REFIT OR EXPAND? FINITE-SAMPLE CERTIFICATION OF INTERACTION SURROGATES UNDER COMPOUND SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Assessing multistage inference under compound distribution shifts requires repeated evaluations across parameter settings because interacting parameters and noncommuting operators can produce non-additive performance changes. Interaction surrogates approximate these joint effects using low-order marginal moments. However, when surrogate error exceeds a tolerance, empirical point estimates cannot establish whether the discrepancy can be resolved by coefficient refitting or requires additional features. We introduce a finite-sample certification framework based on empirical Bernstein confidence intervals. We solve two linear programs: a robust refit program provides an upper bound on the error achievable in the surrogate class, while a dual program provides a lower bound for every class member through the design matrix's left nullspace. These bounds can certify parameter misfit or feature inadequacy and identify minimal adequate subsets among candidate terms. For bilinear surrogates, the same guarantees hold uniformly over continuous product mixtures of evaluated marginals. Vision and language experiments show when interactions can be omitted, when refitting suffices, and when feature expansion is required. For ResNet-50 on ImageNetV2 under resampling/JPEG shifts, the one-term class has certified error at least 0.185 pp while a two-term surrogate has certified error at most 0.135 pp. For blur/JPEG, order reversal changes error by over 6 pp. The interaction surrogates correct all three order selection errors made by the additive approximation.

1 INTRODUCTION

When a compact surrogate fails under compound distribution shifts, can its coefficients be refitted to meet tolerance, or are additional features required? When multiple operators are composed, jointly varying their parameter distributions can produce changes beyond their separate marginal effects. To reduce repeated evaluations across joint settings, compact interaction surrogates approximate the interaction response using low-order marginal moments.

A surrogate error above tolerance can reflect *parameter misfit* if another coefficient choice within the same class meets tolerance, or *feature inadequacy* if no class member does. Validation residuals cannot resolve this distinction because they mix surrogate error with sampling variability and assess only the fitted surrogate. By contrast, feature inadequacy is a class-wide conclusion that can be established only by ruling out every coefficient choice in the representation.

We construct simultaneous empirical Bernstein confidence intervals (Maurer & Pontil, 2009) for the interaction responses over a target family. A fitted surrogate certificate applies to one coefficient choice. A robust refit linear program (LP) determines whether the class contains an adequate coefficient choice. A dual LP uses vectors in the left nullspace to obtain an error lower bound that applies to every class member. Together, these certificates separate parameter misfit from feature inadequacy.

We evaluate the method on language model decoding and vision preprocessing. In language decoding with fixed operator order, the zero surrogate meets the tolerance across shifts in temperature and nucleus (top-p) distributions. In vision, refitting resolves some failures, whereas class-wide lower bounds establish when additional features are required. In particular, the ImageNet-R evaluation

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

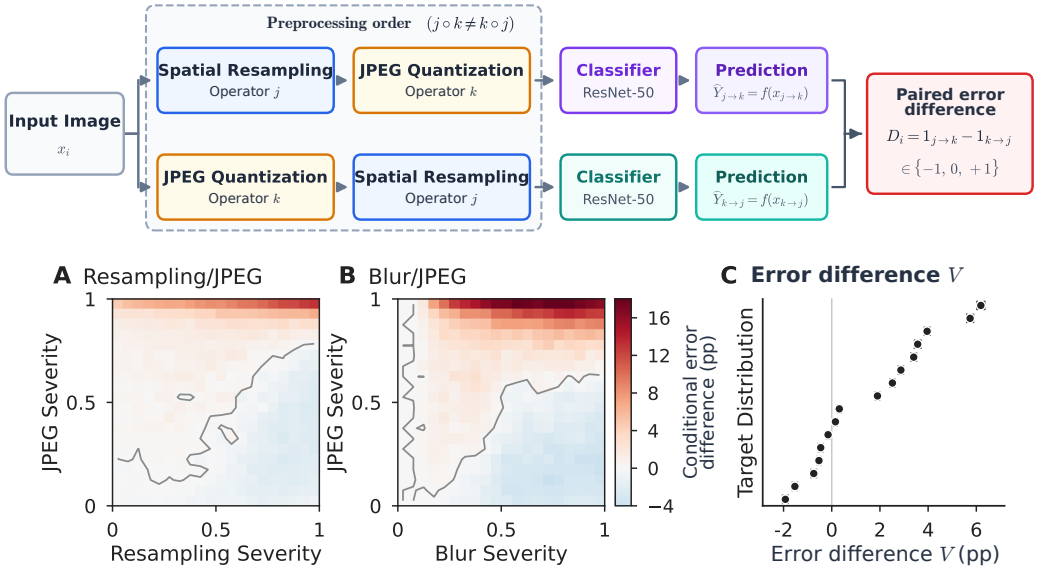


Figure 1: **Why interaction surrogates require certification.** Operator order can alter performance under compound shifts. The top panel presents an example composition of noncommuting image transformations ($A \circ B \neq B \circ A$). (A-B) Binned conditional error difference between the two operator orders for ResNet-50 on ImageNetV2. Solid zero contours indicate equal error. (C) Execution-order error difference $V(P)$ under compound target distributions.

separates failure of the transferred coefficients from class inadequacy. Refitting the same one-term class reduces the certified error interval from $[0.450, 0.496]$ pp for transferred coefficients to $[0.078, 0.125]$ pp. The certificates also provide guarantees for operator order selection, for which the ordering margin determines the required surrogate accuracy. On ImageNetV2, reversing blur and JPEG changes classification error by more than 6 pp under compound shifts, and the certified interaction surrogate corrects all three order selection errors made by the additive baseline (Figures 1 and 5).

Contributions. We introduce a finite-sample framework that certifies when coefficient refitting suffices and when interaction surrogates require feature expansion under compound shifts (Section 3, Algorithm 1). We develop certificates that identify required features and minimal adequate subsets within a candidate set (Section 3.2). For bilinear surrogates, we extend error upper bounds uniformly to continuous product mixtures of evaluated marginals without additional model evaluations (Section 5.1). We establish guarantees for operator order selection and demonstrate the framework across vision and language pipelines, including correction of ordering errors made by additive baselines (Section 4,6).

2 Interaction Structure and Distributional Features

2.1 Functional ANOVA and Interaction Structure

Let $A_j, A_k \in [0, 1]$ denote two normalized operator parameters. The reference distribution is $P_0 = P_{0,j} \otimes P_{0,k}$, with $P_{0,\ell} = \text{Uniform}[0, 1]$ for $\ell \in \{j, k\}$. For each of N fixed evaluation instances, let $D_i(A_j, A_k, \xi)$ be the bounded response, where ξ represents the random draws used during evaluation. In vision, A_j and A_k control two image transformation severities. With label y_i and prediction $Y_{i,j\to k}$ under operator order $j \to k$, the response is

$$D_i(A_j, A_k, \xi) := \mathbf{1}\{Y_{i,j\to k} \neq y_i\} - \mathbf{1}\{Y_{i,k\to j} \neq y_i\} \in \{-1, 0, 1\}. \tag{1}$$

Both orders use the same crop and random draws. In the language evaluation, A_j controls temperature and A_k controls nucleus (top-p) truncation. Their order is fixed, and $D_i(A_j, A_k, \xi) := \mathbf{1}\{Y_i(A_j, A_k, \xi) \neq y_i\} \in \{0, 1\}$ is the classification error indicator (Appendix D).

Define $f(a,b) \coloneqq N^{-1}\sum_i \mathbb{E}_\xi[D_i(a,b,\xi)]$ and $\mathbb{E}_0[\cdot] \coloneqq \mathbb{E}_{P_0}[\cdot]$. Functional ANOVA (Hooker, 2007, Lengerich et al., 2020) decomposes f as $f(a,b) = M_0 + f_j(a) + f_k(b) + g(a,b)$, where $M_0 = \mathbb{E}_0[f]$, $\mathbb{E}_{P_{0,j}}[f_j] = \mathbb{E}_{P_{0,k}}[f_k] = 0$, and $g \in H_{jk} \coloneqq \{h \in L^2(P_0) : \mathbb{E}_0[h \mid A_j] = \mathbb{E}_0[h \mid A_k] = 0\}$.

For a product target $P = P_j \otimes P_k$, define $V(P) \coloneqq \mathbb{E}_P[f(A_j, A_k)]$, $V_{\mathrm{add}}(P) \coloneqq M_0 + \mathbb{E}_{P_j}[f_j] + \mathbb{E}_{P_k}[f_k]$, and $F(P) \coloneqq \mathbb{E}_P[g]$. The interaction response $F(P)$ is the contribution beyond the two marginal effects, so $V(P) = V_{\mathrm{add}}(P) + F(P)$. The equivalent marginal-reference expression for V_{add} appears in Appendix G.2.

Let $w_\ell \coloneqq dP_\ell/dP_{0,\ell} \in L^2(P_{0,\ell})$. The conditional mean-zero property of g gives $F(P) = \langle g, h_P\rangle_{L^2(P_0)} = \mathbb{E}_0[f h_P]$, where $h_P(a,b) \coloneqq (w_j(a)-1)(w_k(b)-1) \in H_{jk}$. Hence, $F(P) = 0$ whenever either target marginal equals its reference distribution.

2.2 Distributional features and surrogate classes

Under the uniform reference distribution, tensor products $\{q_u \otimes q_v : u, v \geq 1\}$ of shifted Legendre polynomials form an orthonormal basis of H_{jk} (Sobol', 1993, Xiu & Karniadakis, 2002). We use

$$q_1(a) = \sqrt{12}\left(a - \frac{1}{2}\right), \quad q_2(a) = \sqrt{5}(6a^2 - 6a + 1),$$
$$\lambda_{\ell u} \coloneqq \mathbb{E}_{P_\ell}[q_u], \quad m_\ell \coloneqq \lambda_{\ell 1}, \quad s_\ell \coloneqq \lambda_{\ell 2}, \quad x_{uv}(P) \coloneqq \lambda_{ju}\lambda_{kv}. \tag{2}$$

The $\lambda_{\ell u}$ are the shifted-Legendre coefficients of $w_\ell - 1$, so h_P has tensor coefficients $x_{uv}(P)$. With $\beta_{uv} \coloneqq \langle g, q_u \otimes q_v\rangle_{L^2(P_0)} = \mathbb{E}_0[f(A_j, A_k) q_u(A_j) q_v(A_k)]$, orthogonality gives

$$F(P) = \sum_{u,v \geq 1} \beta_{uv} x_{uv}(P).$$

For marginal degrees one and two, the tensor features are $x_{11} = m_j m_k$, $x_{12} = m_j s_k$, $x_{21} = s_j m_k$, and $x_{22} = s_j s_k$.

For a finite index set $S \subset \{1, 2, \ldots\}^2$, abbreviate $(1,1), (1,2), (2,1), (2,2)$ as $11, 12, 21, 22$. Define $\mathcal{G}_S \coloneqq \{P \mapsto \sum_{r \in S} c_r x_r(P) : c \in \mathbb{R}^{|S|}\}$. The basis coefficients β_{uv} are determined by g and P_0. Refitting optimizes coefficients within $\mathcal{G}_S$ (Definition 1). Appendix E.1 gives the corresponding geometric decomposition into omitted basis terms and coefficient error.

We consider the one-term class $\mathcal{G}_{\mathrm{sc}} = \{c\, m_j m_k : c \in \mathbb{R}\}$ with $S = \{11\}$. The exchange-symmetric class $\mathcal{G}_{\mathrm{sym}}$ contains surrogates whose predictions are invariant to exchanging the two marginal distributions between the two operators (Appendix E). The two-term surrogate is $C_2(P) = \widehat{\beta}_{11} m_j m_k + \widehat{\beta}_{12} m_j s_k$ with $S = \{11, 12\}$, where $\widehat{\beta}_{11}$ and $\widehat{\beta}_{12}$ are estimated from fitting data. For a finite target set $\mathcal{T} = \{P_t\}_{t=1}^T$, the feature evaluations yield the design matrix $X \in \mathbb{R}^{T \times |S|}$, with $X_{t,r} = x_r(P_t)$ for $r \in S$. Section 3 analyzes the class $\mathcal{G}_X = \{Xb : b \in \mathbb{R}^{|S|}\}$.

3 Primal-dual certification of surrogate adequacy and feature inadequacy

3.1 Certification from simultaneous response bounds

For $\mathcal{T} = \{P_t\}_{t=1}^T$, write $\mathbf{F} = (F(P_t))_{t=1}^T$ and, for any surrogate C, $\mathbf{C} = (C(P_t))_{t=1}^T$. We identify surrogate classes with their evaluation vectors on $\mathcal{T}$ when computing class distances. Let $\mathcal{B} = \prod_{t=1}^T [L_t, U_t]$ denote a simultaneous $(1-\alpha)$ confidence box constructed from the finite-sample bounds in Section 3.3.

Lemma 1 (Surrogate error bounds)**.** *On the event $\mathbf{F} \in \mathcal{B}$, the bounds hold simultaneously for every $\mathbf{C} \in \mathbb{R}^T$,*

$$\mathcal{L}_C \coloneqq \max_t \max\{0, L_t - C_t, C_t - U_t\} \leq E_C \leq \max_t \max\{|L_t - C_t|, |U_t - C_t|\} \coloneqq \mathcal{U}_C, \tag{3}$$

where $E_C \coloneqq \|\mathbf{F} - \mathbf{C}\|_\infty$.

