
KINDER: Kernel-based Independence for Fair Representation Learning via Prototype-space Erasure

Abtin Mogharabin^{1,2} Jiaee Cheong^{3,4} Alp Toykan Kaplan⁵ Sinan Kalkan⁵

¹University of Tübingen ²Zuse School ELIZA

³Harvard University ⁴MIT ⁵Middle East Technical University

Abstract

Concept erasure is a prominent approach to achieving fairness in machine learning through the removal of sensitive attributes from learned representations. Prior concept erasure methods typically define debiasing indirectly through the failure of a chosen adversary, probe, or fairness penalty, which makes the target of erasure dependent on a particular decoder family or optimization setup. Rather than depending on a particular architecture, task type, or training paradigm, we introduce KINDER which directly removes sensitive-attribute information from intermediate representations with respect to a chosen family of prototype functionals, and is compatible with any deep model containing an intermediate feature space. KINDER operates in a random Fourier feature space, where it estimates a sensitive subspace from protected group prototypes and projects representations onto its orthogonal complement. Thus, it forms an explicit representation cleaning mechanism distinct from prior fairness control studies. We conduct extensive experiments across diverse settings and show the effectiveness of KINDER on (i) supervised and self-supervised settings, (ii) classification, regression, and image segmentation tasks, and (iii) diverse data modalities, including visual, textual, and tabular data.

1 Introduction

Extensive prior work in fair representation learning (FRL) has shown effectiveness in promoting fairness by learning intermediate representations that retain task-relevant information while reducing dependence on protected attributes [18, 34, 39, 67]. This can be pursued through adversarial invariance, information-theoretic objectives, kernel/ Maximum Mean Discrepancy (MMD) regularization, disentanglement, and certified downstream fairness guarantees [14, 17, 30, 40, 58]. While a stream of FRL studies optimize the representations for a specific downstream task [9, 34], numerous FRL methods learn more *general* representations that are fair [11, 39, 45, 67]. This distinction is essential because learned representations are often reused with unknown downstream tasks, prediction heads, modalities, or deployment settings, where a fairness constraint optimized for one task may not transfer to another [40, 56, 58]. To achieve this reusable form of FRL, a substantial body of research focuses on learning representations that exclude sensitive concepts, such as demographic attributes like gender or race [3, 11, 51]. Such methods are known as *concept erasure* methods [3, 37] and typically entail balancing two competing objectives: (i) removing concept-related information from the representations to ensure they are no longer predictive of the concept after erasure [50, 51], and (ii) preserving as much information as possible from the original representation space [12, 54, 58, 68].

A prominent family of concept erasure methods rely on projections [3, 46, 49–51, 55, 57], and they typically aim to identify a subspace associated with the protected concept [49–51] and remove it from the representation through an explicit projection operator [3, 46, 55, 57] (Fig. 1(a)). However, a key limitation is that the target of erasure is typically defined indirectly through restricted recovery families,

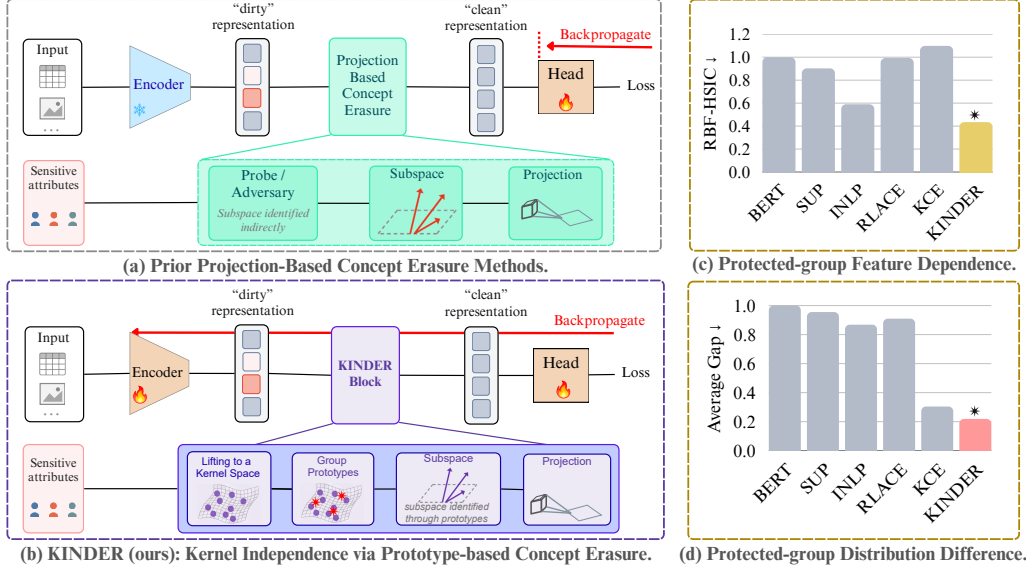


Figure 1: Overview of prior projection-based concept erasure and **KINDER**. (a) Prior methods identify an erasable subspace indirectly through decoder-specific objectives. (b) **KINDER** instead builds protected-group prototypes in a lifted feature space and removes the directions along which those group prototypes differ. (c–d) After cleaning, the representation should contain less protected-group information and smaller group-level distribution differences, which is reflected by lower HSIC and average gap values. Significance tests for these results are reported in Appendix E.2.

such as linear probes or classifiers [3, 49], linear or kernelized adversarial games [50, 51], quadratic decoders [46], or other selected decoder families, rather than through independence at the level of the cleaned representation (**Observation 1 - O1**). A second limitation concerns *how* the erasable subspace is identified. Existing projection-based methods typically recover the removable directions indirectly, through probe-based optimization [49], minmax or adversarial objectives [13, 50, 51], covariance-based closed forms [3], sufficient or spectral projections [55, 57], or kernel/rate-distortion objectives [2]. This leaves open whether group-conditioned nuisance structure can instead be estimated directly from the representation geometry itself, rather than inferred through the selected decoder family used to test erasure (**O2**). Moreover, concept erasure methods are often applied as post-hoc transforms on fixed representations (**O3**) [3, 49, 50, 55]. However, in supervised settings, once the representation is projected, the original task head is no longer matched to the cleaned features, and post-hoc projections can alter the statistical geometry of the feature distribution [27]. This impedes direct comparison to native end-to-end models and can introduce an additional utility bottleneck [12, 68]. Further, realistic fairness settings often require simultaneous control of multiple protected attributes [44, 61], including intersectional or structured subgroup settings where marginal fairness can hide subgroup violations [14, 19, 31, 63], and imbalanced groups [29]. Yet many concept erasure methods are predominantly evaluated in single-attribute regimes [3, 49–51, 55], thus calling into question whether they remain stable and effective when multiple protected attributes are removed simultaneously [57, 66] (**O4**).

Based on observations O1–O4, we address three main research questions: **Research Question 1 (RQ1)**: Can a projection-based operator estimate erasable group-conditioned structure directly at the representation level? **RQ2**: Can projection-based concept erasure be defined at the cleaned-representation level, rather than through the failure of a selected decoder family? **RQ3**: Can such an operator remain useful across post-hoc and end-to-end training, diverse task types, and single, multiple, intersectional, or imbalanced protected-attribute settings?

To address these gaps, we introduce **KINDER** (see Fig. 1), a modular projection-based concept erasure operator that (i) lifts representations into a prototype space, (ii) estimates group-conditioned nuisance structure from explicit group prototypes, and (iii) removes that structure through an orthogonal projection. Unlike prior projection-based methods whose guarantees are tied to linear probes or minmax games [49, 50], kernelized adversaries [51], closed-form linear or quadratic erasure [3, 46], sufficient or spectral projections [55, 57], or iterative RKHS functional erasure [1], **KINDER** defines erasure directly at the level of the cleaned representation itself. This makes **KINDER**

Table 1: Prior methods guarantee erasure only for restricted decoder families. KINDER instead proves a representation-level guarantee, $R \perp_{\mathfrak{F}} S$, meaning independence up to the degree captured by the chosen prototype space $\mathfrak{F}$. Thus under KINDER, removal works at the level of the cleaned representation itself. Further, KINDER has been evaluated on a wider set of data, tasks and settings. Abbreviations: **C**: classification. **R**: regression. **S**: segmentation. **Set.**: settings. **Inter.**: intersectional fairness. **E**: end-to-end. **P**: post-hoc. **NR**: Not Reported. **Rep.**: Reported. **T**: Text. **Tab**: Tabular. **V**: Visual.

