s)$ and $\lambda_\mu > 0$, the minimizing intercept satisfies

$$\beta_0(s) = \mathbb{E}_\nu\{\mu_\theta(X, s)\} - \bar{x}^\top \beta(s).$$

Substituting this expression into the mean loss gives, up to a term independent of β,

$$\int_{\mathcal{S}} \{\beta(s) - \beta_{\text{mean}}^c(s)\}^\top \Sigma_X \{\beta(s) - \beta_{\text{mean}}^c(s)\} \, \mathrm{d}s.$$

If $\lambda_g = 0$, this quadratic is minimized by all slopes satisfying $\Sigma_X\{\beta(s) - \beta_{\text{mean}}^c(s)\} = 0$ for almost every $s \in \mathcal{S}$. These slopes induce the same fitted centered map, and Proposition 5 selects β_{mean}^c as the minimum-norm centered representative. When Σ_X is nonsingular, this representative is the unique slope. It remains to consider $\lambda_g > 0$. By the proof of Theorem 3, the effect loss is, up to a term independent of β,

$$\int_{\mathcal{S}} \|\beta(s) - \bar{g}_\theta(s)\|_2^2 \, \mathrm{d}s.$$

For almost every $s \in \mathcal{S}$, the combined population target minimizes the quadratic function

$$\lambda_\mu \{\beta - \beta_{\text{mean}}^c(s)\}^\top \Sigma_X \{\beta - \beta_{\text{mean}}^c(s)\} + \lambda_g \|\beta - \bar{g}_\theta(s)\|_2^2.$$

The first-order condition is

$$(\lambda_\mu \Sigma_X + \lambda_g I_p)\beta = \lambda_\mu \Sigma_X \beta_{\text{mean}}^c(s) + \lambda_g \bar{g}_\theta(s).$$

Because Σ_X is positive semidefinite and $\lambda_g > 0$, $\lambda_\mu \Sigma_X + \lambda_g I_p$ is positive definite, and the solution is the stated expression for $\beta_\lambda(s)$. If $\beta_{\text{mean}}^c(s) = \bar{g}_\theta(s)$, substituting this common value into the formula gives the final claim. This completes the proof.

G.15 Proof of Proposition 10

Because $V_0^{-1} = \Psi^{-1} + G_0$, we have $I_p - H_0 = V_0(V_0^{-1} - G_0) = V_0\Psi^{-1}$. The matrix V_0 is positive definite because $\Psi^{-1} \succ 0$ and $G_0 \succeq 0$. The matrix H_0 is similar to $V_0^{1/2} G_0 V_0^{1/2} = I_p - V_0^{1/2}\Psi^{-1}V_0^{1/2}$, which is positive semidefinite because $G_0 \succeq 0$ and has eigenvalues smaller than one because $V_0^{1/2}\Psi^{-1}V_0^{1/2} \succ 0$. Hence the eigenvalues of H_0 lie in $[0, 1)$, and $I_p + \omega H_0$ is invertible for $\omega \geq 0$.

The marginal distribution of Y is $N(0, \Sigma_Y)$ with $\Sigma_Y = B_0\Psi B_0^\top + \Sigma_\varepsilon$, and the Woodbury identity gives $\Sigma_Y^{-1} = \Sigma_\varepsilon^{-1} - \Sigma_\varepsilon^{-1}B_0V_0B_0^\top\Sigma_\varepsilon^{-1}$. Up to an additive term that does not depend on y,

$$\log p_\omega(y \mid x) = -\frac{1+\omega}{2}(y-B_0x)^\top\Sigma_\varepsilon^{-1}(y-B_0x)+\frac{\omega}{2}y^\top\Sigma_Y^{-1}y = -\frac{1}{2}y^\top J_\omega y+(1+\omega)y^\top\Sigma_\varepsilon^{-1}B_0x,$$

where

$$J_\omega = (1+\omega)\Sigma_\varepsilon^{-1} - \omega\Sigma_Y^{-1} = \Sigma_\varepsilon^{-1} + \omega\Sigma_\varepsilon^{-1}B_0V_0B_0^\top\Sigma_\varepsilon^{-1}.$$