Definition 1 (Surrogate adequacy, parameter misfit, and feature inadequacy). *Let $\mathcal{G} \subseteq \mathbb{R}^T$ be a nonempty closed surrogate class, let $\mathbf{C} \in \mathcal{G}$, and define $d_{\mathcal{G}}(\mathbf{F}) \coloneqq \inf_{\mathbf{y}\in\mathcal{G}} \|\mathbf{F} - \mathbf{y}\|_\infty$. At tolerance ϵ, C is **adequate** if $\|\mathbf{F} - \mathbf{C}\|_\infty \le \epsilon$. The surrogate C has **parameter misfit** if $\|\mathbf{F} - \mathbf{C}\|_\infty > \epsilon$ and $d_{\mathcal{G}}(\mathbf{F}) \le \epsilon$. The class $\mathcal{G}$ has **feature inadequacy** if $d_{\mathcal{G}}(\mathbf{F}) > \epsilon$.*

By Lemma 1, $\mathcal{U}_C \le \epsilon$ certifies adequacy. The condition $\mathcal{L}_C > \epsilon$ certifies that C exceeds tolerance. In particular, adequacy of $C = 0$ certifies that the interaction term may be omitted at tolerance ϵ.

Summary-based approximation bounds. For a summary mapping $\phi\colon \mathcal{T} \to \mathsf{Z}$, let $\mathcal{G}_\phi$ be the class of all surrogates that assign the same prediction to targets with the same summary. The approximation error of this class satisfies $d_{\mathcal{G}_\phi}(\mathbf{F}) = \frac{1}{2}\max_{z\in\phi(\mathcal{T})} \operatorname{diam}\{F_t : \phi(P_t) = z\}$ (Appendix E.3, Proposition E.2). On the event $\mathbf{F} \in \mathcal{B}$, the bounds certify feature inadequacy of $\mathcal{G}_\phi$ if two targets t_1, t_2 with $\phi(P_{t_1}) = \phi(P_{t_2})$ satisfy $\max\{L_{t_1} - U_{t_2}, L_{t_2} - U_{t_1}\} > 2\epsilon$.

3.2 Robust refitting and dual class lower bounds

For the linear class $\mathcal{G}_X = \{Xb : b \in \mathbb{R}^d\}$, where $d \coloneqq |S|$, any $\nu \in \operatorname{null}(X^\top)$ with $\|\nu\|_1 \le 1$ is orthogonal to every class prediction and satisfies, for every b,

$$|\nu^\top \mathbf{F}| = |\nu^\top(\mathbf{F} - Xb)| \le \|\nu\|_1 \|\mathbf{F} - Xb\|_\infty \le \|\mathbf{F} - Xb\|_\infty$$

by Hölder's inequality. Taking the infimum over b gives $|\nu^\top \mathbf{F}| \le d_{\mathcal{G}_X}(\mathbf{F})$.

Proposition 1 (Class lower bound and robust refit). *Let $\mathbf{L} \coloneqq (L_t)_{t=1}^T$ and $\mathbf{U} \coloneqq (U_t)_{t=1}^T$, and define the box center $\mathbf{z} \coloneqq \frac{1}{2}(\mathbf{L} + \mathbf{U})$ and radius $\mathbf{r} \coloneqq \frac{1}{2}(\mathbf{U} - \mathbf{L})$. Strong linear programming duality gives the class lower bound*

$$\begin{aligned} \underline{d}_X^{\mathrm{box}} &\coloneqq \inf_{\mathbf{y}\in\mathcal{B}} \inf_{b\in\mathbb{R}^d} \|\mathbf{y} - Xb\|_\infty \\ &= \min_{b\in\mathbb{R}^d,\, e\ge 0} \left\{ e : \mathbf{L} - e\mathbf{1} \le Xb \le \mathbf{U} + e\mathbf{1} \right\} \qquad (4) \\ &= \max_{\nu\in\operatorname{null}(X^\top),\, \|\nu\|_1\le 1} \left\{ \nu^\top \mathbf{z} - \sum_{t=1}^T |\nu_t| r_t \right\}. \qquad (5) \end{aligned}$$

The robust refit value is

$$u_* = \overline{e}_X(\mathcal{B}) \coloneqq \inf_{b\in\mathbb{R}^d} \sup_{\mathbf{y}\in\mathcal{B}} \|\mathbf{y} - Xb\|_\infty = \min_{b\in\mathbb{R}^d,\, u\ge 0} \left\{ u : Xb - \mathbf{L} \le u\mathbf{1}, \quad \mathbf{U} - Xb \le u\mathbf{1} \right\}. \qquad (6)$$

On $\mathbf{F} \in \mathcal{B}$, $u_* \le \epsilon$ certifies an adequate robust refit, and the class approximation error satisfies

$$\underline{d}_X^{\mathrm{box}} \le d_{\mathcal{G}_X}(\mathbf{F}) \le \overline{e}_X(\mathcal{B}).$$

The class lower bound minimizes over compatible responses and coefficients, whereas robust refitting selects one coefficient vector against every response in $\mathcal{B}$. The gap between these bounds is at most $2\|\mathbf{r}\|_\infty$ (Appendix E.2). Alternatively, a Chebyshev refit minimizes $\|\widehat{\mathbf{F}}_{\mathrm{fit}} - Xb\|_\infty$ on the fitting response estimates. The resulting surrogates are evaluated using Eq. (3). Appendix E.4 gives explicit dual vectors.

Combining class lower bounds. To exploit cross-coordinate dependence, we construct confidence intervals for prespecified dual scores. Let $\mathcal{V}_X \subseteq \{\nu \in \operatorname{null}(X^\top) : \|\nu\|_1 \le 1\}$ be a finite set fixed before observing the evaluation sample, possibly using independent fitting data. With probability at least $1 - \alpha$, all response intervals and dual score intervals cover their respective quantities simultaneously. In particular, $F_t \in [L_t, U_t]$ for every t and $\nu^\top \mathbf{F} \in [L_\nu, U_\nu]$ for every $\nu \in \mathcal{V}_X$. The combined confidence set is

$$\mathcal{K} = \mathcal{B} \cap \bigcap_{\nu\in\mathcal{V}_X} \{\mathbf{y} : L_\nu \le \nu^\top \mathbf{y} \le U_\nu\}.$$

The combined class lower bound satisfies

$$\underline{d}_X \coloneqq \max\left\{ \underline{d}_X^{\mathrm{box}}, \max_{\nu\in\mathcal{V}_X} \max\{0, L_\nu, -U_\nu\} \right\} \le \inf_{\mathbf{y}\in\mathcal{K}} \inf_b \|\mathbf{y} - Xb\|_\infty, \qquad (7)$$

where the second maximum is taken as 0 when $\mathcal{V}_X = \emptyset$.

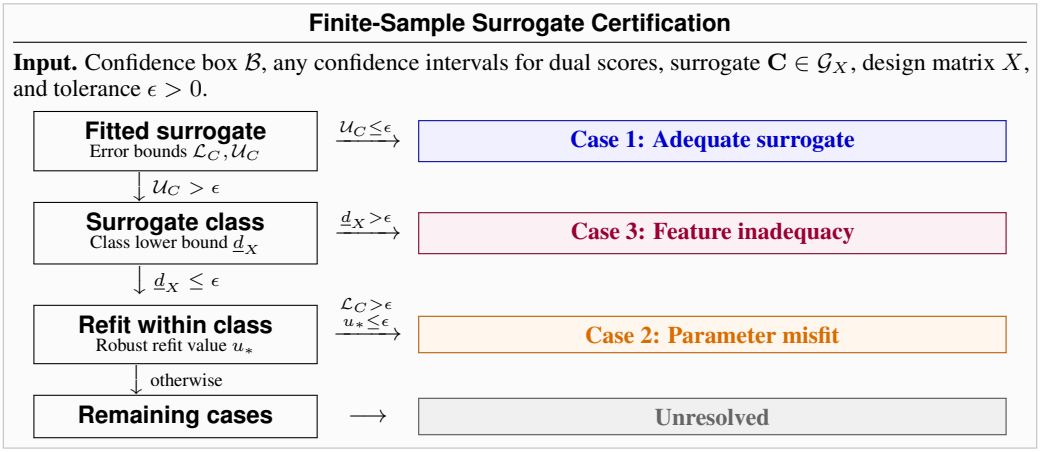


Figure 2: **Finite-sample certification procedure.** Algorithm 1 computes the bounds and applies the decision rule.

Certifying a minimal feature subset. For a finite candidate index set $\mathcal{D}$, let $\underline{d}_{-r}$ be a certified lower bound on $d_{\mathcal{G}_{\mathcal{D}\setminus\{r\}}}(\mathbf{F})$. On the joint coverage event, if a surrogate $C_S \in \mathcal{G}_S$, with $\varnothing \neq S \subseteq \mathcal{D}$, satisfies

$$\mathcal{U}_{C_S} \le \epsilon < \min_{r \in S} \underline{d}_{-r},$$

then every adequate subset of $\mathcal{D}$ must contain S. Because C_S is adequate, S is the unique minimal adequate subset.

3.3 Finite-sample construction of simultaneous response bounds

Proposition 2 (Finite-sample simultaneous empirical Bernstein coverage)**.** *Fix M evaluation coordinates and sample counts $n_m \ge 2$ before evaluation. For each $m \in \{1, \dots, M\}$, let $S_{u,m}$, $u = 1, \dots, n_m$, be independent over u, not necessarily identically distributed, with $|S_{u,m}| \le B_m$. Scores may be dependent across coordinates. Define $\mu_m = n_m^{-1} \sum_u \mathbb{E}[S_{u,m}]$, $\bar{S}_m = n_m^{-1} \sum_u S_{u,m}$, and $\hat{\sigma}_m^2 = (n_m - 1)^{-1} \sum_u (S_{u,m} - \bar{S}_m)^2$. With probability at least $1 - \alpha$, $|\bar{S}_m - \mu_m| \le r_m^{\mathrm{stat}}$ simultaneously for all m, where*

$$r_m^{\mathrm{stat}} = \sqrt{\frac{2\hat{\sigma}_m^2 \log(4M/\alpha)}{n_m}} + \frac{14 B_m \log(4M/\alpha)}{3(n_m - 1)}. \tag{8}$$

The M coordinates include the interaction responses and dual scores used for certification, auxiliary responses, and coordinates from settings evaluated jointly (Appendix C.2).

For vision, each independent unit u pairs an image with an evaluation replicate. Its score is $S_{u,t} = D_u h_t(A_{j,u}, A_{k,u})/\rho_u$, where $h_t := h_{P_t}$, $\rho_u := \rho(A_{j,u}, A_{k,u})$, and $\rho = dQ/dP_0$. Equal replicate counts for each image give $n_t^{-1} \sum_u \mathbb{E}[S_{u,t}] = F_t$ under this sampling design. The proposal satisfies $\rho \ge \frac{1}{4}$ (Appendix C). Language evaluations use independent question batches with the rescaling in Appendix D.

For each dual vector ν, the response coordinates with $\nu_t \neq 0$ use shared evaluation units. We form $S_{u,\nu} = \sum_t \nu_t S_{u,t}$ before computing its empirical variance. This preserves the cross-coordinate covariance induced by shared evaluations. The intervals combine the statistical and numerical error bounds (Appendix C.3).

4 Empirical results

Tolerance and reporting. We set $\epsilon = 0.15$ percentage points (pp) for vision and $\epsilon = 2.0$ pp for language, approximately three nominal standard errors of order differences and four of classification error, respectively. This choice gives a practical reference scale, while empirical Bernstein intervals

Table 1: **Certified outcomes at 0.15 pp for vision and 2.0 pp for language.** The Upper bound column reports $\epsilon^*(C)$. Case labels use class-wide lower bounds and refit bounds. The Terms column reports the fewest terms certified adequate. A dash denotes none.

Setting	Terms	Upper bound (pp)	Surrogate
Case 1: Adequate surrogate			
ResNet-50 / Tone/Noise	**0**	**0.121**	Zero
ViT-B/16 / Tone/Noise	**1**	**0.090**	$m_j m_k$
Qwen2.5-1.5B / Temp./Top-p	**0**	**1.02**	Zero
Case 2: Parameter misfit			
ConvNeXt-Tiny / Resamp./JPEG	**1**	**0.120**	Refit $m_j m_k$
Unresolved at 0.15 pp			
ViT-B/16 / Resamp./JPEG	–	0.161	$\{11, 12\}$
Case 3: Feature inadequacy			
ResNet-50 / Resamp./JPEG	**2**	**0.135**	$\{11, 12\}$
ResNet-50 / Blur/JPEG	**4**	**0.135**	$\{11, 12, 21, 22\}$
ResNet-20 / Resamp./JPEG	**2**	**0.134**	Refit $\{11, 12\}$
ViT-B/16 / Blur/JPEG	**3**	**0.148**	$\{11, 12, 21\}$

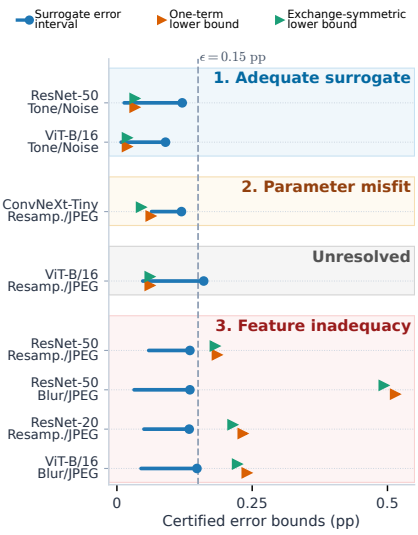


Figure 3: **Vision certification outcomes at 0.15 pp.** Bounds have simultaneous 95% coverage within each evaluation (Appendix C.2).

provide uncertainty bounds for the interaction estimators. For a surrogate C, let $\epsilon^*(C) := \mathcal{U}_C$ denote the smallest tolerance at which the bounds certify adequacy. Table 1 reports this value for each surrogate.