Method	Independence type	Erasure target	Subspace source	Data	Task	Set.	Inter.
INLP [49]	$f(R) \perp S, \forall \text{ linear } f$	Linear probe failure	Iterative probes; nullspace	T	C	P	NR
RLACE [50]	$f(R) \perp S, \forall \text{ linear } f$	Linear adversary failure	Linear minmax subspace	T	C	P	NR
KCE [51]	$f(R) \perp S$, for some kernel f	Kernel adversary failure	Kernelized minmax subspace	T	C	P	NR
SAL [55]	$\text{Cov}(f(R), g(S)) \approx 0$, for top spectral linear pairs (f, g)	High-covariance directions	Covariance SVD projection	T	C	P	NR
LEACE [3]	$f(R) \perp S, \forall \text{ linear } f$	Linear classifier failure	Covariance projection	T	C	P	NR
SUP [57]	$f(R) \perp S$ for some chosen f	Unbiased subspace	Sufficient projection	T	C	P	NR
QLEACE [46]	$f(R) \perp S, \forall \text{ quadratic } f$	Quad. decoder failure	Closed-form quad. erasure	T	C	P	NR
KINDER (ours)	Pop. guarantee (affine model): Theorem: $R \perp_{\mathfrak{F}} S \Rightarrow \text{Cor.}: f(R) \perp_{\mathfrak{F}} S \forall \text{ measurable } f$	Prototype-space invariance	Prototype SVD projection	T Tab V	C R S	E P	Rep.

decoder-agnostic and model-agnostic, while preserving the explicit and interpretable structure of projection-based concept erasure.

Our key contributions (KC) are as follows. **KC1 (addresses RQ1 and RQ2)**: We improve projection-based concept erasure generalizability by moving the target of removal from a restricted decoder family to the cleaned representation itself. **KC2 (addresses RQ2 and RQ3)**: We introduce a prototype-space projection mechanism that estimates erasable group-conditioned structure directly from explicit group prototypes in a nonlinear lifted space and makes it easier to recover the shared nuisance structure across *multiple, intersectional*, or *unevenly represented groups*. **KC3 (addresses RQ3)**: KINDER supports both end-to-end training and retrofit adaptation within a unified formulation. **KC4 (addresses RQ1–RQ2)**: We provide a population-level theory, summarized in Section 4.5, showing that KINDER removes group-specific nuisance perturbations in prototype space and reaches prototype-space independence of the cleaned representation from the protected attributes.

2 Related Work

The simplest setting for concept erasure involves categorical concepts, such as the removal of demographic biases related to gender or race in text embeddings [2, 5, 11] and can be achieved by both adversarial training [13, 18, 33, 64] or projection-based methods [2, 5, 13, 20, 49–51, 55]. Post-hoc methods directly modify the learned representations without altering the parameters of the backbone model [1, 49]. In contrast, fine-tuned approaches require access to the model and enforce erasure by finetuning its parameters [55]. Earlier projection-based methods typically rely on linear projection techniques [5, 20, 49, 50] applied in a post-hoc manner, whereas recent work has attempted to extend them to non-linear settings [2, 13, 51, 55].

Comparative Summary: Table 1 compares KINDER against projection-based concept-erasure methods with removal guarantees. KCE extends minimax erasure to a selected kernel family [51]; SAL removes guarded information through spectral covariance directions [55]; LEACE and QLEACE provide closed-form guarantees for linear and quadratic information [3, 46]; and SUP removes information through sufficient projection [57]. Thus, within projection-based concept erasure, existing guarantees are typically tied to restricted decoder families, post-hoc editing regimes, or limited protected-attribute evaluations. KINDER instead proves a more general, decoder-agnostic representation-level guarantee.

Kernel and Reproducing kernel Hilbert space (RKHS) tools have also been used for fair representation learning and prediction debiasing, including fair subspaces [60], MMD-based objectives [17], and CLIP debiasing [16]. These methods show that kernel machinery is useful for fairness, but focus on statistical tests/objectives or prediction-level debiasing, whereas KINDER uses RKHS/ Random Fourier Features (RFF) prototypes to estimate and project out group-conditioned nuisance directions. Orthogonal to this comparison, recent work studies information-theoretic erasure–utility limits of perfect concept erasure [12]. KINDER does not bypass these limits, but targets a different limitation of projection-based erasure, namely, the decoder-family dependence [3, 46].

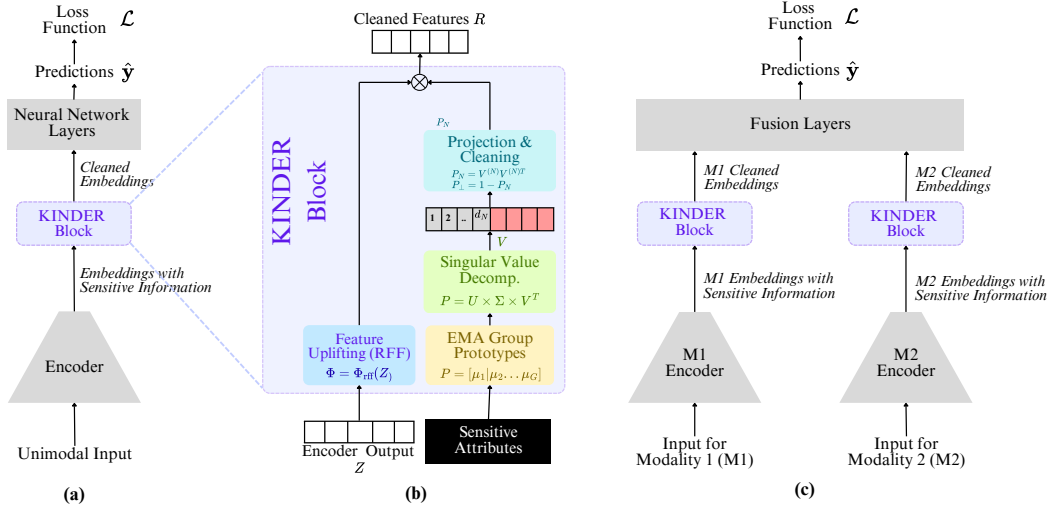


Figure 2: **The KINDER block.** (a) **Unimodal use:** KINDER is inserted between the encoder and the downstream network to transform embeddings containing sensitive variation into cleaned embeddings used by the task loss. (b) **Inside the block,** the encoder output Z is first lifted into a prototype space through random Fourier features (RFF); group-conditioned prototypes are then maintained by exponential moving averages (EMA) using the sensitive attributes, and their dominant variation is summarized by a singular value decomposition (SVD) to estimate the nuisance subspace N . The cleaned representation is obtained by projecting onto its orthogonal complement, $P_{\perp} = I - P_N$, while the removed component corresponds to the estimated group-associated nuisance structure. (c) **Multimodal use:** the same block is applied independently to each modality branch before fusion, so downstream layers operate on cleaned modality-specific embeddings.

3 Preliminaries and Background

Notation and Problem Setting. Let $x \in \mathcal{X}$ be an input, f_{θ} an encoder, and $z = f_{\theta}(x) \in \mathbb{R}^d$ its representation; for two-view or multimodal settings, the same notation is applied to each branch. Let $S \in \{1, \dots, G\}$ be a discrete sensitive attribute, with $s^{(i)}$ denoting the protected-group label for sample i . We use mini-batches $\{(x^{(i)}, s^{(i)})\}_{i=1}^B$, and add $y^{(i)}$ when task supervision is available. For a batch of features $\mathbf{Z} \in \mathbb{R}^{B \times d}$, we define the centering matrix $\mathbf{H}_B = \mathbf{I}_B - \frac{1}{B} \mathbf{1}\mathbf{1}^T$. We define $\mathbf{Z}_c = \mathbf{H}_B \mathbf{Z}$ for the column centered version of $\mathbf{Z}$.