Because $\omega \geq 0$ and $V_0 \succ 0$, $J_\omega \succeq \Sigma_\varepsilon^{-1} \succ 0$, and hence $p_\omega(\cdot \mid x)$ is the Gaussian density with precision matrix J_ω and mean $(1+\omega)J_\omega^{-1}\Sigma_\varepsilon^{-1}B_0x$. Using $B_0^\top\Sigma_\varepsilon^{-1}B_0 = G_0$ and $V_0G_0 = H_0$, we obtain

$$J_\omega B_0 = \Sigma_\varepsilon^{-1}B_0 + \omega\Sigma_\varepsilon^{-1}B_0V_0G_0 = \Sigma_\varepsilon^{-1}B_0(I_p + \omega H_0),$$

which implies $J_\omega^{-1}\Sigma_\varepsilon^{-1}B_0 = B_0(I_p + \omega H_0)^{-1}$ and the first identity in (23). For the second identity,

$$\begin{aligned} B_\omega - B_0 &= B_0(I_p + \omega H_0)^{-1}\{(1+\omega)I_p - (I_p + \omega H_0)\} \\ &= \omega B_0(I_p + \omega H_0)^{-1}(I_p - H_0) = \omega B_0(I_p + \omega H_0)^{-1}V_0\Psi^{-1}. \end{aligned}$$

If $\omega > 0$, the matrix $(I_p + \omega H_0)^{-1}V_0\Psi^{-1}$ is invertible, and hence $B_\omega = B_0$ if and only if $B_0 = 0$. This completes the proof.

H Interpretation of the proof components

The proof of Theorem 1 separates population bias from finite-query estimation. Proposition 8 is the only step in which the true conditional mean μ_0 is compared with the teacher mean μ_θ. The theorem uses the exact projected bias d_θ and, when desired and when $\lambda_{\min}(\Gamma_X) \geq \kappa > 0$, upper-bounds it by $\kappa^{-1/2}\delta_\theta$ through Proposition 8. Lemma 1 is purely algebraic and relates errors in the empirical normal equations to errors in the coefficient maps through Γ_M^{-1}. Lemma 2 then supplies the two stochastic rates, namely the $M^{-1/2}$ query-design term and the $(MK)^{-1/2}$ teacher Monte Carlo term. For a fixed teacher, these bounds quantify error from the finite query design and teacher sampling. The projected teacher bias remains a separate term.

I Additional implementation details

Spatial representation. The coefficient maps can be represented voxelwise with smoothing penalties, with splines or wavelets, or with coordinate networks. Coordinate networks are well suited to high-resolution images and to applications that require evaluating the maps at arbitrary spatial locations. Sparse or total-variation penalties can be added when effect maps are expected to occupy localized regions.

Teacher mean estimation. For diffusion teachers, the conditional mean can be estimated by repeated sampling. A deterministic sampler used consistently across query covariates, or a denoised low-noise prediction, defines a separate access query unless the noised-query distribution and averaging scheme identify $\mu_\theta(x)$. Repeated sampling is slower but separates teacher uncertainty from student estimation.

Choice of design distribution. The design distribution should be reported with the maps. If ν is the empirical distribution, the map describes the observed population. If ν is balanced, the map describes a target population chosen by design. If ν is localized around x_0, the output is a local effect map.

Choice of empirical weights and access modes. The weights in (16) define the empirical target being estimated and should be reported with the maps. If only samples from the teacher are available, mean distillation is the default access mode. If the teacher is differentiable with respect to x, effect distillation directly targets the average local teacher sensitivity in Theorem 3. If a conditional score is available, score distillation should be accompanied by sensitivity to the student covariance $C_{\phi,\tau}$ and by comparison with a mean-distilled map when conditional mean queries are feasible. Large discrepancies among these maps indicate that the chosen discrepancy or student covariance is changing the projected teacher summary.