Experimental design. We evaluate ResNet-50, ViT-B/16, and ConvNeXt-Tiny on ImageNetV2 MatchedFrequency (Recht et al., 2019), and ResNet-20 on CIFAR-10. We evaluate Qwen2.5-1.5B-Instruct on a stratified MMLU sample (Hendrycks et al., 2021b). For vision, the primary targets are products of exponential tilt marginals relative to the uniform reference. Importance weighting reuses evaluation data across targets. Fitting and certification use independent data. Appendices C and D provide evaluation details.

4.1 Case 1: Adequate surrogate

For tone/noise shifts, the zero surrogate is adequate for ResNet-50 and the one-term surrogate for ViT-B/16. For MMLU temperature/top-p shifts (Holtzman et al., 2020), the zero surrogate is adequate at the language tolerance (Table 1, Appendix B.4).

4.2 Case 2: Refitting resolves parameter misfit

For ConvNeXt-Tiny under resampling/JPEG, the fitted one-term surrogate is certified to exceed the tolerance, while a Chebyshev refit in the same class has $\mathcal{U}_{Xb_{\mathrm{Ch}}} \leq 0.120$ pp on independent data. Since $u_* \leq \mathcal{U}_{Xb_{\mathrm{Ch}}}$, the class contains an adequate refit, establishing parameter misfit. The fitted coefficient estimates the $L^2(P_0)$ projection coefficient β_{11}, whereas the Chebyshev refit minimizes the maximum error over the fitting responses.

ImageNet-R evaluation. We transfer the one-term coefficient fitted on ImageNetV2 to ImageNet-R (Hendrycks et al., 2021a). The ResNet-50 model, resampling/JPEG operators, and target family remain unchanged. The transferred surrogate exceeds the tolerance, whereas a Chebyshev refit within the same one-term class meets it (Table 2, Appendix B.6). The one-term class is inadequate on ImageNetV2 but admits adequate refits on ImageNet-R. Because the target family is

Table 2: **One-term surrogate on ImageNet-R.** Surrogate error intervals with simultaneous 95% coverage.

Fit	Error interval (pp)
ImageNetV2 fit	$[0.450, 0.496]$
Chebyshev refit	$[0.078, 0.125]$

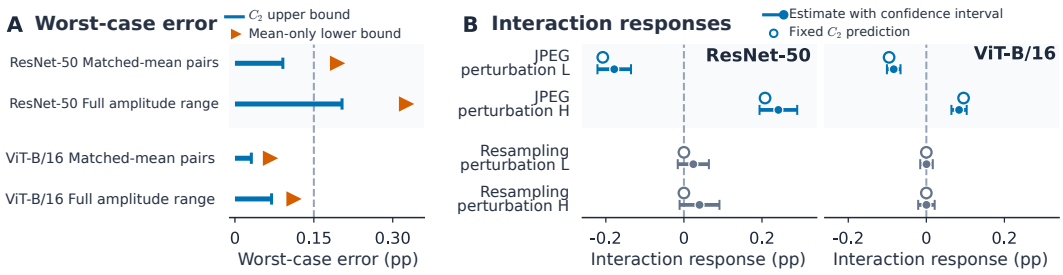


Figure 4: **Matched-mean perturbations on ImageNetV2.** (A) Error bounds for C_2 and class lower bounds for unrestricted mean-only surrogates on the matched-mean pairs and full amplitude range. (B) Interaction responses under second-moment perturbations at a fixed marginal mean. Intervals have simultaneous 95% coverage within each model's perturbation evaluation, with L and H denoting low and high severity of the fixed marginal.

common, both datasets have the same design matrix X. Transferring a coefficient vector evaluates one point Xb, whereas class adequacy depends on the distance from each dataset's response $\mathbf{F}$ to $\mathrm{col}(X)$.

4.3 Case 3: Feature inadequacy requires additional terms

For ResNet-50/ImageNetV2 under resampling/JPEG, the nested classes $\mathcal{G}_{\mathrm{sc}} \subset \mathcal{G}_{\mathrm{sym}}$ are both inadequate at 0.15 pp. Exchanging the marginals changes the response, and every exchange-symmetric surrogate gives exchanged targets the same prediction. The two-term surrogate has an upper bound of 0.135 pp, while the symmetric and one-term classes have lower bounds of 0.181 and 0.185 pp, respectively. Therefore, the bounds satisfy

$$\mathcal{U}_{C_2} < 0.15\,\mathrm{pp} < \min\{\underline{d}_{\mathrm{sym}}, \underline{d}_{\mathrm{sc}}\}. \tag{9}$$

Required feature. Within $\mathcal{D}_4 = \{11, 12, 21, 22\}$, $\{11, 12\}$ is the unique minimal adequate subset for $\epsilon \in [0.135, 0.168)$ pp (Appendix B.1). Thus, $m_j s_k$ is required alongside $m_j m_k$, where j denotes resampling and k denotes JPEG compression. At a fixed marginal mean, changes in s_k are proportional to changes in the variance of JPEG severity, so $m_j s_k$ retains variance information while retaining operator identity (Appendix F.1). Appendix C explains the possible noncommutativity of resampling and JPEG compression through the input dependence of JPEG reconstruction error.

5 Beyond the primary target family

Section 4 analyzes surrogates on the primary exponential tilt target set. We next consider product mixtures and use matched-mean perturbations outside the tilt family to assess which basis terms are required.

5.1 Continuous product mixtures

Let $\{P_j^p\}_{p=1}^K$ and $\{P_k^q\}_{q=1}^L$ denote evaluated marginal distributions whose $K \times L$ product targets are included in the certificate. Let $P^{pq} := P_j^p \otimes P_k^q$. For $\pi \in \Delta^{K-1}$ and $\tau \in \Delta^{L-1}$, define $P_j^\pi := \sum_p \pi_p P_j^p$, $P_k^\tau := \sum_q \tau_q P_k^q$, and $P^{\pi,\tau} := P_j^\pi \otimes P_k^\tau$. Because F and any surrogate C built from products of marginal moments are bilinear in (π, τ), the residual at a product mixture is a convex combination of the residuals at the evaluated pairs. Hence,

$$\sup_{\pi,\tau} |F(P^{\pi,\tau}) - C(P^{\pi,\tau})| = \max_{p,q} |F(P^{pq}) - C(P^{pq})| \le \mathcal{U}_C. \tag{10}$$

The error bound certified on the finite target set holds uniformly over the entire family of product mixtures without additional model evaluations (Appendix G.4).

5.2 CERTIFYING FEATURE REQUIREMENTS BEYOND MARGINAL MEANS

Testing whether marginal means retain enough information requires targets outside the primary tilt family. Within the exponential tilt family, the marginal mean determines the distribution, so results on that family cannot exclude arbitrary functions of the ordered means. Therefore, we construct $P_\pm$ with density ratios $w_\pm = 1 \pm c_0 q_2$, where $c_0 = \frac{2}{5}$. Both distributions have $\mathbb{E}[A] = \frac{1}{2}$ and mean feature $m = 0$, while their second-moment features are $s = \pm c_0$. Every probability mixture of exponential tilts has $s \geq 0$. Because P_- has $s = -c_0$, any product target with marginal P_- lies outside the product mixture family in Section 5.1 (Appendices F.1 and F.2).

To turn this construction into a class test, we perturb one operator at a time while fixing the other marginal at low or high severity. The perturbations have opposite deviations from the reference density. Linearity of F gives opposite interaction responses at identical ordered marginal means. A mean-only surrogate $\psi(m_j, m_k)$ assigns both targets the same prediction and incurs worst-case error at least half the absolute difference between their responses (Appendix F.1).

For ResNet-50, C_2 has certified error at most 0.092 pp on these pairs, while the unrestricted mean-only class has minimax error at least 0.193 pp and is excluded at 0.15 pp. JPEG perturbations produce larger response magnitudes than resampling perturbations (Figure 4B). On the JPEG pairs, $m_k = 0$ and s_k changes sign, so the $m_j s_k$ term changes sign with the response. By contrast, the zero surrogate is adequate for ViT-B/16 on the same matched-mean pairs (Appendix B.2).

To extend the analysis beyond the fixed amplitude, we vary $w_c = 1 + cq_2$ over its full admissible interval $\mathcal{I}_2 = [-\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}]$ while holding the other marginal distributions fixed. Linearity in c extends the matched-mean bounds. The worst-case error of C_2 scales with the largest absolute amplitude, while the mean-only minimax error scales with the interval width (Appendix F.3). Over this interval, C_2 is no longer certified adequate for ResNet-50, although the mean-only class is ruled out. For ViT-B/16, C_2 remains certified adequate, whereas the mean-only class is not ruled out (Figure 4A).

6 OPERATOR ORDER SELECTION

We use surrogate error certificates to bound excess classification error in operator order selection. The required approximation accuracy depends on the ordering margin.

Recall $V(P) = \mathrm{Err}_{j\to k}(P) - \mathrm{Err}_{k\to j}(P) = \mathbb{E}_P[D] \in [-1, 1]$, where D is the paired error difference from Eq. (1). We predict $\widetilde{V}(P) = \widehat{V}_{\mathrm{add}}(P) + C(P)$, set $\delta = \mathrm{sign}(\widetilde{V}(P))$, and break ties at zero deterministically. The value $\delta = -1$ selects $j \to k$, and $\delta = +1$ selects $k \to j$. The excess classification error relative to the better order is

$$\mathrm{ExcessError}(\delta) = \mathrm{Err}_{\mathrm{chosen}}(P) - \min\{\mathrm{Err}_{j\to k}(P), \mathrm{Err}_{k\to j}(P)\} = \max\{0, -\delta V(P)\}.$$

Proposition 3 (Excess classification error bound)**.** *If $|F(P) - C(P)| \leq \epsilon$ and $|\widehat{V}_{\mathrm{add}}(P) - V_{\mathrm{add}}(P)| \leq \eta$, then*

$$|\widetilde{V}(P) - V(P)| \leq \epsilon + \eta, \quad \mathrm{ExcessError}(\delta) \leq \big(\epsilon + \eta - |\widetilde{V}(P)|\big)_+ \leq \epsilon + \eta, \tag{11}$$

where $(x)_+ \coloneqq \max\{x, 0\}$.

If $|\widetilde{V}(P)| > \epsilon + \eta$, the predicted sign equals the true sign and the excess classification error is zero. For bilinear C and $\widehat{V}_{\mathrm{add}}$, Corollary G.1 gives a uniform bound on excess classification error over the product mixture family in Section 5.1.

Order selection with interaction terms. For the additive confidence interval $[L_t^{\mathrm{add}}, U_t^{\mathrm{add}}]$, set $\widehat{V}_{\mathrm{add}}(P_t) = \frac{1}{2}(L_t^{\mathrm{add}} + U_t^{\mathrm{add}})$ and $\eta_t = \frac{1}{2}(U_t^{\mathrm{add}} - L_t^{\mathrm{add}})$. For each target, the order is certified whenever $|\widetilde{V}(P_t)| > \max\{|L_t - C(P_t)|, |U_t - C(P_t)|\} + \eta_t$.

For both blur/JPEG models, the surrogates in Table 1 satisfy the margin condition at every target and certify which order has lower error. They correct all three order selection errors made by the additive approximation. The corrected choices have zero certified excess error (Figure 5B, Appendix B.5). The target-specific margin condition also certifies the orders selected by the two-term surrogates, showing that decision certification can require less uniform accuracy than surrogate adequacy.

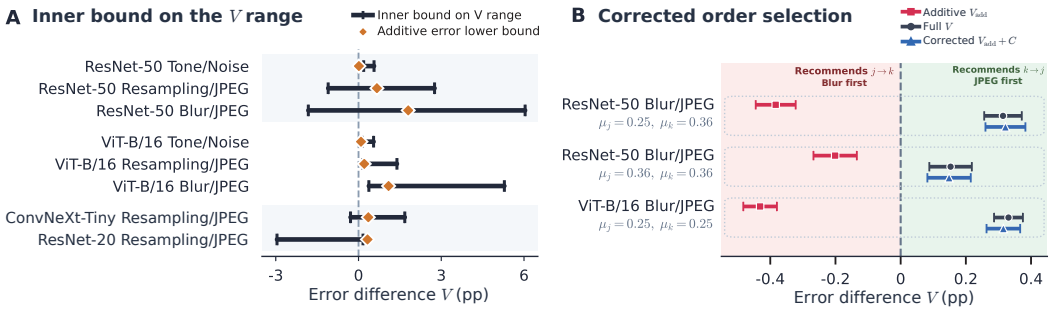


Figure 5: **Consequences for operator order.** (A) Inner bounds for the range of V and lower bounds for the maximum additive approximation error. (B) Blur/JPEG targets where the additive baseline selects the suboptimal order. The interaction surrogates in Table 1 correct each decision.