Reproducing Kernel Hilbert Space (RKHS). A Hilbert space $\mathcal{H}$ is a vector space with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, which allows geometric notions such as distance and orthogonality to be defined in possibly infinite-dimensional spaces. An RKHS is a Hilbert space of functions where evaluation is continuous and satisfies $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$, where k is the reproducing kernel. In our setting, bounded measurable kernels k_Z and k_S induce RKHSs $\mathcal{H}_Z$ and $\mathcal{H}_S$, with feature maps $\varphi_Z(z) = k_Z(\cdot, z)$ and $\varphi_S(s) = k_S(\cdot, s)$ for encoder features z and sensitive attributes s . This lifts Z and S into Hilbert spaces where *distributional differences can be expressed through inner products*.

4 KINDER: Methodology

Intuitively, our KINDER block removes group-specific nuisance structure in the lifted prototype space and maps each encoder output to a “cleaned” representation retaining only the shared component of the representation (see Figure 2 – see Appendix B.1 for the full algorithm). We first introduce KINDER for a unimodal setting and extend it for a pair of representations (z_a, z_b) (corresponding to e.g., two views, augmentations, or modalities). KINDER executes three steps:

1. Step 1: Lifting the encoder output into an RKHS space (Section 4.1).
2. Step 2: Prototype Estimation in the RKHS space (Section 4.2).
3. Step 3: Representation projection and cleaning in the RKHS space (Section 4.3).

In Section 4.5, we provide a theoretical analysis showing that, under the corresponding population-level structural model, the cleaned representation becomes independent of the protected attribute at the prototype-space level, which is certified by zero pairwise MMD across groups.

4.1 Step 1: Uplifting the Encoder Features into an RKHS

Kernels on Z and S . On encoder features Z , we use the Gaussian RBF kernel with bandwidth $\gamma > 0$:

$$k_Z(z, z') = \exp\left(-\frac{\|z - z'\|_2^2}{2\gamma^2}\right) = \langle \varphi_Z(z), \varphi_Z(z') \rangle_{\mathcal{H}_Z}, \quad (1)$$

which is a standard choice because it is universal and characteristic on $\mathbb{R}^d$, allowing the lifted representation to capture rich geometric differences between group-conditional distributions in the encoder feature space [42, 59]. On the sensitive attribute, we use the Dirac kernel: $k_S(s, s') = \mathbf{1}\{s = s'\}$, where $\varphi_S(s) = e_s \in \mathbb{R}^G$.

For discrete sensitive attributes, this kernel coincides with the one-hot embedding and makes the RKHS mean $\mu_S = \sum_s \pi_s e_s$, with $\pi_s = \Pr(S = s)$, and the class conditional means $\mathbb{E}[\varphi_Z(z) \mid S = s]$ directly interpretable as group prototypes [42]. This also makes group-conditioned prototype statistics directly interpretable in the lifted space, which is the basis for the nuisance-subspace construction used by KINDER.

Random Fourier features for k_Z . We approximate k_Z with random Fourier features (RFF) [48]. For bandwidth $\gamma > 0$ and RFF dimension M ,

$$\Phi_{\text{rff}}(Z) = \sqrt{\frac{2}{M}} [\cos(\omega_1^\top Z + b_1), \dots, \cos(\omega_M^\top Z + b_M)]^\top \in \mathbb{R}^M, \quad (2)$$

with $\omega_j \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \gamma^{-2} I_d)$ and $b_j \stackrel{\text{iid}}{\sim} \text{Unif}[0, 2\pi]$, where I_d is the $d \times d$ identity matrix. For notational simplicity we write Φ for Φ_{rff} in what follows. All linear operators used in the KINDER block implementation (prototypes, projectors, and cleaned features) act in this lifted space $\mathbb{R}^M$.

4.2 Step 2: Prototype Estimator of the Sensitive Subspace via SVD

Let $X = \Phi_{\text{rff}}(z) \in \mathbb{R}^M$ denote the lifted encoder feature. Rather than fixing the removable structure to a single statistic, KINDER estimates *group-associated variation* through a chosen prototype family. For an order parameter K , let $\mathcal{F}_K = \{\psi_a^{(q)} : q = 1, \dots, K, a \in \mathcal{A}_q\}$ be a finite collection of prototype functionals, where each $\psi_a^{(q)}$ maps lifted features to a vector in $\mathbb{R}^M$. For each group s , order q , and prototype index a , we maintain an EMA prototype

$$p_{s,a}^{(q)} \leftarrow (1 - \rho) p_{s,a}^{(q)} + \rho \cdot \frac{1}{|\mathcal{B}_s|} \sum_{i \in \mathcal{B}_s} \psi_a^{(q)}(X^{(i)}; \mu_s), \quad (3)$$

$$\mathcal{B}_s := \{i : s^{(i)} = s\}.$$

Here, $\rho \in (0, 1]$ controls the EMA update rate. The argument μ_s is used for central higher-order moments; for first-order prototypes it is not needed.

For each prototype statistic, we define its empirical global average $\bar{p}_a^{(q)} = \sum_{s=1}^G \hat{\pi}_s p_{s,a}^{(q)}$, where $\hat{\pi}_s$ is the empirical group frequency. The order- K nuisance subspace is then defined as the span of all centered group-prototype differences:

$$\mathcal{N}_K := \text{span}\left\{p_{s,a}^{(q)} - \bar{p}_a^{(q)} : s = 1, \dots, G, q = 1, \dots, K, a \in \mathcal{A}_q\right\} \subset \mathbb{R}^M. \quad (4)$$

To estimate this subspace, KINDER processes the prototype orders sequentially. For each order q , we form a centered prototype matrix $\tilde{P}_q$ by stacking the rows $\left(p_{s,a}^{(q)} - \bar{p}_a^{(q)}\right)^\top$ over all groups s and prototype indices $a \in \mathcal{A}_q$. For the first order, we directly compute the thin SVD of $\tilde{P}_1$. For higher orders, before applying SVD, we remove the directions already selected from lower-order prototypes:

$$\tilde{P}_{q|<q} = \tilde{P}_q (I_M - Q_{<q} Q_{<q}^\top), \quad (5)$$

where $Q_{<q}$ is the orthonormal basis selected from orders $1, \dots, q-1$. We then compute

$$\tilde{P}_{q|<q} = U^{(q)} \Sigma^{(q)} V^{(q)\top} \quad (6)$$

and add the leading right singular vectors to the nuisance basis. After processing orders $1, \dots, K$, let $Q_{\leq K}$ denote the resulting basis. We define

$$P_N^{(K)} = Q_{\leq K} Q_{\leq K}^\top, \quad P_\perp^{(K)} = I_M - P_N^{(K)}. \quad (7)$$

Thus, higher-order prototypes add new nuisance directions beyond the lower-order subspace, while the cleaning operation remains a single orthogonal projection.

Prototype-order instantiations. The simplest instantiation is first-order KINDER. It uses only the mean prototype $\psi^{(1)}(X) = X$, $\mu_s = \mathbb{E}[X \mid S = s]$, so that Eq. (4) reduces to

$$\mathcal{N}_1 = \text{span}\{\mu_s - \bar{\mu} : s = 1, \dots, G\}.$$

Higher-order variants address residual nuisance directions through additional prototype functionals while using the same projection pipeline. Second-order covariance prototypes are further detailed in Appendix B.4. A feature-level ablation over higher-order prototype families using MOJI dataset is also provided in Appendix E.1.

4.3 Step 3: Representation Projection and Cleaning

Given a batch of lifted features $\mathbf{X} = \Phi_{\text{rff}}(\mathbf{Z}) \in \mathbb{R}^{B \times M}$, we define the cleaned representations as

$$\mathbf{R}^{(K)} = \mathbf{X} P_\perp^{(K)} \in \mathbb{R}^{B \times M}. \quad (8)$$

Only $\mathbf{R}$ is passed to the remainder of the model (e.g., a projection head, classifier, or regressor). R denotes a generic cleaned representation random variable with the same distribution as any R_m .