Score-based GED on a conditional residual DDPM. Score access on a conditional residual DDPM with ϵ-prediction is implemented by `experiments/celeba_attribute_score_ged.py`, which loads the same trained CelebA conditional residual DDPM checkpoint as the mean-distillation analysis and reads off a denoised query from the teacher score through Tweedie's formula at a single noise level τ. Let $\bar{\alpha}_\tau$ denote the cumulative signal coefficient of the DDPM forward noising process at level τ, let $\widehat{b}_\theta(x_m)$ denote the fitted condition-specific base mean, and let $\widehat{s}_{\mathrm{res}}$ denote the residual scale saved with the teacher checkpoint. Concretely, at each query covariate x_m the script draws a noised residual $\tilde{Y}_\tau = \sqrt{\bar{\alpha}_\tau}\,\bar{y} + \sqrt{1-\bar{\alpha}_\tau}\,\xi$ about the conditional residual base $\bar{y} = 0$ with $\xi \sim \mathcal{N}(0, I)$, evaluates the teacher noise prediction $\epsilon_\theta(\tilde{Y}_\tau, x_m, \tau)$, and defines the clipped residual estimate

$$\widehat{r}^{\mathrm{Tw}}_{\theta,\tau}(x_m) = \mathrm{clip}_{[-1,1]}\left\{\frac{\tilde{Y}_\tau - \sqrt{1-\bar{\alpha}_\tau}\,\epsilon_\theta(\tilde{Y}_\tau, x_m, \tau)}{\sqrt{\bar{\alpha}_\tau}}\right\},$$

where $\mathrm{clip}_{[-1,1]}$ clips each component to $[-1, 1]$. The image-scale score-access query used for OLS is $\widehat{Y}^{\mathrm{Tw}}_{\theta,\tau}(x_m) = \widehat{b}_\theta(x_m) + \widehat{s}_{\mathrm{res}}\widehat{r}^{\mathrm{Tw}}_{\theta,\tau}(x_m)$, clipped to the image range. It coincides with a teacher conditional mean query only when the noised input distribution and averaging scheme identify $\mu_\theta(x_m)$. A single OLS fit on $(x_m, \widehat{Y}^{\mathrm{Tw}}_{\theta,\tau}(x_m))$ produces the score-based GED estimate. This procedure uses one ϵ evaluation per query, not the 30-step DDIM sampler used by the mean-distillation analysis and produces a score-access denoised query before the linear projection step. The script writes `results/celeba_score_ged_summary.csv`, which provides the score-access row of Table 12: with 15 repetitions at $M = 768$ empirical-design covariates and $\tau = 20$, score-access GED attains SISE 0.009 ± 0.000, map correlation 0.996 ± 0.000, and signed region F1 0.936 ± 0.001, compared with mean-access GED at 0.090, 0.955, and 0.769. These values are closer to the teacher-pool OLS reference (0.005, 0.997, 0.932) than the mean-access values.

Coordinate-network, Integrated Gradients, and SmoothGrad baselines. The neural image-response regression baseline of Zhang et al. (2023) fits $\beta(s)$ as a coordinate network indexed by the spatial location. We implement this baseline as a four-layer 128-unit coordinate MLP that maps each pixel $s \in [-1, 1]^2$ to $(\beta_0(s), \beta_1(s), \beta_2(s))$ and is trained on the same $n = 120$ observed image-covariate pairs as the direct-OLS row. We also include two widely used saliency-based interpretation baselines: Integrated Gradients (Sundararajan et al., 2017), which integrates the teacher gradient $\partial\mu_\theta/\partial x$ along the path from a zero baseline to each observed covariate using a 20-step trapezoidal quadrature, and SmoothGrad (Smilkov et al., 2017), which averages the teacher gradient over 25 Gaussian-jittered inputs at $\sigma = 0.15$. For per-instance attribution maps $A_{ij}(s)$, we report $\widehat{\beta}^{\mathrm{sal}}_j(s) = n^{-1}\sum_{i=1}^n A_{ij}(s)$; this is intercept-only least squares across instances and does not regress attribution maps on the covariates. Reported in Table 9, the coordinate network attains SISE 0.009 ± 0.002, Integrated Gradients 0.007 ± 0.002, and SmoothGrad 0.008 ± 0.001. All three have lower SISE than direct OLS (0.012 ± 0.002), using spatial smoothness or teacher gradients. Their SISE values exceed those of mean-access GED (0.003 ± 0.001) and effect-access GED (0.002 ± 0.000) on the same teacher. These comparisons are consistent with design-distribution mismatch contributing to the remaining gap, while arguing against spatial parametrization or per-instance gradient aggregation as the sole explanation; they do not by themselves separate this contribution from the larger target-design query set and the teacher-access form used by GED.