7 Related Work and Discussion

7.1 Related Work

Prior work predicts out-of-distribution performance from accuracy relationships (Miller et al., 2021, Liang et al., 2023), model agreement (Baek et al., 2022), confidence information (Garg et al., 2022, Guillory et al., 2021), and statistics of feature distributions (Deng & Zheng, 2021). Vision studies also document interactions among augmentations and corruptions (Hendrycks & Dietterich, 2019, Hendrycks et al., 2020, Cubuk et al., 2019, Mintun et al., 2021). Data augmentation methods learn compositions of domain-specific transformations (Ratner et al., 2017), and theoretical work on combinatorial distribution shift examines product structure and bilinear embeddings (Simchowitz et al., 2023). We study finite-sample certification of interaction surrogates and their feature classes.

Functional ANOVA and related response models characterize interactions (Hooker, 2007, Hutter et al., 2014, Lengerich et al., 2020, Friedman & Popescu, 2008, Chastaing et al., 2012). Polynomial chaos and variance-based sensitivity analysis motivate orthogonal expansions (Sobol', 1993; 2001, Xiu & Karniadakis, 2002, Sudret, 2008). Sparse expansions commonly select terms through fitting criteria (Blatman & Sudret, 2011), whereas our class and term-omission lower bounds certify when a restricted surrogate class cannot meet tolerance.

The procedure combines empirical Bernstein bounds (Maurer & Pontil, 2009, Audibert et al., 2009) with Chebyshev approximation and LP duality (Cheney, 1966, Boyd & Vandenberghe, 2004). The robust refit builds on worst-case optimization, set membership, and model invalidation under bounded uncertainty (Bertsimas et al., 2011, Milanese & Vicino, 1991, Smith & Doyle, 1992, Harirchi & Ozay, 2018). Related work calibrates predictive risk (Angelopoulos et al., 2025), transfers generalization guarantees through certifiable surrogates (Bazinet et al., 2026), and separates training error from capacity error in verification and validation (Jakeman et al., 2026). Our certificates distinguish parameter misfit from feature inadequacy.

7.2 Discussion and Conclusion

Establishing feature inadequacy requires ruling out all coefficient choices, but our duality framework reduces this to bounding a single dual score. Because the certifying vector is orthogonal to every feature column, a score certified above tolerance isolates response variation that no refit can capture and requires any adequate expansion to add a feature column nonorthogonal to that vector. Exchanged and matched-mean target pairs provide these certificates by testing discarded distinctions. More importantly, what a representation must retain depends on the response and its downstream use. The one-term class excluded on ImageNetV2 suffices on ImageNet-R, while two-term surrogates certify blur/JPEG order selection despite error upper bounds above the surrogate tolerance. Requiring only linearity in the coefficients and simultaneous response bounds, the certification framework provides a basis for extensions to interactions among three or more pipeline parameters and correlated target distributions.

REPRODUCIBILITY STATEMENT

Proofs and experimental details are provided in the appendices. Code will be made publicly available upon acceptance.

AI USE STATEMENT

In this work, ChatGPT was used for language proofreading, language and format polishing, and generating plotting and code for figures and certain experiments. Generative AI was not used for research ideation, theoretical derivations, or experimental designs. All AI-assisted text and code were reviewed, tested, and verified by the authors. The authors take full responsibility for the final content of this paper.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009. doi: 10.1016/j.tcs.2009.01.016. URL https://www.sciencedirect.com/science/article/pii/S030439750900067X.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J. Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19274–19289, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7a8d388b7a17df480856dff1cc079b08-Abstract-Conference.html.

Mathieu Bazinet, Valentina Zantedeschi, and Pascal Germain. Bound to disagree: Generalization bounds via certifiable surrogates. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 491–520. PMLR, 2026. URL https://proceedings.mlr.press/v337/bazinet26a.html.

Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011. doi: 10.1137/080734510. URL https://doi.org/10.1137/080734510.

Géraud Blatman and Bruno Sudret. Adaptive sparse polynomial chaos expansion based on least angle regression. *Journal of Computational Physics*, 230(6):2345–2367, 2011. doi: 10.1016/j.jcp.2010.12.021.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Gaëlle Chastaing, Fabrice Gamboa, and Clémentine Prieur. Generalized hoeffding–sobol decomposition for dependent variables—application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448, 2012. doi: 10.1214/12-EJS749. URL https://doi.org/10.1214/12-EJS749.

E. W. Cheney. *Introduction to Approximation Theory*. McGraw-Hill, New York, 1966.

Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Cubuk_AutoAugment_Learning_Augmentation_Strategies_From_Data_CVPR_2019_paper.html.

Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15069–15078, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/

html/Deng_Are_Labels_Always_Necessary_for_Classifier_Accuracy_Evaluation_CVPR_2021_paper.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008. doi: 10.1214/07-AOAS148.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2201.04234.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1134–1144, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Guillory_Predicting_With_Confidence_on_Unseen_Distributions_ICCV_2021_paper.html.

Farshad Harirchi and Necmiye Ozay. Guaranteed model-based fault detection in cyber-physical systems: A model invalidation approach. *Automatica*, 93:476–488, 2018. doi: 10.1016/j.automatica.2018.03.040.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1gmrxHFvB.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021b. URL https://arxiv.org/abs/2009.03300.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Giles Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007. doi: 10.1198/106186007X237892.

Qi Huangfu and J A Julian Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018.

Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32(1) of *Proceedings of Machine Learning Research*, pp. 754–762. PMLR, 2014. URL https://proceedings.mlr.press/v32/hutter14.html.

John D. Jakeman, Lorena A. Barba, Joaquim R. R. A. Martins, and Thomas O'Leary-Roseberry. Verification and validation for trustworthy scientific machine learning. *Machine Learning: Science and Technology*, 7(2):025055, 2026. doi: 10.1088/2632-2153/ae59ec. URL https://doi.org/10.1088/2632-2153/ae59ec.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://cave.cs.toronto.edu/kriz/cifar.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165.

Benjamin Lengerich, Sarah Tan, Chun-Hao Chang, Giles Hooker, and Rich Caruana. Purifying interaction effects with the functional ANOVA: An efficient algorithm for recovering identifiable additive models. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2402–2412. PMLR, 2020.

Weixin Liang, Yining Mao, Yongchan Kwon, Xinyu Yang, and James Zou. Accuracy on the curve: On the nonlinear correlation of ML performance between data subpopulations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20706–20724. PMLR, 2023. URL https://proceedings.mlr.press/v202/liang23d.html.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. doi: 10.1109/CVPR52688.2022.01167.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Mario Milanese and Antonio Vicino. Optimal estimation theory for dynamic systems with set membership uncertainty: An overview. *Automatica*, 27(6):997–1009, 1991. doi: 10.1016/0005-1098(91)90134-N.

John P. Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7721–7735, 2021. URL https://proceedings.mlr.press/v139/miller21b.html.

Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2102.11273.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8026–8037. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115.

Alexander J. Ratner, Henry R. Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. Learning to compose domain-specific transformations for data augmentation. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3239–3249. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f26dab9bf6a137c3b6782e562794c2f2-Abstract.html.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400, 2019.

Max Simchowitz, Abhishek Gupta, and Kaiqing Zhang. Tackling combinatorial distribution shift: A matrix completion perspective. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 3356–3468. PMLR, 2023. URL https://proceedings.mlr.press/v195/simchowitz23a.html.

Roy S. Smith and John C. Doyle. Model validation: A connection between robust control and identification. *IEEE Transactions on Automatic Control*, 37(7):942–952, 1992. doi: 10.1109/9.148346.

Ilya M. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiment*, 1(4):407–414, 1993.

Ilya M. Sobol'. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1–3):271–280, 2001. doi: 10.1016/S0378-4754(00)00270-6.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2106.10270.

Bruno Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, 93(7):964–979, 2008. doi: 10.1016/j.ress.2007.04.002. URL https://doi.org/10.1016/j.ress.2007.04.002.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Ross Wightman, Hugo Touvron, and Hervé Jégou. ResNet strikes back: An improved training procedure in timm. In *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*, 2021. URL https://openreview.net/forum?id=NG6MJnVl6M5.

Dongbin Xiu and George Em Karniadakis. The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002. doi: 10.1137/S1064827501387826.

APPENDIX CONTENTS

A Notation **15**

B Experimental results **16**
B.1 Class lower bounds and minimal adequate basis subsets 16
B.2 Perturbation results 17
B.3 Sensitivity to sample size and failure probability 17
B.4 Language results and omission of interactions 17
B.5 Operator order results 18
B.6 ImageNet-R refitting results 19

C Experimental setup **19**
C.1 Datasets, models, and operators 19
C.2 Target distributions and estimation 20
C.3 Numerical implementation 20

D Language evaluation setup **21**
D.1 Sampling parameters and target 21
D.2 Statistical units 21

E Proofs and constructions for Sections 2 and 3 **21**
E.1 Projection error bound 21
E.2 Class lower bound and robust refit 22
E.3 Summary-based approximation bounds and structured classes 23
E.4 Left-nullspace dual vectors and structured class bounds 24

F Perturbation constructions and proofs **24**
F.1 Quadratic perturbations, positivity, and the mean-only optimum 24
F.2 Separation from all mixtures of exponential tilts 25
F.3 Extension to all admissible quadratic amplitudes 25
F.4 Cubic perturbations and quadratic truncation 26

G Statistical and decision guarantees **26**
G.1 Empirical Bernstein bounds for independent, nonidentically distributed observations 26
G.2 Marginal-reference form of the additive component 27
G.3 Proof of the excess classification error bound 27
G.4 Product mixtures of evaluated marginals 27

A NOTATION

Table 3: Core notation.

Symbol	Meaning
P_0, P, Q, w_ℓ, ρ	Reference, target, and proposal measures with their density ratios.
q_u, $\lambda_{\ell u}$, m_ℓ, s_ℓ	Shifted-Legendre basis and coefficients. The mean and second-moment features are expectations of q_1 and q_2.
h_P, $F(P)$, $\mathbf{F}$, $V(P)$, $V_{\mathrm{add}}(P)$	Interaction weighting function, interaction response, response vector, expected response, and additive component.
β_{uv}, $x_{uv}(P)$, S, $\mathcal{D}_4$	Basis coefficient, distributional product feature, selected subset, and candidate feature set.
C, $\mathbf{C}$, C_2, $\mathcal{G}_S$, $\mathcal{G}_X$	Surrogate, prediction vector, two-term surrogate, basis class, and its design-matrix representation.
$\mathcal{G}_{\mathrm{sc}}$, $\mathcal{G}_{\mathrm{sym}}$, $\mathcal{G}_\phi$	One-term, exchange-symmetric, and summary-measurable surrogate classes.
$\mathcal{T}$, T, M	Target set, its size, and the number of coordinates in a joint evaluation.
$\mathcal{B}$, $\mathbf{z}$, $\mathbf{r}$, $\mathcal{K}$	Response confidence box, its center and radius, and its intersection with dual-score interval constraints.
E_C, $\mathcal{L}_C$,$\mathcal{U}_C$	True worst-case surrogate error and certified lower and upper bounds.
$d_{\mathcal{G}_X}(\mathbf{F})$, $\underline{d}_X^{\mathrm{box}}$, $\underline{d}_X$, $\overline{e}_X(\mathcal{B})$	Class distance, response-box and combined class lower bounds, and robust-refit error upper bound.
ν	Dual vector in the left nullspace of X.
$S_{u,m}$, B_m, n_m, μ_m	Coordinate score, deterministic bound, number of independent units, and average score expectation.
$\widetilde{V}$, $\widehat{V}_{\mathrm{add}}$, δ, η, $\mathrm{ExcessError}(\delta)$	Predicted error difference, estimated additive component, order choice, additive-component error bound, and excess classification error.

Algorithm 1 Finite-sample surrogate certification

Input. Response and auxiliary scores satisfying Proposition 2, score bounds B_m, numerical error bounds Δ_m, $X \in \mathbb{R}^{T\times d}$, $\mathbf{C} \in \mathcal{G}_X$, tolerance $\epsilon > 0$, failure probability α, and a finite dual vector set $\mathcal{V}_X$ fixed before evaluation as in Section 3.2.

Output. Certification status, bounds, and any computed refit.