4.4 Training Objective (Loss)

KINDER is loss-agnostic: after cleaning, the representation is passed to the downstream head or objective without modifying the task loss. For a supervised input x with target y , an encoder f_θ first produces $z = f_\theta(x) \in \mathbb{R}^d$. KINDER then lifts and cleans this feature,

$$R = \text{KINDER}(z, s) = \Phi_{\text{rff}}(z) P_\perp^{(K)} \in \mathbb{R}^M,$$

and the downstream head h_ω is trained with the original supervised objective, $\mathcal{L}_{\text{sup}}^{\text{KINDER}} = \mathcal{L}_{\text{task}}(h_\omega(R), y)$. Thus, KINDER can be used with standard classification, regression, or segmentation losses without changing the task objective. For two-view self-supervised settings, the same block is applied to each view or modality before the contrastive objective; explicit uni-/multimodal SSL formulations are provided in Appendix B.5.

4.5 Theoretical Analysis

In Section C, we provide rigorous theoretical guarantees for KINDER, using a shared-base affine structural model. Our theoretical analysis assumes that a generic representation Z is generated by applying group-specific affine transformations to a latent vector W that is strictly independent of S ($W \perp S$). Under this model, we identify the canonical affine discrepancy subspace, $\mathcal{N}_{\text{aff}}$, which captures the precise linear directions along which group-specific affine images of the same latent point diverge. As shown in Theorem C.3, projecting Z onto the orthogonal complement of $\mathcal{N}_{\text{aff}}$ yields a cleaned representation R^* that is strictly independent of S . We prove $\mathcal{N}_{\text{aff}}$ is the minimal linear subspace whose removal equalizes all groups, thereby preserving maximum latent information.

KINDER extends this affine intuition into a reproducing kernel Hilbert space (RKHS) via prototype families. By mapping representations to a Hilbert space $\mathcal{H}_{\mathfrak{F}}$ using a prototype map $\Psi_{\mathfrak{F}}(Z)$, KINDER forms a cleaned representation R by projecting away the subspace $\mathcal{N}_{\mathfrak{F}}$ spanned by the differences in group prototype means. Theorem C.5 establishes that this projection achieves prototype-level independence, i.e. the cleaned representation is invariant to the protected attribute under the linear kernel of the prototype family ($R \perp_{\mathfrak{F}} S$). Further, Corollary C.6 guarantees that if the prototype functionals are rich enough to identify the relevant group laws in the cleaned representation space, zero MMD implies full probabilistic independence, $R \perp S$. Complete proofs for all statements, including the minimality of the discrepancy subspace, are provided in Appendix C.

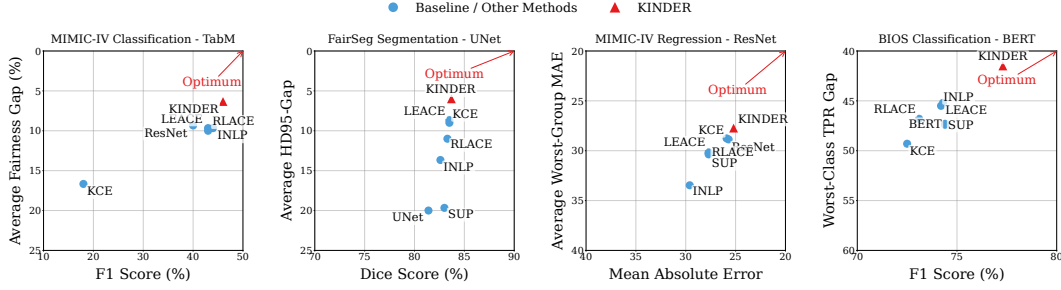


Figure 3: **Performance-fairness tradeoff plots** across classification, segmentation & regression. **Red** triangles denote KINDER; **blue** circles denote baselines and competing post-hoc methods.

5 Experiment settings

Datasets used: *Supervised classification:* Bias in Bios [15] and Moji [69] (for post-hoc evaluation), MIMIC-IV [28], *Supervised regression:* Communities and Crime [41], MIMIC-IV [28], *Supervised segmentation:* FairSeg [62], *SSL classification:* Harvard-FairVLMed [38] (multimodal). CelebA [35] (unimodal).

Fairness metrics: For segmentation, we report

worst-group Dice for overlap fairness [43] and HD95 gap for boundary fairness [6]. For text classification, we report worst-group macro-F1 [7], worst-class TPR gap [26, 52], and RMS class TPR gap [49, 50]. For other classification tasks, we use the aggregated fairness score from [10], averaging statistical parity, equal accuracy, equalized odds, and equal opportunity. For regression, we report MAE gap for group error disparity [24]. Further training, dataset, implementation, and details are provided in Appendix D. Hyperparameters are selected by validation utility (i.e., AUC for classification, MAE for regression, and Dice for segmentation), rather than fairness or erasure scores. Our goal is to evaluate how much protected-attribute dependence can be removed while preserving performance.

6 Results

Across Tables 2–7, we evaluate whether KINDER remains effective under dataset imbalance (Appendix D.1, Table D.8) while preserving utility and improving fairness across single-/multiple-attribute, intersectional, supervised, and self-supervised settings.

Single attribute classification results. On BIOS and MOJI (Table 2), which evaluate KINDER in post-hoc settings, KINDER gives the strongest fairness-sensitive classification results. On BIOS, it improves accuracy and worst-group F1 while reducing the worst-class TPR gap. On MOJI, the accuracy is slightly lower than the BERT baseline, but F1 worst-group, F1, and class-level disparity metrics improve by a larger margin.

Intersectional fairness results. A similar pattern appears on MIMIC-IV (Table 3), where we test the same claim in a setting with multiple protected attributes. The joint KINDER variant gives the best overall balance across the table. It preserves the baseline accuracy, improves F1, reduces the race gap, and improves worst-group accuracy. Further multi-attribute ablations of sequential attributes removal over the baseline methods are provided in Appendix E.3.

In the MIMIC-IV classification results we observe the potential issues that might arise from head mismatch in post-hoc projection methods. When the native KINDER head is replaced with linear

Table 2: **BIOS & MOJI classification results.** WgF1 denotes worst-group macro-F1, WCTG denotes worst-class TPR gap, and RCTG denotes RMS class TPR gap. Arrows indicate whether higher or lower values are better. First row is the baseline model. Rows prefixed with “+” apply the corresponding method on top of the baseline backbone. Best values are in **bold**, second-best values are underlined for all tables.

Method	BIOS					MOJI				
	Acc ↑	F1 ↑	WgF1 ↑	WCTG ↓	RCTG ↓	Acc ↑	F1 ↑	WgF1 ↑	WCTG ↓	RCTG ↓
BERT	80.9	74.4	70.7	47.3	16.2	80.1	74.7	<u>62.1</u>	25.6	17.5
+ INLP	80.7	74.2	70.3	45.5	14.9	79.8	74.7	61.5	25.3	17.9
+ RLACE	79.8	73.1	69.1	46.8	<u>14.8</u>	79.1	73.7	61.0	23.9	16.9
+ LEACE	80.7	74.3	70.5	<u>45.2</u>	14.5	<u>79.9</u>	74.8	61.1	26.3	18.6
+ KCE	79.8	72.5	68.6	49.3	16.5	79.2	73.9	61.4	23.6	<u>16.7</u>
+ SUP	80.8	<u>74.4</u>	<u>71.0</u>	47.4	15.9	79.8	74.6	61.8	<u>23.8</u>	16.8
+ KINDER	81.9	77.3	73.9	41.5	15.2	79.2	77.5	65.3	16.7	14.1

Table 3: **MIMIC-IV classification** results using TabM. Acc and F1 are performance metrics and the remaining columns are fairness metrics. Arrows indicate the preferred direction. KINDER (NH) uses the native end-to-end task head, while KINDER replaces this head with a linear probe as an ablation comparison with other post-hoc erasure baselines. Post-hoc ablations are reported in Appendix E.3.