Computational cost. Table 28 reports the wall-clock cost of each experimental block on the configurations released with this manuscript. The controlled, semi-real, FFHQ measured-attribute, and real-image scripts run on CPU within tens of minutes per repetition; the SDXL-Turbo and CelebA conditional-diffusion scripts use Apple MPS when available. The dominant cost in the diffusion

blocks is the per-query DDIM sampler, which scales linearly in the number of query covariates M and the number of denoising steps. In the SDXL-Turbo block, GED with $M = 512$ queries at 4 denoising steps requires roughly the same wall-clock cost as $4M = 2048$ teacher forward evaluations; in the CelebA block, GED with $M = 768$ queries at 30 DDIM steps corresponds to $\approx 2.3 \times 10^4$ forward evaluations per repetition. The score-based GED procedure described above replaces the 30-step sampler with a single ϵ evaluation per query, reducing diffusion-teacher querying cost by approximately the number of DDIM steps.

Table 28: Compute summary for the experimental blocks. Wall-clock times are approximate per-repetition costs on Apple Silicon CPU (or MPS where noted) with two PyTorch CPU threads.

Block	Hardware	Cost per repetition
Controlled synthetic	CPU	$<$ 1 minute
Semi-real digit teacher	CPU	$\approx$ 2–3 minutes
Learned synthetic teacher	CPU	$\approx$ 5 minutes (teacher fitting)
FFHQ measured attributes	CPU	$\approx$ 2 minutes
Real-image handwriting / face	CPU	$\approx$ 3–5 minutes
Latent-diffusion digit teacher	CPU	$\approx$ 10 minutes
CIFAR-10 pixel diffusion	CPU	$\approx$ 30 minutes
CelebA conditional diffusion	MPS or CPU	$\approx$ 30–60 minutes
Pretrained SDXL-Turbo/FFHQ-256	MPS preferred	$\approx$ 2–3 hours

Train/test split and the full-data reference. The full-data OLS reference used in the real-image blocks is itself a spatial regression on the same image pool and hence shares estimation noise with the direct-regression baseline. We report the sample-size sensitivity tables (Table 17 and Table 19), which fix the reference to the full-data fit and vary n for direct regression and augmentation. In these tables, the GED estimator is determined by the fixed learned teacher and the target-design queries, whereas the direct-regression estimator changes with the number of observed images. The narrowing difference in SISE between GED and direct regression is caused by the decreasing estimation error of direct regression as n increases, while GED error remains governed by the teacher-defined projection, the finite-query terms, and the teacher-bias term d_θ in (20). Splitting a real-image pool into a teacher-training subset and a held-out reference subset can separate teacher overfitting from target-design effects. The controlled synthetic and finite-query scaling tables avoid this dependence on a same-pool reference by using an analytic teacher with a known projection target, while the real-image sample-size sensitivity tables hold the full-data reference fixed and report the resulting dependence on n.

Sample size and statistical power. The SDXL-Turbo block uses $n = 12$ direct observed samples and $M = 512$ teacher queries, reflecting scientific imaging settings in which observed images are expensive to acquire while pretrained generators are available; the regime in which direct regression matches GED, that is, the regime in which direct OLS attains a SISE no larger than that of GED, is identified empirically by the morphology and face sample-size sensitivity tables. The five and eight repetitions used in the diffusion and ridge-teacher blocks were chosen because of the wall-clock cost of training a conditional teacher; the standard errors reported alongside the means in Tables 11–22 are smaller than the gaps between competing estimators, and the CIFAR-10 pixel-diffusion and latent-diffusion teacher blocks provide corroborating evidence on natural-image and digit-image domains at 6 and 8 repetitions, respectively.