1. **for** each $\nu \in \mathcal{V}_X$ **do**
2. Form $S_{u,\nu} = \sum_t \nu_t S_{u,t}$ using shared evaluation units.
3. **end for**
4. Let M count all coordinates in the joint evaluation.
5. **for** $m = 1, \dots, M$ **do**
6. Compute $\bar{S}_m$, $\hat{\sigma}_m^2$, and r_m^{stat} using Eq. (8) and sample count n_m.
7. Set $[L_m, U_m] \leftarrow [\bar{S}_m - r_m^{\mathrm{stat}} - \Delta_m, \bar{S}_m + r_m^{\mathrm{stat}} + \Delta_m]$.
8. **end for**
9. For language batch scores, rescale the intervals using Eq. (13).
10. Form $\mathcal{B} = \prod_{t=1}^{T}[L_t, U_t]$ and compute $\mathcal{L}_C$,$\mathcal{U}_C$ using Eq. (3).
11. **if** $\mathcal{U}_C \le \epsilon$ **then**
12. **return** Adequate surrogate and $[\mathcal{L}_C, \mathcal{U}_C]$.
13. **end if**
14. Solve Eq. (5) for $\underline{d}_X^{\mathrm{box}}$ and a maximizing dual vector.
15. Combine this bound with the dual score intervals using Eq. (7) to obtain $\underline{d}_X$. Retain a certifying dual vector ν_*.
16. **if** $\underline{d}_X > \epsilon$ **then**
17. **return** Feature inadequacy, $[\mathcal{L}_C, \mathcal{U}_C]$, $\underline{d}_X$, and ν_*.
18. **end if**
19. Solve the robust refit LP (6) for (b_*, u_*).
20. **if** $\mathcal{L}_C > \epsilon$ **and** $u_* \le \epsilon$ **then**
21. Set status to Parameter misfit.
22. **else**
23. Set status to Unresolved.
24. **end if**
25. **return** status, $[\mathcal{L}_C, \mathcal{U}_C]$, $\underline{d}_X$, ν_*, and (b_*, u_*).

Unresolved records uncertainty in the diagnosis. A refit satisfying $u_* \le \epsilon$ is certified adequate.

B Experimental results

B.1 Class lower bounds and minimal adequate basis subsets

For ResNet-50 on ImageNetV2 under resampling/JPEG shifts, ν_{sym} is orthogonal to every exchange-symmetric surrogate, including every one-term surrogate cm_jm_k. Its dual score interval gives $|\nu_{\mathrm{sym}}^{\top}\mathbf{F}| \ge 0.181$ pp. The diagonal dual vector ν_{diag} tightens the one-term class bound to $\underline{d}_{\mathrm{sc}} \ge 0.185$ pp.

The response box gives lower bounds of 0.165 and 0.170 pp. Accounting for covariance between response coordinates raises these bounds to 0.181 and 0.185 pp.

A dual vector that isolates a basis term provides an error lower bound for every surrogate that omits the term after refitting the remaining coefficients. The omission lower bounds are $\underline{d}_{-11} \ge 0.439$ pp and $\underline{d}_{-12} \ge 0.168$ pp. The two-term surrogate C_2 has $\mathcal{U}_{C_2} \le 0.135$ pp. Both omission lower bounds exceed 0.15 pp, while C_2 meets the tolerance. Applying the same omission test to ResNet-50 under blur/JPEG shifts rules out the class that omits s_jm_k at 0.15 pp.

Comparison by number of terms. For ResNet-50 under resampling/JPEG shifts, least squares, Chebyshev fitting, and lasso (Tibshirani, 1996) all select the two-term basis $\{11, 12\}$. For each term count, Table 4 reports the smallest error upper bound among candidate surrogates using at most that many terms. The final two columns give lower bounds for the one-term and exchange-symmetric classes.

Setting	Best error upper bound 0	1	2	3	4	64	Class lower bound One-term	Exchange-symmetric
ResNet-50 / Tone/Noise	0.121	0.118	0.118	0.118	0.118	0.118	0.032	0.032
ResNet-50 / Resampling/JPEG	0.749	0.336	0.135	0.110	0.110	0.110	0.185	0.181
ViT-B/16 / Tone/Noise	0.172	0.090	0.090	0.090	0.090	0.090	0.018	0.018
ViT-B/16 / Resampling/JPEG	0.332	0.235	0.161	0.161	0.161	0.161	0.060	0.060
ConvNeXt-Tiny / Resampling/JPEG	0.420	0.120	0.057	0.044	0.040	0.037	0.062	0.044
ResNet-20 / Resampling/JPEG	0.405	0.161	0.134	0.124	0.108	0.091	0.232	0.213
ResNet-50 / Blur/JPEG	1.895	0.621	0.177	0.163	0.135	0.127	0.514	0.493
ViT-B/16 / Blur/JPEG	1.167	0.329	0.201	0.148	0.148	0.148	0.240	0.222
ConvNeXt-Tiny / Resampling/JPEG initial	0.418	0.256	0.231	0.224	0.224	0.224	0.068	0.050
ResNet-20 / Resampling/JPEG initial	0.413	0.309	0.249	0.249	0.249	0.249	0.212	0.173

Table 4: **Surrogate bounds by number of terms.** Values are in pp and rounded outward to three decimals. Rows labeled "initial" use coefficients before refitting.

Surrogate	ResNet-50	ViT-B/16
Zero (best mean-only)	$[0.193, 0.291]$	$[0.066, 0.104]$
C_2 (terms 11 and 12)	$[0.000, 0.092]$	$[0.000, 0.032]$

Table 5: **Surrogate error intervals on the ImageNetV2 matched-mean pairs (pp).** The first row is also the minimax error interval for the unrestricted mean-only class. Each model's intervals have simultaneous 95% coverage within its evaluation.

B.2 Perturbation results

Cubic perturbations. Appendix F.4 gives the cubic perturbation construction, and Table 6 reports the resulting bounds. The upper bounds remain below the 0.15 pp tolerance at both amplitudes.

B.3 Sensitivity to sample size and failure probability

Changing α from 0.05 to 0.01 changes the vision surrogate upper bounds in Table 4 by at most 0.006 pp. The one-term and exchange-symmetric classes remain ruled out, but the ViT blur/JPEG three-term surrogate is no longer certified adequate at 0.15 pp. At $\alpha = 0.001$, the maximum upper-bound change is 0.012 pp. For ResNet-50 under resampling/JPEG shifts, the one-term and exchange-symmetric classes remain ruled out.

Evaluation sample size. Figure 6 gives the ResNet-50 resampling/JPEG bounds for prefixes with increasing numbers of evaluation replicates. The fitted C_2, confidence allocation, and numerical error bounds are held fixed. Both class lower bounds exceed 0.15 pp at 192 replicates, while C_2 is first certified at 384 replicates. Therefore, the class can be ruled out with fewer replicates than are needed to certify the two-term surrogate.

B.4 Language results and omission of interactions

Table 7 reports how the language assessment changes with tolerance. The zero surrogate is certified adequate at 2.0 pp and remains Unresolved at the smaller tabulated tolerances. At the target that

Model	Error interval at $c = \pm 0.3$ (pp)	Upper bound at $\lvert c\rvert = \frac{1}{\sqrt{7}}$ (pp)
ResNet-50	$[0.015, 0.086]$	0.108
ViT-B/16	$[0.023, 0.076]$	0.095

Table 6: **Matched-moment cubic perturbations.**

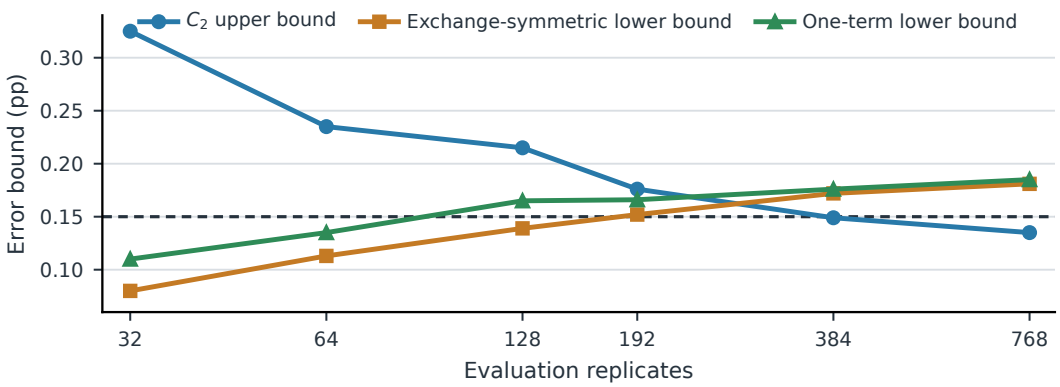


Figure 6: **Error bounds by evaluation sample size for ResNet-50 under resampling/JPEG shifts.** Each point uses a prefix of the evaluation replicates. Coverage is assessed separately at each sample size.

Tolerance (pp)	Terms	One-term class	Exchange-symmetric class	Zero surrogate
0.25	–	Unresolved	Unresolved	Unresolved
0.50	–	Unresolved	Unresolved	Unresolved
1.00	64 (spline)	Unresolved	Adequate	Unresolved
2.00	0 (zero)	Adequate	Adequate	Adequate

Table 7: **Language assessment across tolerances.** The Terms column reports the fewest terms among surrogates certified adequate. A dash denotes none.

determines its error upper bound, the bounded range term in Eq. (8) is the largest component of the confidence radius.

B.5 Operator order results

Order selection with interaction terms. With the tie rule of Section 6, the paired improvement from the additive choice $\delta_t^{(0)} = \mathrm{sign}(\widehat{V}_{\mathrm{add}}(P_t))$ to the corrected choice $\delta_t^{(1)} = \mathrm{sign}(\widetilde{V}(P_t))$ is

$$\mathrm{Err}_{\delta_t^{(0)}}(P_t) - \mathrm{Err}_{\delta_t^{(1)}}(P_t) = \frac{\delta_t^{(1)} - \delta_t^{(0)}}{2} V(P_t).$$

This expression preserves cancellation when the selected orders agree ($\delta_t^{(1)} = \delta_t^{(0)}$) and avoids widening the interval through separate bounds. The margin condition holds at every blur/JPEG target under its target-specific bound, whereas the global bound leaves some ResNet-50 margins uncertified.

Setting	Inner bound on V range (pp)	Certified orders $j \to k$ / $k \to j$	Additive error lower bound (pp)	Interaction signs LL LH HL HH
ResNet-50 / Tone/Noise	$[0.19, 0.56]$	$0/16$	0.013	$0\ 0\ -\ 0$
ResNet-50 / Resampling/JPEG	$[-1.10, 2.75]$	$5/10$	0.666	$+\ -\ -\ +$
ViT-B/16 / Tone/Noise	$[0.14, 0.54]$	$0/16$	0.089	$+\ 0\ -\ +$
ViT-B/16 / Resampling/JPEG	$[0.17, 1.39]$	$0/15$	0.207	$+\ -\ -\ +$
ConvNeXt-Tiny / Resampling/JPEG	$[-0.29, 1.67]$	$2/13$	0.354	$+\ -\ -\ +$
ResNet-20 / Resampling/JPEG	$[-2.95, 0.16]$	$13/2$	0.326	$-\ -\ +\ +$
ResNet-50 / Blur/JPEG	$[-1.82, 6.04]$	$6/10$	1.802	$+\ -\ -\ +$
ViT-B/16 / Blur/JPEG	$[0.38, 5.28]$	$0/16$	1.089	$+\ -\ -\ +$

Table 8: **Execution-order variation across vision settings.** The V interval is an inner bound on its range over targets. Counts report certified order preferences. Interaction signs follow LL, LH, HL, and HH, with 0 marking an uncertified sign.

Table 9: **Order selection under blur and JPEG.** In each row, the additive approximation selects blur before JPEG even though $V > 0$ favors JPEG before blur.

Setting	$(\mathbb{E}[A_j], \mathbb{E}[A_k])$	V (pp)	V_{add} (pp)	C (pp)	$V_{\mathrm{add}} + C$ interval (pp)
ResNet-50	$(0.25, 0.36)$	$[0.256, 0.372]$	$[-0.445, -0.322]$	0.705	$[0.260, 0.383]$
ResNet-50	$(0.36, 0.36)$	$[0.087, 0.219]$	$[-0.268, -0.134]$	0.350	$[0.081, 0.216]$
ViT-B/16	$(0.25, 0.25)$	$[0.286, 0.376]$	$[-0.483, -0.379]$	0.746	$[0.263, 0.367]$

B.6 ImageNet-R refitting results

On the primary target set, the two transferred surrogates exceed the 0.15 pp tolerance. Both two-term refits have error upper bounds below the one-term class lower bound of 0.080 pp (Table 10). On the matched-mean targets, all surrogates in Table 10 meet the tolerance, and both two-term refits have upper bounds below the unrestricted mean-only class lower bound. One-term surrogates predict zero on these targets.

Table 10: **Evaluation and refitting on ImageNet-R.** Error intervals in pp for the primary and matched-mean target sets. All intervals share a simultaneous 95% confidence box. One-term surrogates use $m_j m_k$, and two-term surrogates also use $m_j s_k$.