Method	Acc $\uparrow$	F1 $\uparrow$	Gender Gap $\downarrow$	Race Gap $\downarrow$	Age Gap $\downarrow$	Inter. Gap $\downarrow$	WG Inter. $\uparrow$
TabM	<u>91.9</u>	39.9	2.1	23.2	2.5	<u>46.6</u>	<u>65.8</u>
+ INLP	<u>79.5</u>	43.6	1.6	22.1	5.2	53.3	58.3
+ RLACE	78.5	43.0	2.4	<u>20.7</u>	7.1	53.1	58.3
+ LEACE	78.3	42.6	2.3	21.4	6.2	54.4	58.3
+ KCE	80.1	18.0	0.4	46.2	4.1	73.6	35.7
+ KINDER	80.6	<u>44.7</u>	1.2	22.5	7.1	<u>46.6</u>	68.7
+ KINDER (NH)	92.0	45.5	<u>0.7</u>	15.4	<u>3.3</u>	44.8	68.7

Table 5: **FairSeg segmentation** results on UNet. Dice and HD95 measure performance, Dice measures overlap and HD95 measures boundary error. D-WG and H-Gap measure fairness, where D-WG denotes worst-group Dice and H-Gap denotes HD95 gap.

Method	Perf.		Race		Gender		Ethnicity		Inter.	
	Dice $\uparrow$	HD95 $\downarrow$	D-WG $\uparrow$	H-Gap $\downarrow$	D-WG $\uparrow$	H-Gap $\downarrow$	D-WG $\uparrow$	H-Gap $\downarrow$	D-WG $\uparrow$	H-Gap $\downarrow$
UNet	81.4	6.8	77.4	4.1	81.3	0.7	81.3	1.2	67.5	30.7
+ INLP	82.6	5.5	78.7	3.0	82.6	1.0	82.6	0.1	70.5	20.0
+ RLACE	83.3	5.1	80.0	2.8	83.2	<u>0.3</u>	83.2	<u>0.2</u>	73.1	17.0
+ LEACE	83.5	4.7	80.3	1.2	83.5	0.6	83.4	0.8	<u>76.6</u>	6.0
+ SUP	83.0	5.5	79.1	3.7	83.0	0.8	82.9	1.4	68.7	30.8
+ KCE	83.5	<u>4.5</u>	80.2	1.9	83.3	<u>0.3</u>	83.4	0.5	75.7	6.9
+ KINDER	83.7	4.1	80.4	<u>1.6</u>	83.7	0.1	83.7	0.1	78.2	5.4

probing for comparability with prior erasure baselines, performance drops sharply even though fairness remains competitive. The gap between KINDER (NH) and KINDER suggests that replacing the original task head can impose a substantial utility bottleneck on some datasets. Similar observations were made in our sequential attribute removal ablation in Appendix E.3. This is the setting in which an end-to-end compatible erasure operator is practically advantageous.

Segmentation results. On FairSeg (Table 5), the setting is different from classification as fairness improvements cannot be attributed only to threshold shifts or changes in class scores and the method must preserve key spatial information as well. KINDER improves the overall Dice and HD95 scores relative to the UNet baseline, while improving worst-group Dice and reducing the largest boundary disparities. The strongest gains appear on the intersectional metrics, where the baseline has the largest gap. This supports that KINDER is not tied to any particular classifier head or task loss.

Table 4: **I-JEPA** results on CelebA. Acc. and WG-Inter. measure utility and worst-group intersectional performance, where higher is better; gap metrics measure protected-attribute disparity, where lower is better.

Method	Acc. $\uparrow$	Gender Gap $\downarrow$	Age Gap $\downarrow$	Inter. Gap $\downarrow$	WG-Inter. $\uparrow$
I-JEPA	91.1	6.6	2.7	11.5	87.7
+ INLP	85.7	3.4	8.5	13.7	82.7
+ LEACE	90.0	6.3	5.8	12.0	86.3
+ RLACE	91.8	6.1	2.6	10.8	88.5
+ KCE	91.8	7.0	<u>2.5</u>	11.4	88.3
+ SUP	<u>91.9</u>	<u>6.0</u>	2.7	<u>11.0</u>	88.5
+ KINDER	92.2	6.3	2.3	11.4	89.4

Regression results. On MIMIC-IV regression (Table 6), KINDER improves ResNet MAE and marginal worst-group errors, with intersectional worst-group MAE as the main exception. Similar observations appear for MLP-PLR in the same table. On Communities and Crime (Appendix Table E.12), KINDER gives the best utility and worst-group error, while remaining close to the strongest race-gap result. This is important as some post-hoc baselines, especially INLP and LEACE, reduce the representation aggressively and lose a large amount of predictive information.

Multimodal SSL results. Looking at FairVLMed (Table 7), on FactorCL, KINDER improves accuracy while reducing all reported gaps and gives the strongest worst-intersectional balanced accuracy. On CLIP, KINDER gives the best race, gender, and ethnicity gaps, and matches the best worst-intersectional balanced accuracy. CelebA (Table 4) shows the same unimodal SSL pattern: KINDER gives the best accuracy, age gap, and worst-intersectional performance, while keeping gender and intersectional gaps close to the baseline.

Table 6: **MIMIC-IV regression results on ResNet and MLP-PLR.** MAE and R^2 measure performance; worst-group MAE columns measure fairness. Arrows indicate the preferred direction. Best values are in **bold** and second-best values are underlined.

Method	MAE ↓	R^2 ↑	Race WG-MAE ↓	Gender WG-MAE ↓	Age WG-MAE ↓	Inter. WG-MAE ↓
ResNet	2.59	0.15	3.24	2.61	<u>2.78</u>	<u>5.55</u>
+ INLP	2.96	0.00	3.75	3.01	3.28	10.88
+ LEACE	2.77	0.11	3.25	2.79	3.01	5.91
+ RLACE	2.77	0.11	3.27	2.79	3.02	5.93
+ KCE	<u>2.57</u>	0.15	<u>3.23</u>	<u>2.59</u>	2.84	5.10
+ SUP	2.77	0.11	3.29	2.79	3.03	5.90
+ KINDER	2.52	0.13	3.07	2.53	2.72	7.01
MLP-PLR	2.59	0.11	<u>3.35</u>	<u>2.62</u>	<u>2.84</u>	5.65
+ INLP	2.67	0.11	<u>3.45</u>	2.72	2.83	5.87
+ LEACE	2.69	0.09	3.40	2.73	2.83	5.25
+ RLACE	2.68	0.09	3.43	2.71	2.89	5.05
+ KCE	2.67	0.10	3.50	2.70	2.89	5.29
+ SUP	2.68	0.09	3.50	2.72	2.88	<u>5.22</u>
+ KINDER	2.59	0.12	3.30	2.61	2.85	6.83

Table 7: **FairVLMed multimodal self-supervised learning results** on CLIP and FactorCL. Accuracy measures performance; gap and worst group intersectional accuracy measures fairness. Arrows indicate the preferred direction.

Method	Acc ↑	Race Gap ↓	Gender Gap ↓	Ethnicity Gap ↓	Inter. Gap ↓	WG-Inter. ↑
FactorCL	60.4	12.8	4.3	15.6	95.8	0.0
+ INLP	60.2	7.9	5.6	<u>10.3</u>	76.0	<u>25.0</u>
+ RLACE	60.0	13.7	4.2	16.9	95.9	0.0
+ LEACE	57.3	4.3	3.7	7.3	84.8	0.0
+ KCE	58.8	9.1	5.8	15.7	93.5	0.0
+ SUP	<u>60.6</u>	12.3	2.3	17.7	95.8	0.0
+ KINDER	61.2	<u>6.9</u>	<u>3.0</u>	12.4	<u>83.7</u>	41.5
CLIP	73.6	<u>6.2</u>	4.0	14.2	80.7	41.9
+ INLP	70.5	12.0	5.4	13.5	74.6	50.0
+ RLACE	73.0	7.4	2.8	16.1	80.7	41.9
+ LEACE	69.3	6.7	6.2	18.0	87.1	0.0
+ KCE	<u>73.4</u>	6.3	4.1	18.9	83.6	44.5
+ SUP	<u>73.4</u>	6.3	3.1	<u>12.8</u>	80.2	44.9
+ KINDER	71.6	5.7	1.6	5.9	<u>79.1</u>	50.0

7 Discussion and Conclusion

RQ 1: Representation-level estimation: We show that in KINDER, a projection-based operator is able to estimate erasable group-conditioned structure directly at the representation level by estimating the removable structure from the explicit protected-group prototypes in a lifted prototype space.