Reproducibility and compute. The following Python commands run the supplied experiment and sensitivity implementations from the repository root:

```
python experiments/sim_controlled_ged.py
python experiments/real_digit_ged.py
python experiments/stable_diffusion_ffhq_ged.py --cached-only
python experiments/celeba_attribute_diffusion_ged.py
python experiments/celeba_attribute_score_ged.py --reps 15
python experiments/celeba_attribute_ged.py
python experiments/cifar_diffusion_ged.py
python experiments/diffusion_teacher_ged.py
```

```
python experiments/real_ffhq_attribute_ged.py
python experiments/real_morphology_ged.py
python experiments/real_face_ged.py
python experiments/rebuttal_targeted_checks.py --analysis highfreq
--reps 100 --tag production_highfreq_r100 --seed 20260724
python experiments/rebuttal_targeted_checks.py --analysis matched
--reps 100 --teacher-epochs 600 --reference-m 50000 --tag
production_matched_r100 --seed 20260724
python experiments/rebuttal_teacher_diagnostics.py --reps 100
--holdout-n 400 --query-m 400 --sigma 0.25 --tag
production_teacher_diagnostics_r100
python experiments/rebuttal_misalignment_check.py --reps 100 --tag
production_misalign_r100 --seed 20260801
python experiments/rebuttal_bias_localization_audit.py --reps 100
--holdout-n 400 --query-m 400 --teacher-k 8 --bias-rho 0.75 --seed
20260803 --tag production_bias_localization_r100
```

The supplementary material provides experiment programs, teacher checkpoints, cached queries, and stored results. Its README gives execution and verification commands. The scripts listed above generate synthetic data when needed, load the scikit-learn handwritten-digit, CelebA, FFHQ-256, CIFAR-10, and Olivetti face data for the image analyses, fit or query the conditional teachers used in the experiments, and write the CSV summaries, TeX table fragments, and figures included in the manuscript. The sample sizes, query sizes, teacher sample counts, image sizes, and repetition counts are stated in Appendix F. Neural teachers are fully connected networks trained by AdamW on squared image loss. The learned synthetic teacher has three 256-unit hidden layers, SiLU activations, learning rate 8×10^{-4}, weight decay 3×10^{-5}, batch size 128, and 600 epochs. The semi-real digit teacher has two 256-unit hidden layers, learning rate 10^{-3}, weight decay 5×10^{-5}, batch size 128, and 300 epochs; this block uses 800 teacher-training samples, $n = 120$ observed images, eight observed-covariate augmentation queries per observed covariate, $M = 800$ GED queries, and 8 repetitions. The SDXL-Turbo face experiment uses the pretrained model `stabilityai/sdxl-turbo` through the `diffusers` package and the public `bitmind/ffhq-256` dataset. It uses 100 base faces, generates 512×512 image-to-image outputs, downscales them to 64×64 RGB coefficient arrays, uses 4 denoising steps with strength 0.6 and guidance scale 0, fits a reference projection from 512 target-design queries, fits GED from 512 target-design queries per repetition, uses 12 direct observed teacher samples and 12 additional observed-covariate augmentation samples, and repeats the comparison 5 times. The scalar-to-prompt rule uses thresholds ± 0.33: $x_1 > 0.33$ uses "wearing large black sunglasses and smiling," $x_1 < -0.33$ uses "without sunglasses and with a neutral expression," and the middle bin uses "wearing clear eyeglasses and with a gentle smile." The corresponding x_2 phrases are "warm golden studio lighting and a bright white background," "cool blue studio lighting and a dark gray background," and "neutral studio lighting and a plain gray background." The CelebA semantic-attribute diffusion analysis uses the public `flwrlabs/celeba` dataset with configuration `img_align+identity+attr`, the official attributes Smiling, Eyeglasses, Male, and Blond Hair, 2000 aligned images downsampled to 32×32 RGB coefficient arrays, 80 observed images for direct OLS, a teacher-training pool of 1800 images, 768 GED query covariates per repetition, 5 mean-access repetitions, 15 score-access repetitions, and a conditional residual DDPM teacher trained for 1000 gradient steps with 200 diffusion training timesteps and 30 DDIM sampling steps. The FFHQ measured-attribute analysis uses 100 FFHQ-256 images, four image-measured covariates, 32×32 RGB coefficient arrays, 25 observed images for direct OLS, a teacher-training pool of 80 images, 1200 GED query covariates per repetition, and 8 repetitions. Its conditional mean teacher is a degree-two polynomial ridge regression with penalty parameter 5. The four FFHQ covariates are standardized versions of luminance-weighted horizontal and vertical centers, mean luminance, and RGB warmth, with luminance computed as $0.299R + 0.587G + 0.114B$ and RGB warmth computed as the image average of $R - B$. The handwritten morphology covariates are standardized versions of the intensity-weighted horizontal center and mean intensity, and the face morphology covariates are standardized versions of intensity-weighted horizontal and vertical centers and mean intensity. The handwritten morphology and face teachers have two hidden layers of widths 128 and 192, respectively, learning rate 10^{-3}, weight decay 10^{-4}, batch size 64, and 600 and 450 epochs, respectively. The handwritten morphology block uses 145 teacher-training images, observed sample sizes $n \in \{30, 60, 90, 120\}$, eight augmentation queries per observed covariate, $M = 800$ GED queries, and 8 repetitions. The face morphology block uses 320 teacher-training images, observed sample sizes