Coefficients	Surrogate	Primary targets $[\mathcal{L}_C, \mathcal{U}_C]$ (pp)	Matched-mean targets $[\mathcal{L}_C, \mathcal{U}_C]$ (pp)
ImageNetV2 fit	One-term	$[0.450, 0.496]$	$[0.088, 0.113]$
ImageNetV2 fit	Two-term	$[0.458, 0.512]$	$[0.123, 0.146]$
Chebyshev refit	One-term	$[0.078, 0.125]$	$[0.088, 0.113]$
Chebyshev refit	Two-term	$[0.001, 0.046]$	$[0.013, 0.034]$
Robust refit	One-term	$[0.099, 0.145]$	$[0.088, 0.113]$
Robust refit	Two-term	$[0.034, 0.077]$	$[0.026, 0.049]$

C Experimental setup

C.1 Datasets, models, and operators

The vision experiments evaluate pretrained ResNet-50, ViT-B/16, and ConvNeXt-Tiny models on ImageNetV2 MatchedFrequency (Recht et al., 2019, Wightman et al., 2021, Steiner et al., 2022, He et al., 2016, Dosovitskiy et al., 2021, Liu et al., 2022). The ViT-B/16 model is pretrained on ImageNet-21k and fine-tuned on ImageNet-1k. We use the preprocessing configuration distributed with each pretrained model. Both operator orders begin from the same input crop and share all applicable random draws. We evaluate a public ResNet-20 on the CIFAR-10 test set of 10000 images (Krizhevsky, 2009) using its supplied preprocessing configuration.

The ImageNet-R evaluation uses a stratified sample of 10000 images selected before fitting and certification. It retains the ResNet-50 model and resampling/JPEG operators used for ImageNetV2. Labels retain their ImageNet class indices in the model's 1000-class output space.

Tone applies $\mathrm{Tone}_a(x) = \mathrm{clip}(x)^{2^{2a-1}}$, and noise applies $\mathrm{Noise}_b(x, Z) = \mathrm{clip}(x + 0.10bZ)$ with a shared pixelwise Gaussian array Z. Resampling uses bicubic downsizing to $\lfloor D_{\mathrm{px}}(1 - 0.75a) + 0.5 \rfloor$ pixels per side before resizing to the original input size, where D_{px} is the model input width. JPEG maps severity to quality $\lfloor 100 - 95b + 0.5 \rfloor \in [5, 100]$. Gaussian blur uses $\sigma = 3a$ pixels. Pixel values are clipped to the valid image range where required by the transformation.

Resampling/JPEG order. For fixed severities a and b, let R_a denote bicubic downsampling followed by resizing. Write $J_b(x) = x + e_b(x)$, where $e_b(x)$ is the JPEG reconstruction error. Before clipping and rounding, R_a is linear, so

$$J_b(R_a x) - R_a(J_b x) = e_b(R_a x) - R_a e_b(x).$$

Because e_b depends on its input, the equality $e_b(R_a x) = R_a e_b(x)$ need not hold.

C.2 Target distributions and estimation

The exponential tilt density on $[0, 1]$ is $w_\gamma(a) = \frac{e^{\gamma a}}{\mathcal{Z}(\gamma)}$, where $\mathcal{Z}(\gamma) = \frac{e^\gamma - 1}{\gamma}$ for $\gamma \neq 0$ and $\mathcal{Z}(0) = 1$. We use two positive and two negative tilts with mean features ± 0.48 and ± 0.88. Their 16 products and the reference distribution define the primary set of 17 targets. A randomized replicate evaluates the dataset once with parameter pairs and image-level randomness drawn independently.

The vision proposal Q approximates an equally weighted mixture of the extreme product tilts. Let γ_- and γ_+ denote the negative and positive extreme parameters. The ideal mixture density is $\frac{1}{4}(w_{\gamma_-}(a) + w_{\gamma_+}(a))(w_{\gamma_-}(b) + w_{\gamma_+}(b))$. The proposal density ratio satisfies $\rho = \frac{dQ}{dP_0} \geq \frac{1}{4}$. This lower bound supplies the deterministic score bounds in Proposition 2. Numerical approximation error from the proposal is included in the confidence bounds. Evaluations sampled from the reference distribution use $Q = P_0$ and $\rho = 1$.

Tensor coefficients are estimated from fitting data drawn from the uniform reference distribution. Candidate sets and coefficients are fixed using data separate from certification. The candidates include low-order basis subsets, refits, and a flexible spline benchmark.

Evaluation design and sample sizes. Fitting or refitting data are disjoint from certification data. Each pair lists fitting and certification replicate counts, in that order. The joint evaluation of four vision settings uses $(4, 768)$ for ResNet-50 resampling/JPEG and $(4, 192)$ for the other settings. ConvNeXt-Tiny uses $(4, 1152)$ initially and $(1152, 768)$ after refitting. CIFAR-10 uses $(4, 192)$ initially and $(192, 768)$ after refitting. Each of the two blur/JPEG settings uses $(16, 384)$.

The matched-mean study uses 64 certification replicates for ResNet-50 and 192 for ViT-B/16. The cubic study uses 64 certification replicates per model. Both perturbation studies use coefficients fitted in the primary vision experiment. The language evaluation uses $(4, 277)$. ImageNet-R refits use 16 fitting replicates and 768 certification replicates.

Response and dual score intervals have joint 95% coverage within each evaluation. Coefficients fitted using one evaluation are certified using a separate evaluation. The primary vision evaluation, language evaluation, perturbation studies, and ImageNet-R evaluation use separate coverage events.

C.3 Numerical implementation

Experiments use PyTorch/timm (Paszke et al., 2019, Wightman et al., 2021), Pillow, and vLLM (Kwon et al., 2023) with deterministic random seeds. LPs are solved with HiGHS (Huangfu & Hall, 2018). Numerical approximation error from feature moments, importance weights, and LP computations is included in the interval bounds.

D LANGUAGE EVALUATION SETUP

D.1 SAMPLING PARAMETERS AND TARGET

We evaluate Qwen2.5-1.5B-Instruct (Qwen et al., 2024) on a stratified MMLU sample (Hendrycks et al., 2021b) using zero-shot multiple-choice prompting. The parser extracts the first standalone answer letter, and the response is the error indicator $D_i = \mathbf{1}\{Y_i \neq y_i\} \in \{0, 1\}$.

We apply pre-softmax temperature scaling followed by nucleus (top-p) truncation (Holtzman et al., 2020). The parameters are $T_{\text{temp}} = 0.1 + 1.4A_{ij} \in [0.1, 1.5]$ and $p = 0.5 + 0.5A_{ik} \in [0.5, 1.0]$, where A_{ij} and A_{ik} are independent $\text{Uniform}[0, 1]$ draws for each question and evaluation replicate.

Questions are evaluated in a fixed batching scheme. To define $\bar{f}_i(a, b)$, we fix question i's sampling parameters at (a, b) and average over the other questions' sampling parameters and the generation randomness. The interaction target is

$$F_{\text{text}}(P_j, P_k) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{(A_{ij},A_{ik})\sim P_0}\left[\bar{f}_i(A_{ij}, A_{ik})(w_j(A_{ij}) - 1)(w_k(A_{ik}) - 1)\right]. \tag{12}$$

This construction isolates each question's sampling parameters while averaging over reference draws for the other questions in its vLLM batch.

D.2 STATISTICAL UNITS

The $N = 10000$ questions form $J = \lceil \frac{N}{b} \rceil$ batches in each of $R = 277$ replicates. With nominal batch size $b = 64$, this gives $n = JR$ sampling units at the batch level. Weighted sums for each batch are divided by b. To compute the batch scores, the final incomplete batch is padded with zero contributions. If $\bar{S}_t$ is the pooled batch mean and Δ_t is the numerical error bound in batch-score units, the question-level interval is

$$[L_t, U_t] = \kappa\left[\bar{S}_t - r_t^{\text{stat}} - \Delta_t,\ \bar{S}_t + r_t^{\text{stat}} + \Delta_t\right], \quad \kappa = \frac{Jb}{N}. \tag{13}$$

Proposition 2 applies to the batch scores without assuming independence among questions within a batch.

E PROOFS AND CONSTRUCTIONS FOR SECTIONS 2 AND 3

E.1 PROJECTION ERROR BOUND

For a finite index set $S \subset \{1, 2, \dots\}^2$, let $H_S := \text{span}\{q_u \otimes q_v : (u, v) \in S\}$. Let g_S and $h_{P,S}$ denote the orthogonal projections of g and h_P onto H_S, and define $\widehat{C}_S(P) = \sum_{(u,v)\in S} \widehat{\beta}_{uv} x_{uv}(P)$. The absolute surrogate error is $\mathcal{E}(P, \widehat{C}_S) := |F(P) - \widehat{C}_S(P)|$.

Proposition E.1 (Geometric error decomposition)**.** *Surrogate error decomposes as*

$$\mathcal{E}(P, \widehat{C}_S) \leq \underbrace{\|g - g_S\|_{L^2(P_0)} \varrho_S(P)}_{\textit{truncation error bound}} + \underbrace{\sum\nolimits_{(u,v)\in S} |x_{uv}(P)|\,|\beta_{uv} - \widehat{\beta}_{uv}|}_{\textit{coefficient error bound}}, \tag{14}$$

where $\varrho_S(P)^2 = \chi^2(P_j\|P_{0,j})\chi^2(P_k\|P_{0,k}) - \sum_{(u,v)\in S} x_{uv}(P)^2 \geq 0$.

Proof of Proposition E.1. Because g is the interaction component in the functional ANOVA decomposition, it is orthogonal to the constant and to functions of either coordinate alone. Hence, for every $(u, v) \in S$,

$$\beta_{uv} = \langle g, q_u \otimes q_v\rangle_{L^2(P_0)} = \langle f, q_u \otimes q_v\rangle_{L^2(P_0)}.$$

The tensor coefficient of $h_P = (w_j - 1)(w_k - 1)$ is

$$\begin{aligned}\langle h_P, q_u \otimes q_v\rangle_{L^2(P_0)} &= \mathbb{E}_{P_{0,j}}[(w_j - 1)q_u]\,\mathbb{E}_{P_{0,k}}[(w_k - 1)q_v] \\ &= \mathbb{E}_{P_j}[q_u]\,\mathbb{E}_{P_k}[q_v] = x_{uv}(P),\end{aligned}$$

where the zero mean of each q_u, $u \geq 1$, under the reference marginal removes the -1 terms. Therefore, the orthogonal projections onto $H_S = \operatorname{span}\{q_u \otimes q_v : (u,v) \in S\}$ are

$$g_S = \sum_{(u,v)\in S} \beta_{uv}(q_u \otimes q_v), \quad h_{P,S} = \sum_{(u,v)\in S} x_{uv}(P)(q_u \otimes q_v).$$

Using $F(P) = \langle g, h_P \rangle_{L^2(P_0)}$ and orthogonality between H_S and $H_S^{\perp}$ gives

$$\begin{aligned} F(P) - \widehat{C}_S(P) &= \langle g - g_S, h_P - h_{P,S} \rangle_{L^2(P_0)} \\ &\quad + \sum_{(u,v)\in S} (\beta_{uv} - \widehat{\beta}_{uv}) x_{uv}(P). \end{aligned}$$

Taking absolute values and applying the triangle and Cauchy–Schwarz inequalities gives

$$\mathcal{E}(P, \widehat{C}_S) \leq \|g - g_S\|_{L^2(P_0)} \, \|h_P - h_{P,S}\|_{L^2(P_0)} + \sum_{(u,v)\in S} |x_{uv}(P)| \, |\beta_{uv} - \widehat{\beta}_{uv}|.$$

It remains to identify the norm of the omitted part of h_P. Product structure gives

$$\begin{aligned} \|h_P\|^2_{L^2(P_0)} &= \mathbb{E}_{P_{0,j}}[(w_j - 1)^2] \, \mathbb{E}_{P_{0,k}}[(w_k - 1)^2] \\ &= \chi^2(P_j \| P_{0,j}) \, \chi^2(P_k \| P_{0,k}). \end{aligned}$$

Because the tensor basis is orthonormal,

$$\|h_{P,S}\|^2_{L^2(P_0)} = \sum_{(u,v)\in S} x_{uv}(P)^2.$$

Finally, $h_{P,S}$ is the orthogonal projection of h_P onto H_S, so Pythagoras gives

$$\begin{aligned} \|h_P - h_{P,S}\|^2_{L^2(P_0)} &= \|h_P\|^2_{L^2(P_0)} - \|h_{P,S}\|^2_{L^2(P_0)} \\ &= \chi^2(P_j \| P_{0,j}) \chi^2(P_k \| P_{0,k}) - \sum_{(u,v)\in S} x_{uv}(P)^2 \\ &= \varrho_S(P)^2. \end{aligned}$$

□

E.2 Class lower bound and robust refit

Let $\mathcal{B} = \prod_{t=1}^{T} [L_t, U_t] \subset \mathbb{R}^T$ and $\mathcal{G}_X = \{Xb : b \in \mathbb{R}^d\}$ with design matrix $X \in \mathbb{R}^{T \times d}$.