RQ 2: Generalizable concept erasure: As a result, KINDER is able to achieve *generalizable concept erasure* by moving the target of removal from a restricted decoder family to the cleaned representation itself. *Empirically*, KINDER is shown to work across a variety of training paradigms (i.e. during training and post training) and architectures (BERT, IJEPa, ResNet, UNet, CLIP, etc.). *Theoretically*, our analysis shows that KINDER removes the nuisance directions responsible for protected-group variation and yields independence from the protected attribute at the level identified by the prototype space. Further, under stronger prototype-exactness conditions, our prototype-level guarantee recovers probabilistic independence, thus separating the main target of erasure from a particular decoder family while preserving the explicit structure of projection-based concept erasure.

RQ 3: All round utility across a variety of training paradigms, tasks and settings: KINDER improves the empirical fairness–utility tradeoff under our utility-first selection protocol [12, 68]. KINDER achieves strong performance across supervised (e.g., MIMIC-IV dataset, Bios & Moji), self-supervised (e.g., FairVLMed, CelebA), unimodal, and multimodal settings, and consistently improves both fairness and performance (see Fig. 3) for a variety of tasks even under *multiple* and *intersectional* protected attributes. Our findings support KINDER as an all-round practical and theoretically grounded approach to *generalizable task-agnostic fair representation learning*.

Limitations: Our experiments are restricted to datasets that contain sensitive attributes which is the common setting in fairness analysis. We only focus on KINDER’s moment order functionals and not other forms of functionals. Future work can extend the analysis on other datasets and new variations of prototype functionals. We provide an extended discussion of the limitations in Appendix F.

Acknowledgments and Disclosure of Funding

Abtin Mogharabin is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. Jiaee Cheong is supported by a DAAD Fellowship. We also gratefully acknowledge the computational resources kindly provided by METU-ROMER, Center for Robotics and Artificial Intelligence.

References

- [1] Ramin Akbari, Milad Afshari, and Vishnu Naresh Boddeti. Oblivator reveals the cost of nonlinear guardedness in concept erasure. *arXiv preprint arXiv:2603.07529*, 2026.
- [2] Somnath Basu Roy Chowdhury, Nicholas Monath, Kumar Avinava Dubey, Amr Ahmed, and Snigdha Chaturvedi. Robust concept erasure via kernelized rate-distortion maximization. *Advances in Neural Information Processing Systems*, 36:43284–43306, 2023.
- [3] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.
- [4] Su Lin Blodgett, Lisa Green, and Brendan O’Connor. Demographic dialectal variation in social media: A case study of African-American English. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics.
- [5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NeurIPS*, 29, 2016.
- [6] Kate Čevora, Ben Glocker, and Wenjia Bai. Quantifying the impact of population shift across age and sex for abdominal organ segmentation. In *MICCAI Workshop on Fairness of AI in Medical Imaging*, pages 88–97. Springer, 2024.
- [7] Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, 2022.
- [8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.
- [9] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Fairrefuse: referee-guided fusion for multimodal causal fairness in depression detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7224–7232, 2024.
- [10] Jiaee Cheong, Abtin Mogharabin, Paul Liang, Hatice Gunes, and Sinan Kalkan. Fairwell: Fair multimodal self-supervised learning for wellbeing prediction. *arXiv preprint arXiv:2508.16748*, 2025.
- [11] Somnath Basu Roy Chowdhury and Snigdha Chaturvedi. Learning fair representations via rate-distortion maximization. *Transactions of the Association for Computational Linguistics*, 10:1159–1174, 2022.
- [12] Somnath Basu Roy Chowdhury, Kumar Avinava Dubey, Ahmad Beirami, Rahul Kidambi, Nicholas Monath, Amr Ahmed, and Snigdha Chaturvedi. Fundamental limits of perfect concept erasure. In *International Conference on Artificial Intelligence and Statistics*, pages 901–909. PMLR, 2025.
- [13] Somnath Basu Roy Chowdhury, Sayan Ghosh, Yiyuan Li, Junier Oliva, Shashank Srivastava, and Snigdha Chaturvedi. Adversarial scrubbing of demographic information for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 550–562, 2021.
- [14] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR, 2019.

- [15] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.
- [16] Sepehr Dehdashtian, Lan Wang, and Vishnu Boddeti. Fairerclip: Debiasing clip’s zero-shot predictions using functions in rkhs. In *The Twelfth International Conference on Learning Representations*.
- [17] Namrata Deka and Danica J. Sutherland. Mmd-b-fair: Learning fair representations with statistical testing. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 9564–9576. PMLR, 25–27 Apr 2023.
- [18] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.
- [19] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th international conference on data engineering (ICDE)*, pages 1918–1921. IEEE, 2020.
- [20] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, 2019.
- [21] Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations*, 2025.
- [22] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pages 18932–18943, 2021.
- [23] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.
- [24] Xudong Han, Aili Shen, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Systematic evaluation of predictive fairness. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 68–81, 2022.
- [25] 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*, pages 770–778, 2016.
- [26] Léo Hemamou and William Coleman. Delivering fairness in human resources ai: Mutual information to the rescue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 867–882, 2022.
- [27] Richard Johansson. What happens to a dataset transformed by a projection-based concept removal method? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17486–17492, 2024.
- [28] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet*. Available online at: [https://physionet.org/content/mimiciv/1.0/\(accessed August 23, 2021\)](https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)), 2020.

- [29] Precious Jones, Weisi Liu, I-Chan Huang, and Xiaolei Huang. Examining imbalance effects on performance and demographic fairness of clinical language models. In *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*, pages 58–68. IEEE, 2025.
- [30] Nikola Jovanović, Mislav Balunovic, Dimitar Iliev Dimitrov, and Martin Vechev. FARE: Provably fair representation learning with practical certificates. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15401–15420. PMLR, 23–29 Jul 2023.
- [31] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 10–15 Jul 2018.
- [32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, 2014.
- [33] Yitong Li, Timothy Baldwin, and Trevor Cohn. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, 2018.
- [34] Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1088–1097, 2022.
- [35] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.
- [36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- [37] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024.
- [38] Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301, 2024.
- [39] Yuhong Luo, Austin Hoag, Xintong Wang, Philip S Thomas, and Przemyslaw A Grabowicz. Fair representation learning with controllable high confidence guarantees via adversarial inference. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- [40] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3384–3393. PMLR, 10–15 Jul 2018.
- [41] Lawrence McClendon and Natarajan Meghanathan. Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal*, 2(1):1–12, 2015.
- [42] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.
- [43] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, 2025.

- [44] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.
- [45] Amisha Priyadarshini and Sergio Gago-Masague. Learning fair representations with kolmogorov-arnold networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 24864–24872, 2026.
- [46] Lucia Quirke and Nora Belrose. Slowing learning by erasing simple features. *arXiv preprint arXiv:2502.02820*, 2025.
- [47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [48] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.
- [49] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*, 2020.
- [50] Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR, 2022.
- [51] Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055, 2022.
- [52] Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. What’s in a name? reducing bias in bios without access to protected attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4187–4195, 2019.
- [53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer, 2015.
- [54] Antoine Saillenfest and Pirmin Lemberger. Nonlinear concept erasure: A density matching approach. In *ECAI 2025*, pages 2138–2145. IOS Press, 2025.
- [55] Shun Shao, Yftah Ziser, and Shay B Cohen. Gold doesn’t always glitter: Spectral removal of linear and nonlinear guarded attribute information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1611–1622, 2023.
- [56] Xudong Shen, Yongkang Wong, and Mohan Kankanhalli. Fair representation: Guaranteeing approximate multiple group fairness for unknown tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):525–538, 2023.
- [57] Enze Shi, Lei Ding, Linglong Kong, and Bei Jiang. Debiasing with sufficient projection: A general theoretical framework for vector representations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5960–5975, 2024.
- [58] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2164–2173. PMLR, 16–18 Apr 2019.