$n \in \{40, 80, 160, 240\}$, eight augmentation queries per observed covariate, $M = 1200$ GED queries, and 8 repetitions. The scripts fix base seeds 20260430, 20260501, 20260901, 20261021, 20261020, 20260801, 20260715, 20261015, 20260601, and 20260630 across the empirical blocks and use consecutive seed offsets across repetitions. The non-SDXL scripts require only CPU computation; the SDXL-Turbo script uses Apple MPS when available and otherwise uses CPU execution, with at most two PyTorch CPU threads per script.

J Ethics, assets, and broader impact

GED produces spatial coefficient maps that summarize a conditional image generator by linear projection of its conditional distribution onto image-valued covariate effects. This can support scientific analysis when the teacher, target design distribution, and calibration diagnostics are reported. The main risk is overinterpretation: distilled maps can contain teacher bias, acquisition artifacts, demographic imbalance, or spurious correlations, and they should not be interpreted as causal effects without a valid causal design. The experiments in this manuscript use synthetic images, CelebA aligned face images with official semantic attributes (Liu et al., 2015), FFHQ-256 face images (Karras et al., 2019), CIFAR-10 images (Krizhevsky, 2009), scikit-learn handwritten digits (Pedregosa et al., 2011), Olivetti face images (Samaria and Harter, 1994; Pedregosa et al., 2011), and the pretrained SDXL-Turbo model (Podell et al., 2023; Sauer et al., 2023); no proprietary image data or newly collected human-subject data are used. The CIFAR-10 project page provides the image archive and asks users to cite the technical report, but does not state a separate dataset license on that page (`https://www.cs.toronto.edu/~kriz/cifar.html`). The scikit-learn package is distributed under the BSD 3-Clause license. The handwritten digits are the scikit-learn copy of the University of California Irvine Optical Recognition test set, which the UCI Machine Learning Repository lists under the Creative Commons Attribution 4.0 International license (`https://archive.ics.uci.edu/dataset/80/optical+recognition+of+handwritten+digits`). The conditional-generation experiments use the open-source `diffusers` library (von Platen et al., 2022), distributed under the Apache 2.0 license (`https://github.com/huggingface/diffusers`). The FFHQ-256 images are loaded from the public Hugging Face dataset `bitmind/ffhq-256`, a 256-pixel version of Flickr-Faces-HQ (`https://huggingface.co/datasets/bitmind/ffhq-256`). The `bitmind/ffhq-256` dataset card does not list a separate dataset license; we rely on the upstream FFHQ terms described next. The NVIDIA FFHQ repository states that the individual images were collected from Flickr under Creative Commons BY 2.0, Creative Commons BY-NC 2.0, Public Domain Mark 1.0, Public Domain CC0 1.0, or U.S. Government Works licenses, while the dataset metadata, download script, and documentation are under the Creative Commons BY-NC-SA 4.0 license (`https://github.com/NVlabs/ffhq-dataset`). The CelebA semantic-attribute analysis uses the public Hugging Face dataset `flwrlabs/celeba` with configuration `img_align+identity+attr`; its dataset card lists the `celeba-dataset-release-agreement` license and provides the official binary attributes used in the experiment (`https://huggingface.co/datasets/flwrlabs/celeba`). The pretrained SDXL-Turbo weights are downloaded from the Hugging Face model id `stabilityai/sdxl-turbo`, whose model card lists the `sai-nc-community` license and directs commercial users to Stability AI licensing (`https://huggingface.co/stabilityai/sdxl-turbo`). Face-image datasets and image generators may contain demographic imbalance, privacy-sensitive face content, and artifacts from the collection source, even when they are used only as public research assets. The Olivetti faces are credited to AT&T Laboratories Cambridge as requested in the data loader documentation; the scikit-learn documentation states this credit term and does not list a separate dataset license.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the proposed teacher-to-map functional, the mean/effect/score distillation objectives, the finite-query theory, and the empirical scope. The stated limitations are also included in Appendix E.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Appendix E discusses noncausal interpretation, teacher bias, global averaging of nonlinear effects, score calibration, registration uncertainty, and the resolution–query trade-off.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: The assumptions are stated before the finite-query theorem and within each population result; complete proofs of all propositions, theorems, corollaries, and auxiliary lemmas are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: The experimental designs, query sizes, repetitions, metrics, teacher-training settings, pretrained SDXL-Turbo/FFHQ-256 settings, reference targets, and additional sensitivity studies are stated in Section F; Appendix I lists the reproduction commands and output artifacts.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The supplementary material provides experiment scripts, teacher checkpoints, cached queries, replication-level results, and numerical validators. The README and Appendix I give commands for running the programs and checking the reported results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The experiment section gives the sample sizes, query sizes, teacher sample counts, repetition counts, held-out teacher errors, SDXL-Turbo reference residual scale, image sizes, and registration perturbations; Appendix I states the neural teacher architectures, query settings, optimizer settings, training lengths, seeds, and compute details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