Proof of Proposition 1. For any coefficient vector $b \in \mathbb{R}^d$, projection onto the product box is coordinatewise. Therefore,

$$\begin{aligned} \min_{\mathbf{y}\in\mathcal{B}} \|\mathbf{y} - Xb\|_\infty &= \max_{1\leq t\leq T} \min_{y_t \in [L_t, U_t]} |y_t - (Xb)_t| \\ &= \max_t \max\{0, \, L_t - (Xb)_t, \, (Xb)_t - U_t\}. \end{aligned}$$

Introducing the epigraph variable $e \geq 0$ and minimizing over b gives the primal LP in Eq. (4),

$$\min_{b\in\mathbb{R}^d, \, e\geq 0} e \quad \text{s.t.} \quad \mathbf{L} - e\mathbf{1} \leq Xb \leq \mathbf{U} + e\mathbf{1}. \tag{15}$$

To derive the dual, attach nonnegative multipliers $\nu_+ \in \mathbb{R}^T_+$ and $\nu_- \in \mathbb{R}^T_+$ to the lower and upper bound constraints, respectively, and $\omega \geq 0$ to $-e \leq 0$.

$$\begin{aligned} \mathcal{L}(b, e, \nu_+, \nu_-, \omega) &= e + \nu_+^\top(\mathbf{L} - Xb - e\mathbf{1}) + \nu_-^\top(Xb - \mathbf{U} - e\mathbf{1}) - \omega e \\ &= \nu_+^\top \mathbf{L} - \nu_-^\top \mathbf{U} + b^\top X^\top(\nu_- - \nu_+) + e\big(1 - \mathbf{1}^\top(\nu_+ + \nu_-) - \omega\big). \end{aligned}$$

The infimum over b and e is finite when $X^\top(\nu_+ - \nu_-) = 0$ and $\mathbf{1}^\top(\nu_+ + \nu_-) \leq 1$. With $\mathbf{z} = \frac{1}{2}(\mathbf{L} + \mathbf{U})$ and $\mathbf{r} = \frac{1}{2}(\mathbf{U} - \mathbf{L})$, the dual objective is $(\nu_+ - \nu_-)^\top \mathbf{z} - (\nu_+ + \nu_-)^\top \mathbf{r}$. Set $\nu := \nu_+ - \nu_-$. Since

$r_t \geq 0$, subtracting $\min(\nu_{+,t}, \nu_{-,t})$ from both multipliers preserves their difference and feasibility without decreasing the objective. Thus, an optimum can be chosen with disjoint supports, for which

$$\nu_{+,t} + \nu_{-,t} = |\nu_{+,t} - \nu_{-,t}| = |\nu_t|.$$

The condition $\mathbf{1}^\top(\nu_+ + \nu_-) \leq 1$ becomes $\|\nu\|_1 \leq 1$, and $X^\top(\nu_+ - \nu_-) = 0$ becomes $X^\top \nu = 0$ ($\nu \in \mathrm{null}(X^\top)$). Under these substitutions, the dual objective becomes

$$\max_{\nu \in \mathrm{null}(X^\top),\, \|\nu\|_1 \leq 1} \left\{ \nu^\top \mathbf{z} - \sum_{t=1}^{T} |\nu_t| r_t \right\}. \tag{16}$$

Taking $b = 0$ and any $e > \max\{0, \max_t L_t, \max_t(-U_t)\}$ makes the primal inequalities strict. Therefore, strong duality gives the same optimum for Eqs. (15) and (16), namely $e^* = \underline{d}_X^{\mathrm{box}}$. For fixed b, the worst-case error over the box is

$$\sup_{\mathbf{y} \in \mathcal{B}} \|\mathbf{y} - Xb\|_\infty = \max_t \max\{|L_t - (Xb)_t|, |U_t - (Xb)_t|\}.$$

This error is at most $u \geq 0$ if and only if $\mathbf{U} - u\mathbf{1} \leq Xb \leq \mathbf{L} + u\mathbf{1}$. Minimizing over b and u gives Eq. (6). □

The robust refit produces a surrogate with a uniform error bound over $\mathcal{B}$. Any feasible pair (b, u) certifies worst-case error at most u. The confidence box lower bound and robust refit upper bound differ in quantifiers,

$$\inf_{\mathbf{y} \in \mathcal{B}} \inf_b \|\mathbf{y} - Xb\|_\infty \leq \inf_b \sup_{\mathbf{y} \in \mathcal{B}} \|\mathbf{y} - Xb\|_\infty.$$

Gap between class bounds. For $r_{\max} \coloneqq \|\mathbf{r}\|_\infty$ and $d_z \coloneqq \inf_b \|\mathbf{z} - Xb\|_\infty$, the identities

$$\sup_{\mathbf{y} \in \mathcal{B}} \|\mathbf{y} - Xb\|_\infty = \max_t \big(|z_t - (Xb)_t| + r_t\big), \quad \inf_{\mathbf{y} \in \mathcal{B}} \|\mathbf{y} - Xb\|_\infty = \max_t \big(|z_t - (Xb)_t| - r_t\big)_+$$

give

$$\max\{0, d_z - r_{\max}\} \leq \underline{d}_X^{\mathrm{box}} \leq d_z, \quad \max\{d_z, r_{\max}\} \leq \overline{e}_X(\mathcal{B}) \leq d_z + r_{\max}.$$

Consequently,

$$0 \leq u_* - \underline{d}_X^{\mathrm{box}} \leq 2r_{\max}, \quad u_* = \overline{e}_X(\mathcal{B}).$$

E.3 Summary-based approximation bounds and structured classes

Proposition E.2 (Summary-based approximation bound)**.** *Let a summary mapping $\phi\colon \mathcal{T} \to \mathsf{Z}$ induce equivalence classes $\phi^{-1}(z) \coloneqq \{P_t \in \mathcal{T} : \phi(P_t) = z\}$. For the class $\mathcal{G}_\phi$ of all ϕ-measurable surrogates,*

$$d_{\mathcal{G}_\phi}(\mathbf{F}) \coloneqq \inf_{\mathbf{y} \in \mathcal{G}_\phi} \|\mathbf{F} - \mathbf{y}\|_\infty = \frac{1}{2} \max_{z \in \phi(\mathcal{T})} \mathrm{diam}\{F_t : P_t \in \phi^{-1}(z)\}. \tag{17}$$

For the linear class $\mathcal{G}_X = \{Xb : b \in \mathbb{R}^d\}$, any $\nu \in \mathrm{null}(X^\top)$ with $\|\nu\|_1 \leq 1$ satisfies $d_{\mathcal{G}_X}(\mathbf{F}) \geq |\nu^\top \mathbf{F}|$.

Proof of Proposition E.2. For each z, every $\mathbf{y} \in \mathcal{G}_\phi$ takes one value v_z on $\phi^{-1}(z)$. If the response range within this equivalence class is $[l_z, u_z]$, then

$$\max_{P_t \in \phi^{-1}(z)} |F_t - v_z| \geq \max\{u_z - v_z, v_z - l_z\} \geq \frac{u_z - l_z}{2},$$

with equality at $v_z = \frac{u_z + l_z}{2}$. Choosing this midpoint for every equivalence class gives Eq. (17). For the linear class, $X^\top \nu = 0$ and Hölder's inequality give

$$|\nu^\top \mathbf{F}| = |\nu^\top (\mathbf{F} - Xb)| \leq \|\nu\|_1 \|\mathbf{F} - Xb\|_\infty \leq \|\mathbf{F} - Xb\|_\infty.$$

□

The exchange-symmetric class. For a target set closed under exchanging marginal distributions, $\phi_{\mathrm{sym}}(P) \coloneqq \{P_j, P_k\}$ omits which operator receives each distribution while leaving the operator order unchanged. If $\sigma(t)$ exchanges the two marginal assignments, Eq. (17) gives

$$d_{\mathrm{sym}}(\mathbf{F}) = \frac{1}{2} \max_{1 \le t \le T} |F_t - F_{\sigma(t)}|, \tag{18}$$

and each pair $(t, \sigma(t))$ supplies the corresponding lower bound.

Unrestricted mean-only surrogate class. The matched-mean evaluation uses the summary $\phi_{\mathrm{mean}}(P) \coloneqq (m_j, m_k)$, which retains ordered marginal means and omits distributional shape beyond them. If targets with the same summary have responses differing by more than 2ϵ, Eq. (17) rules out every mean-only surrogate at tolerance ϵ.

E.4 Left-nullspace dual vectors and structured class bounds

Consider matching low and high marginal distributions on the two operators, with $|x_{11,\mathrm{HH}}| + |x_{11,\mathrm{LL}}| > 0$. With coordinates ordered as $(\mathrm{LL}, \mathrm{LH}, \mathrm{HL}, \mathrm{HH})$, the normalized dual vectors are

$$\nu_{\mathrm{sym}} = \left(0,\ \tfrac{1}{2},\ -\tfrac{1}{2},\ 0\right), \quad \nu_{\mathrm{diag}} = \frac{(-x_{11,\mathrm{HH}},\ 0,\ 0,\ x_{11,\mathrm{LL}})}{|x_{11,\mathrm{HH}}| + |x_{11,\mathrm{LL}}|}, \tag{19}$$

where $x_{11,t} \coloneqq x_{11}(P_t) = X_{t,11}$ is the mean product feature at grid coordinate t. The dual vector ν_{sym} has zero inner product with every prediction symmetric under transposition, and ν_{diag} satisfies $\nu_{\mathrm{diag}}^\top X_{\cdot,11} = 0$.

Reflection of the marginal distributions. For matching tilt marginals $P_L = P_{-\gamma}$ and $P_H = P_\gamma$ with $\gamma > 0$, let $\lambda_u = \mathbb{E}_{P_H}[q_u]$. Reflection gives $\mathbb{E}_{P_L}[q_u] = (-1)^u \lambda_u$. On the equally weighted product grid formed by P_L and P_H, the four features in $\mathcal{D}_4$ are mutually orthogonal. The dual scores satisfy

$$\nu_{\mathrm{sym}}^\top \mathbf{F} = \frac{1}{2} \sum_{u,v \ge 1} \left[(-1)^u - (-1)^v\right] \beta_{uv} \lambda_u \lambda_v,$$

$$\nu_{\mathrm{diag}}^\top \mathbf{F} = \sum_{\substack{u,v \ge 1 \\ u+v \text{ odd}}} \beta_{uv} \lambda_u \lambda_v.$$

Both scores retain interaction terms with one odd degree and one even degree. Within $\mathcal{D}_4$, these are the products $m_j s_k$ and $s_j m_k$.

A normalized Kronecker-product vector that annihilates the retained feature columns can isolate a missing term such as $m_j s_k$.

Using a certificate to guide expansion. Let $Z \in \mathbb{R}^{T \times d_{\mathrm{new}}}$ contain additional feature columns on the same target set. Suppose $X^\top \nu = 0$, $\|\nu\|_1 \le 1$, and

$$\nu^\top \mathbf{z} - \sum_{t=1}^{T} |\nu_t| r_t > \epsilon.$$

If $Z^\top \nu = 0$, then $[X\ Z]^\top \nu = 0$, so the same vector remains feasible in Eq. (5) for the expanded design matrix with unchanged objective value. Hence, $\underline{d}^{\mathrm{box}}_{[X\ Z]} > \epsilon$. If a dual score interval gives $\max\{0, L_\nu, -U_\nu\} > \epsilon$, its lower bound continues to certify feature inadequacy whenever $Z^\top \nu = 0$. Under this condition, ν remains orthogonal to every expanded prediction. Thus, an expansion that meets tolerance must include at least one feature column v with $\nu^\top v \neq 0$.

F Perturbation constructions and proofs

F.1 Quadratic perturbations, positivity, and the mean-only optimum

We construct quadratic perturbations using the shifted Legendre polynomial $q_2(a) = \sqrt{5}(6a^2 - 6a + 1)$,

$$w_\pm(a) = 1 \pm c_0 q_2(a), \quad c_0 = \tfrac{2}{5}. \tag{20}$$

On $[0,1]$, q_2 ranges from $-\frac{\sqrt{5}}{2}$ at $a=\frac{1}{2}$ to $\sqrt{5}$ at $a=0$ and $a=1$. Hence, both densities $w_\pm = 1 \pm \frac{2}{5}q_2$ remain strictly positive. Orthogonality under P_0 gives $\int_0^1 q_2\,da = 0$, $\int_0^1 aq_2\,da = 0$, and $\int_0^1 q_2^2\,da = 1$. Thus, each density integrates to one, preserves $\mathbb{E}[A] = \frac{1}{2}$, and has $m = 0$ and $s = \pm\frac{2}{5}$.

Under the distributions $P_\pm$ with densities $w_\pm$, the normalized severity A has variance

$$\mathrm{Var}_{P_\pm}(A) = \frac{1}{12} \pm \frac{1}{6\sqrt{5}}c_0.$$

At a fixed mean severity, $s_\ell = \sqrt{5}\{6\,\mathrm{Var}(A_\ell) + 6(\mathbb{E}[A_\ell] - \frac{1}{2})^2 - \frac{1}{2}\}$ changes in proportion to severity variance. The matched-mean perturbations differ in s_ℓ because their variances differ.

Let $P_{o,h}$ denote the fixed low or high marginal for operator $o \in \{j,k\}$ and level $h \in \{L,H\}$. Perturbing operator k with $P_{j,h}$ held fixed gives $P^\pm_{k,h}$ with responses $\pm\theta_{k,h}$. Perturbing operator j with $P_{k,h}$ held fixed gives $P^\pm_{j,h}$ with responses $\pm\theta_{j,h}$.

$$\begin{aligned} P^\pm_{k,h} &:= P_{j,h} \otimes P_\pm, \quad \theta_{k,h} := c_0\langle g, (w_{j,h}-1)\otimes q_2\rangle_{L^2(P_0)}, \quad F(P^\pm_{k,h}) = \pm\theta_{k,h}, \\ P^\pm_{j,h} &:= P_\pm \otimes P_{k,h}, \quad \theta_{j,h} := c_0\langle g, q_2 \otimes (w_{k,h}-1)\rangle_{L^2(P_0)}, \quad F(P^\pm_{j,h}) = \pm\theta_{j,h}. \end{aligned} \tag{21}$$

Each perturbation pair has the same ordered marginal means and responses $\pm\theta_{o,h}$. Therefore, Proposition E.2 gives zero as the optimal common prediction and

$$E_{C=0} = d_{\mathrm{mean}}(\mathbf{F}) = \max_{o\in\{j,k\},\, h\in\{L,H\}} |\theta_{o,h}|, \tag{22}$$

which exceeds tolerance ϵ whenever $\max_{o,h}|\theta_{o,h}| > \epsilon$.