- [59] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.
- [60] Zilong Tan, Samuel Yeom, Matt Fredrikson, and Ameet Talwalkar. Learning fair representations for kernel models. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 155–166. PMLR, 26–28 Aug 2020.
- [61] Huan Tian, Bo Liu, Tianqing Zhu, Wanlei Zhou, and Philip S Yu. Multifair: Model fairness with multiple sensitive attributes. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):5654–5667, 2024.
- [62] Yu Tian, Min Shi, Yan Luo, Ava Kouhana, Tobias Elze, and Mengyu Wang. Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In *The Twelfth International Conference on Learning Representations*, 2024.
- [63] Steven Vethman, Quirine TS Smit, Nina M van Liebergen, and Cor J Veenman. Fairness beyond the algorithmic frame: actionable recommendations for an intersectional approach. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 3276–3290, 2025.
- [64] Liwen Wang, Yuanmeng Yan, Keqing He, Yanan Wu, and Weiran Xu. Dynamically disentangling social bias from task-oriented representations with adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3740–3750, 2021.
- [65] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace’s transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.
- [66] Changdi Yang, Zheng Zhan, Ci Zhang, Yifan Gong, J Liu, X Shen, H Tang, G Yuan, P Zhao, X Lin, et al. Fairsmoe: Mitigating multi-attribute fairness problem with sparse mixture-of-experts. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, 2025.
- [67] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, pages 325–333. PMLR, 2013.
- [68] Han Zhao and Geoffrey J. Gordon. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, 23(57):1–26, 2022.
- [69] Xianda Zhou and William Yang Wang. Mojital: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, 2018.

Table of Contents: Appendix

A More Related Work	16
B Further KINDER Details	16
B.1 KINDER Algorithm	16
B.2 Supervised vs. Semi-supervised Implementation	17
B.3 Specific Block Formulation	17
B.4 Higher-order prototype instantiations	18
B.5 Two-view self-supervised loss formulations	18
C Detailed Theoretical Results	18
D More on Experimental Settings	21
D.1 Datasets	21
D.2 Implementation Details	22
D.3 Compared Method Details	24
D.4 Definitions of Metrics Used	24
E Additional Results	26
E.1 Prototype-order ablation	26
E.2 Feature-level bootstrap significance tests	26
E.3 Sequential MIMIC-IV classification ablation	26
E.4 Communities and Crime Dataset	27
F Limitations	27
G Impact Statement	28

A More Related Work

Projection-based concept-erasure methods. INLP relies on iterative linear probes and a joint-nullspace projection for concept erasure [49]. RLACE formulates concept erasure as a linear minimax game where the concept is removed from the learned subspace [50]. KCE kernelizes the linear minimax game and provides erasure against a selected kernelized adversary [51]. LEACE is a closed-form method that prevents all linear classifiers from detecting a concept while changing the embedding as little as possible [3]. SUP obtains debiased representations by projecting them onto the subspace orthogonal to a sufficient span [57]. QLEACE extends LEACE by removing quadratically available information about a concept from a representation [46]. Oblivator targets nonlinear guardedness by minimizing statistical dependence between learned PLM representations and unwanted attributes [1]. These methods motivate Table 1, where the main distinction is whether erasure is defined through a restricted decoder/adversary family or directly at the cleaned-representation level.

B Further KINDER Details

B.1 KINDER Algorithm

We present the full algorithm in Algorithm 1.

Algorithm 1 KINDER for a batch: first-order prototype subspace projection.

Require: Minibatch representations $\mathbf{Z} \in \mathbb{R}^{B \times d}$ and protected-group labels $\mathbf{s} = (s^{(1)}, \dots, s^{(B)})$, with $s^{(i)} \in \{1, \dots, G\}$

Require: RFF feature map Φ_{rff} defined in Eq. (2)

Require: Removal rank $d_N \leq G - 1$, EMA rate $\rho \in (0, 1]$

Require: Prototype-subspace refresh interval $T_{\text{refresh}} \in \mathbb{N}$ (measured in optimization steps) and current optimization step $t_{\text{step}} \in \mathbb{N}$

Require: Empirical group frequencies $(\hat{\pi}_1, \dots, \hat{\pi}_G)$

Require: Stored mean prototypes $\{\mu_s \in \mathbb{R}^M\}_{s=1}^G$ and prototype basis $\mathbf{U}_N \in \mathbb{R}^{M \times d_N}$ with orthonormal columns

Ensure: Cleaned representations $\mathbf{R} \in \mathbb{R}^{B \times M}$

Step 1: RFF lift and projection

- 1: $\mathbf{X} \leftarrow \Phi_{\text{rff}}(\mathbf{Z})$ ▷ RFF-lifted representations from Eq. (2)
- 2: $P_{\perp} \leftarrow \mathbf{I}_M - \mathbf{U}_N \mathbf{U}_N^{\top}$ ▷ Projector from Eq. (7)
- 3: $\mathbf{R} \leftarrow \mathbf{X} P_{\perp}$ ▷ Cleaned representations from Eq. (8)

Step 2: Mean-prototype update

- 4: **for** each protected group $s = 1, \dots, G$ **do**
- 5: $\mathcal{B}_s \leftarrow \{i : s^{(i)} = s\}$
- 6: **if** $|\mathcal{B}_s| > 0$ **then**
- 7: $\bar{\mathbf{x}}_s \leftarrow \frac{1}{|\mathcal{B}_s|} \sum_{i \in \mathcal{B}_s} \mathbf{X}_{i,:}$ ▷ Batch group mean used in Eq. (3)
- 8: $\mu_s \leftarrow (1 - \rho) \mu_s + \rho \bar{\mathbf{x}}_s$ ▷ EMA update from Eq. (3)
- 9: **end if**
- 10: **end for**

Step 3: SVD refresh of the prototype subspace

- 11: **if** $t_{\text{step}} \bmod T_{\text{refresh}} = 0$ **then**
- 12: $\bar{\mu} \leftarrow \sum_{s=1}^G \hat{\pi}_s \mu_s$ ▷ Global mean for Eq. (4)
- 13: Form $\tilde{\mathbf{P}} \in \mathbb{R}^{G \times M}$ with rows $\tilde{\mathbf{P}}_{s,:} \leftarrow (\mu_s - \bar{\mu})^{\top}$ ▷ Top right singular vectors from Eq. (6)
- 14: $\mathbf{U}_N \leftarrow \text{TOPRIGHTSINGVECS}(\tilde{\mathbf{P}}, d_N)$ ▷ Centered prototype matrix for Eq. (6)
- 15: **end if**
- 16: **return** $\mathbf{R}$

Optimization step: compute the chosen objective $\mathcal{L}(\mathbf{R})$ and backpropagate through the RFF map Φ and the linear cleaning map $\mathbf{X} \mapsto \mathbf{X} P_{\perp}$.

B.2 Supervised vs. Semi-supervised Implementation

In supervised settings, the same cleaned representations R_m can be fed h and trained with a standard loss (e.g., cross-entropy or regression loss) without modifying the KINDER block.

We use a two-stage protocol on semi-supervised settings. In Stage 1 (self-supervised pretraining), we train on all available samples using either the unimodal objective $\mathcal{L}_{\text{KINDER}}^{\text{uni-ssl}}$ or the multimodal objective $\mathcal{L}_{\text{KINDER}}^{\text{mm-ssl}}$, while maintaining prototypes indexed only by s in Eq. (3). In Stage 2 (supervised fine-tuning), we train a task head on the labeled subset using the supervised objective $\mathcal{L}_{\text{KINDER}}^{\text{sup}}$. When discrete task labels are available, Stage 2 also admits a label-aware refinement in which prototypes are indexed by (s, y) , i.e., by replacing $s^{(i)}$ with $(s^{(i)}, y^{(i)})$ in Eq. (3). This changes only the prototypes used to form $P_{\perp}$ and leaves the training objective unchanged. The resulting label-aware version conditions the removable nuisance structure on the task labels, reducing the risk of removing task-relevant variation in supervised classification settings.