Justification: Tables report averages over independent repetitions with standard errors after the plus-minus sign; the number of repetitions is stated in the table captions and experiment description. In addition, Table 1 reports a one-sided paired Student t-test against the strongest non-GED baseline in every block, with $p < 0.05$ in all eleven blocks and $p < 10^{-3}$ in five of eleven; the test assumes paired repetitions sharing seed structure, and the standard error is the sample standard deviation divided by the square root of the number of repetitions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Table 28 in Appendix I reports per-block hardware (Apple Silicon CPU or MPS) and approximate per-repetition wall-clock cost for every experimental block; Appendix I also limits PyTorch to two CPU threads per script and states the seeds, sample sizes, query sizes, teacher-sample counts, and repetition counts needed to reproduce the runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The experiments use synthetic data, FFHQ-256, public handwritten digits, Olivetti face images, and a pretrained SDXL-Turbo teacher, and the paper discusses teacher bias, calibration, uncertainty, noncausal interpretation, and the public benchmark status of the image assets.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: Appendix J discusses scientific uses of interpretable maps and the risks caused by teacher bias, spurious correlations, and overinterpretation of noncausal maps.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The released teachers are small experimental models, including a 32-by-32 CelebA checkpoint. The release is intended for coefficient-map experiments on public benchmarks. Raw image datasets and the SDXL-Turbo pretrained weights are obtained from their original sources; their terms and the scope of the experiments are described in Appendix J.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: The paper cites FFHQ, SDXL-Turbo, scikit-learn, and the Olivetti face reference for the public image and model assets, and Appendix J states the available license or terms information for these assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The new assets are reproducibility scripts, experimental teacher checkpoints, and generated result files, which are documented by Appendix I and the supplementary README. The experiment and validation scripts are released under the BSD 3-Clause License included in the archive.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

Justification: The paper does not involve crowdsourcing or newly collected human-subject data.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: The paper uses synthetic images, FFHQ-256, public handwritten digits, and Olivetti face images, with no new human-subject data collection.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A].

Justification: The core method does not use LLMs as a modeling, training, inference, or evaluation component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.