Two-term surrogate on the matched pairs. For the ResNet-50 surrogate C_2,

$$C_2(P_j \otimes P_k) = \widehat{\beta}_{11} m_j m_k + \widehat{\beta}_{12} m_j s_k,$$

orthogonality gives

$$C_2(P^\pm_{k,h}) = \pm c_0 \widehat{\beta}_{12} m_{j,h}, \quad C_2(P^\pm_{j,h}) = 0. \tag{23}$$

When the JPEG marginal is perturbed, the $m_j s_k$ term changes sign, whereas both terms vanish when the resampling marginal is perturbed.

F.2 Separation from all mixtures of exponential tilts

Within the exponential-tilt family, the raw mean is strictly increasing in γ because $\frac{d\mathbb{E}_\gamma[A]}{d\gamma} = \mathrm{Var}_\gamma(A) > 0$. The mean feature determines the second-moment feature within that family. Let $\mathcal{H}$ be the family of all mixtures of P_γ, where P_γ has density $\frac{e^{\gamma a}}{\mathcal{Z}(\gamma)}$ and $\mathcal{Z}(\gamma) = \int_0^1 e^{\gamma a}\,da$.

Set $b(a) = a^2(1-a)^2$. Since $q_2 = \frac{\sqrt{5}}{2}b''$ and b, b' vanish at $a = 0$ and $a = 1$, two integrations by parts give

$$\int_0^1 q_2(a)e^{\gamma a}\,da = \frac{\sqrt{5}}{2}\gamma^2 \int_0^1 b(a)e^{\gamma a}\,da.$$

Dividing by $\mathcal{Z}(\gamma)$ gives $s(P_\gamma) \geq 0$. By linearity, every $Q \in \mathcal{H}$ has $s(Q) \geq 0$. In contrast, the negative matched-mean perturbation has density $w_- = 1 - \frac{2}{5}q_2$ and second-moment feature $s = -\frac{2}{5}$, so its distribution lies outside $\mathcal{H}$.

F.3 Extension to all admissible quadratic amplitudes

The extrema of q_2 are $-\frac{\sqrt{5}}{2}$ at $a = \frac{1}{2}$ and $\sqrt{5}$ at $a = 0$ and $a = 1$. The condition $1 + cq_2 \geq 0$ is equivalent to $-\frac{1}{\sqrt{5}} \leq c \leq \frac{2}{\sqrt{5}}$. Thus, the admissible amplitude interval is $\mathcal{I}_2 := [-\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}]$. Orthogonality under P_0 ensures unit normalization and mean $\frac{1}{2}$ throughout this interval. Let $t \in \{1,2,3,4\}$ index the four perturbation pairs in Eq. (21). At $c_0 = \frac{2}{5}$, let $\theta_t = F_t(c_0)$ and let p_t denote

the surrogate prediction. By linearity of the Hilbert inner product, $F_t(c) = \frac{c}{c_0}\theta_t$ and its surrogate extension $C_t(c) = \frac{c}{c_0}p_t$. For any closed interval $[a, b] \subseteq \mathcal{I}_2$, the largest residual is attained at a or b,

$$\sup_{t,\, c\in[a,b]} |F_t(c) - C_t(c)| = \frac{\max\{|a|, |b|\}}{c_0} \max_t |\theta_t - p_t|.$$

Every mean-only surrogate assigns the same prediction for every $c \in [a, b]$. The true responses have range diameter $\frac{(b-a)|\theta_t|}{c_0}$. Because the low and high fixed marginal means are distinct and nonzero, the four ordered mean pairs $(m_{j,h}, 0)$ and $(0, m_{k,h})$ for $h \in \{L, H\}$ are distinct. Their midpoint predictions can be chosen independently and achieve half the diameter of the response range, so

$$d_{\text{mean}}([a, b]) = \frac{b-a}{2c_0} \max_t |\theta_t|.$$

Substituting $a = -\frac{1}{\sqrt{5}}$ and $b = \frac{2}{\sqrt{5}}$ gives the full interval bounds on the same coverage event.

$$E_{C_2}(\mathcal{I}_2) = \sqrt{5}\, E_{C_2}(\{\pm 0.4\}), \quad d_{\text{mean}}(\mathcal{I}_2) = \frac{3\sqrt{5}}{4}\, d_{\text{mean}}(\{\pm 0.4\}). \tag{24}$$

F.4 Cubic perturbations and quadratic truncation

The shifted Legendre polynomial $q_3(a) = \sqrt{7}(20a^3 - 30a^2 + 12a - 1)$ is orthogonal to $\{1, q_1, q_2\}$ and has $\|q_3\|_\infty = \sqrt{7}$. The densities $w_3(a) = 1 + cq_3(a)$ are nonnegative for $|c| \le \frac{1}{\sqrt{7}} \approx 0.38$, so $c = \pm 0.3$ is admissible. These perturbations preserve raw mean $\frac{1}{2}$ and satisfy $m = s = 0$. Matched-moment pairs have responses $\theta_t^{(3)}$ and $-\theta_t^{(3)}$. Any surrogate based only on $\phi_2 = (m_j, s_j, m_k, s_k)$ must assign them a common prediction. The optimal worst-case error is $|\theta_t^{(3)}|$ and is attained by the zero prediction.

$$\begin{aligned} w_3(a) &= 1 \pm \tfrac{3}{10} q_3(a), \\ d_{\mathcal{G}_{\phi_2}}(\mathbf{F}) &= \max_{t\in\{1,\dots,4\}} |\theta_t^{(3)}| = E_{C_2}. \end{aligned} \tag{25}$$

G Statistical and decision guarantees

G.1 Empirical Bernstein bounds for independent, nonidentically distributed observations

We condition on the fixed evaluation dataset, the target and proposal distributions, and the independent fitting data. Assume observations indexed by image and replicate are independent and that coordinate sample counts are fixed before evaluation. Within one observation, scores across target distributions, models, and dual scores can exhibit arbitrary covariance. For one coordinate t, let N be the dataset size and R the replicate count, so the sample count is $n_t = NR$. Each image $i \in [N]$ has conditional expectation $\mu_{it} := \mathbb{E}[S_{ir,t} \mid i]$ and is evaluated in R replicates, so the sample average targets the fixed-dataset mean $\mu_t := N^{-1} \sum_{i=1}^N \mu_{it}$. The design matrix has $T = |\mathcal{T}|$ target rows. The simultaneous evaluation contains M coordinates, including interaction responses, auxiliary responses, and dual scores selected before evaluation.

Proof of Proposition 2. Theorem 11 of Maurer & Pontil (2009) applies to independent bounded random variables with heterogeneous underlying distributions. For each coordinate m with $n_m \ge 2$ and $B_m > 0$, let $\widetilde{S}_{u,m} = \frac{S_{u,m}+B_m}{2B_m} \in [0, 1]$. For vision interaction scores, $B_t \ge \frac{\sup|h_t|}{\inf\rho}$ suffices. Apply the theorem to $\widetilde{S}_{u,m}$ and $1 - \widetilde{S}_{u,m}$ with failure probability $\frac{\alpha}{2M}$ for each tail. Writing $\Lambda = \log\frac{4M}{\alpha}$ and using $\widehat{\mathrm{Var}}(\widetilde{S}_{\cdot,m}) = \frac{\hat{\sigma}_m^2}{4B_m^2}$, rescaling gives

$$r_m^{\text{stat}} = 2B_m \left(\sqrt{\frac{2\widehat{\mathrm{Var}}(\widetilde{S}_{\cdot,m})\Lambda}{n_m}} + \frac{7\Lambda}{3(n_m - 1)} \right) = \sqrt{\frac{2\hat{\sigma}_m^2\Lambda}{n_m}} + \frac{14B_m\Lambda}{3(n_m - 1)}.$$

If $B_m = 0$, the score is identically zero and $r_m^{\mathrm{stat}} = 0$. A union bound over all $2M$ tails gives simultaneous coverage of the score means. If Δ_m bounds the numerical discrepancy from the target mean, the intervals $[\bar{S}_m - r_m^{\mathrm{stat}} - \Delta_m, \bar{S}_m + r_m^{\mathrm{stat}} + \Delta_m]$ cover those target means on the same event. Therefore, the response and dual-score intervals share the joint coverage event. The language evaluation uses the batch scaling in Appendix D. □

G.2 Marginal-reference form of the additive component

Using the notation of Section 2, the additive component admits the marginal reference form

$$V_{\mathrm{add}}(P) \coloneqq M_0 + \mathbb{E}_{P_j}[f_j] + \mathbb{E}_{P_k}[f_k] = V(P_j \otimes P_{0,k}) + V(P_{0,j} \otimes P_k) - V(P_0). \tag{26}$$

G.3 Proof of the excess classification error bound

Proof of Proposition 3. The triangle inequality gives

$$|\widetilde{V}(P) - V(P)| \le |\widehat{V}_{\mathrm{add}}(P) - V_{\mathrm{add}}(P)| + |C(P) - F(P)| \le \eta + \epsilon. \tag{27}$$

Under either deterministic tie break, $\delta\widetilde{V}(P) = |\widetilde{V}(P)|$, so

$$-\delta V(P) = \delta\big(\widetilde{V}(P) - V(P)\big) - |\widetilde{V}(P)| \le \epsilon + \eta - |\widetilde{V}(P)|.$$

Taking positive parts gives

$$\mathrm{ExcessError}(\delta) = \max\{0, -\delta V(P)\} \le \big(\epsilon + \eta - |\widetilde{V}(P)|\big)_+ \le \epsilon + \eta. \tag{28}$$

If $|\widetilde{V}(P)| > \epsilon + \eta$, then $\delta V(P) \ge |\widetilde{V}(P)| - |\widetilde{V}(P) - V(P)| > 0$, so the predicted sign is correct and excess classification error is zero. □

The intervals for V_{add} and F in each joint evaluation share a simultaneous coverage event, so the sign and excess classification error bounds hold on that event.

G.4 Product mixtures of evaluated marginals

Using the notation of Section 5.1, bilinearity gives

$$h_{P^{\pi,\tau}} = \sum_{p,q} \pi_p \tau_q h_{P^{pq}}, \quad C(P^{\pi,\tau}) = \sum_{p,q} \pi_p \tau_q C(P^{pq}).$$

Hence,

$$F(P^{\pi,\tau}) - C(P^{\pi,\tau}) = \sum_{p,q} \pi_p \tau_q \big\{F(P^{pq}) - C(P^{pq})\big\}. \tag{29}$$

Therefore,

$$|F(P^{\pi,\tau}) - C(P^{\pi,\tau})| \le \max_{p,q} |F(P^{pq}) - C(P^{pq})|.$$

Concentrating π and τ on a maximizing pair attains the bound. The residual identity requires bilinearity in both marginal distributions.

Corollary G.1 (Uniform excess classification error under continuous mixtures)**.** *Let the vertex errors satisfy $|F(P^{pq}) - C(P^{pq})| \le \mathcal{U}_{pq}$ and $|\widehat{V}_{\mathrm{add}}(P^{pq}) - V_{\mathrm{add}}(P^{pq})| \le \eta_{pq}$. Extend $\widehat{V}_{\mathrm{add}}$ to mixtures by $\widehat{V}_{\mathrm{add}}(P^{\pi,\tau}) = \sum_{p,q} \pi_p \tau_q \widehat{V}_{\mathrm{add}}(P^{pq})$ and set $\delta(\pi,\tau) = \mathrm{sign}(\widetilde{V}(P^{\pi,\tau}))$ with the tie rule of Section 6. Then excess classification error is bounded uniformly across $\Delta^{K-1} \times \Delta^{L-1}$ by*

$$\sup_{\pi \in \Delta^{K-1},\, \tau \in \Delta^{L-1}} \mathrm{ExcessError}(\delta(\pi,\tau)) \le \max_{p,q} (\mathcal{U}_{pq} + \eta_{pq}). \tag{30}$$

If

$$|\widetilde{V}(P^{\pi,\tau})| > \sum_{p,q} \pi_p \tau_q (\mathcal{U}_{pq} + \eta_{pq}),$$

then $\mathrm{ExcessError}(\delta(\pi,\tau)) = 0$.

Proof of Corollary G.1. By bilinearity and the triangle inequality,

$$|\widetilde{V}(P^{\pi,\tau}) - V(P^{\pi,\tau})| \leq \sum_{p,q} \pi_p \tau_q (\mathcal{U}_{pq} + \eta_{pq}) \leq \max_{p,q} (\mathcal{U}_{pq} + \eta_{pq}).$$

Proposition 3 gives the uniform excess classification error bound. Under the stated margin condition, the predicted sign agrees with the true sign, so excess classification error is zero. □