B.3 Specific Block Formulation

Unimodal version. For a mini-batch of encoder outputs $\mathbf{Z} = [z^{(1)}, \dots, z^{(B)}]^{\top}$ with labels $\mathbf{s} = (s^{(1)}, \dots, s^{(B)})$, we define the KINDER block operator:

$$\begin{aligned} \text{KINDER}(\mathbf{Z}, \mathbf{s}) &:= \Phi_{\text{rff}}(\mathbf{Z}) P_{\perp} \\ &= \Phi_{\text{rff}}(\mathbf{Z}) (\mathbf{I}_M - \mathbf{U}^{(N)} \mathbf{U}^{(N)\top}). \end{aligned} \tag{B.9}$$

where $U^{(N)} = \mathcal{U}(\{\mu_s\}_{s=1}^G)$ denotes the top d_N right singular vectors obtained from the SVD in Eq. (6). The map $\mathcal{U}$ is defined as

$$\mathcal{U} : (\mathbb{R}^M)^G \rightarrow \mathbb{R}^{M \times d_N} \quad (\text{B.10})$$

$$\{\mu_s\}_{s=1}^G \mapsto [v_1^{(P)} \cdots v_{d_N}^{(P)}].$$

Multimodal version. The KINDER block could also be defined for multimodal methods that benefit from a pair of views (e.g., contrastive learning methods). For the two-view case, we define each view as $m \in \{m_1, m_2\}$. Now, we can define the multi-view version of KINDER block as

$$\text{KINDER}(\mathbf{Z}_{m_1}, \mathbf{Z}_{m_2}, \mathbf{s}) := (\text{KINDER}_{m_1}(\mathbf{Z}_{m_1}, \mathbf{s}), \text{KINDER}_{m_2}(\mathbf{Z}_{m_2}, \mathbf{s})). \quad (\text{B.11})$$

B.4 Higher-order prototype instantiations

The main text defines KINDER for a general order- K prototype family and describes how KINDER works for first order. Here we further the exact second-order case used to motivate higher-order nuisance removal.

A stronger exact second-order version additionally maintains covariance prototypes

$$\Sigma_s = \mathbb{E}[(X - \mu_s)(X - \mu_s)^\top \mid S = s], \quad \bar{\Sigma} = \sum_{s=1}^G \hat{\pi}_s \Sigma_s.$$

The corresponding nuisance subspace is

$$\mathcal{N}_2^{\text{exact}} = \text{span}\{\mu_s - \bar{\mu} : s = 1, \dots, G\} + \sum_{s=1}^G \text{Range}(\Sigma_s - \bar{\Sigma}).$$

Thus, the exact second-order variant removes both location differences and scatter differences across protected groups in the lifted RFF space. Higher-order prototype functionals can be defined analogously by adding further centered group-conditioned statistics before applying the same orthogonal projection step. The selection between the orders during training was done based by a grid search maximizing the validation AUC.

B.5 Two-view self-supervised loss formulations

Let x_a and x_b denote two views of the same instance, corresponding either to two augmentations of one modality or to two paired modalities. Let f_a and f_b be the encoders, producing $z_a = f_a(x_a)$ and $z_b = f_b(x_b)$. For group labels $s = (s^{(1)}, \dots, s^{(B)})$, the KINDER block maps each encoder output to cleaned lifted representations,

$$(\mathbf{R}_a, \mathbf{R}_b) = \text{KINDER}(\mathbf{Z}_a, \mathbf{Z}_b, s), \quad \mathbf{R}_a, \mathbf{R}_b \in \mathbb{R}^M.$$

Any two-view contrastive objective can then be applied without modification:

$$\mathcal{L}_{\text{uni-ssl}}^{\text{KINDER}} = \mathcal{L}_{\text{ctr}}(\mathbf{R}_a, \mathbf{R}_b), \quad \mathcal{L}_{\text{mm-ssl}}^{\text{KINDER}} = \mathcal{L}_{\text{ctr}}(\mathbf{R}_{m_1}, \mathbf{R}_{m_2}).$$

C Detailed Theoretical Results

Let κ be a bounded measurable kernel on the support of a cleaned representation R . For probability measures P, Q on that space, let μ_P^κ denote the corresponding kernel mean embedding, and write

$$\text{MMD}_\kappa(P, Q) := \|\mu_P^\kappa - \mu_Q^\kappa\|_{\mathcal{H}_\kappa}.$$

When κ is characteristic, $\text{MMD}_\kappa(P, Q) = 0$ is equivalent to $P = Q$ [23, 42, 59].

Definition C.1 (Shared-base affine structural model). Take a generic encoder representation $Z \in \mathbb{R}^d$, a discrete protected attribute $S \in \{1, \dots, G\}$ with $\pi_g := \Pr(S = g) > 0$ for every g , and a latent random vector $W \in \mathbb{R}^d$ satisfying

$$W \perp S.$$

For each group $g \in \{1, \dots, G\}$, let $A_g \in \mathbb{R}^{d \times d}$ and $b_g \in \mathbb{R}^d$, and write

$$Z = A_S W + b_S \quad \text{almost surely.}$$

Let

$$\mathcal{H} := \text{span}(\text{supp}(W)) \subseteq \mathbb{R}^d$$

denote the linear span of the support of W .

Definition C.2 (Canonical affine discrepancy subspace). In the setting of Definition C.1, define

$$\mathcal{N}_{\text{aff}} := \text{span}\left(\{b_g - b_h : g, h \in \{1, \dots, G\}\} \cup \bigcup_{g, h \in \{1, \dots, G\}} (A_g - A_h)\mathcal{H}\right).$$

Equivalently,

$$\mathcal{N}_{\text{aff}} = \text{span}\{(A_g - A_h)w + (b_g - b_h) : g, h \in \{1, \dots, G\}, w \in \mathcal{H}\}.$$

Let $P_{\text{aff}, \perp} : \mathbb{R}^d \rightarrow \mathcal{N}_{\text{aff}}^\perp$ denote the orthogonal projector onto $\mathcal{N}_{\text{aff}}^\perp$, and define

$$R^* := P_{\text{aff}, \perp} Z.$$

Definition C.2 collects exactly the linear directions along which two group-specific affine images of the same latent point can differ.

Theorem C.3 (Canonical affine projection theorem). *In the setting of Definition C.1, there exists a measurable map*

$$T : \mathcal{H} \rightarrow \mathcal{N}_{\text{aff}}^\perp$$

such that

$$R^* = T(W) \quad \text{almost surely.}$$

Consequently,

$$R^* \perp S.$$

Moreover, if $\mathcal{M} \subseteq \mathbb{R}^d$ is a linear subspace satisfying

$$P_{\mathcal{M}^\perp}(A_g w + b_g) = P_{\mathcal{M}^\perp}(A_h w + b_h) \quad \text{for every } g, h \in \{1, \dots, G\}, w \in \mathcal{H},$$

then

$$\mathcal{N}_{\text{aff}} \subseteq \mathcal{M}.$$

Hence $\mathcal{N}_{\text{aff}}$ is the smallest linear subspace whose removal equalizes all groups under the affine structural model.

Proof. For every $g, h \in \{1, \dots, G\}$ and $w \in \mathcal{H}$,

$$(A_g w + b_g) - (A_h w + b_h) = (A_g - A_h)w + (b_g - b_h) \in \mathcal{N}_{\text{aff}}.$$

Therefore

$$P_{\text{aff}, \perp}(A_g w + b_g) = P_{\text{aff}, \perp}(A_h w + b_h).$$

Fix a reference group $g_0 \in \{1, \dots, G\}$, and define

$$T(w) := P_{\text{aff}, \perp}(A_{g_0} w + b_{g_0}), \quad w \in \mathcal{H}.$$

The preceding identity shows that this quantity does not depend on the chosen group after projection, so

$$P_{\text{aff}, \perp}(A_g w + b_g) = T(w) \quad \text{for every } g \in \{1, \dots, G\}, w \in \mathcal{H}.$$

Since $W \in \mathcal{H}$ almost surely and $Z = A_S W + b_S$ almost surely,

$$R^* = P_{\text{aff}, \perp} Z = P_{\text{aff}, \perp}(A_S W + b_S) = T(W) \quad \text{almost surely.}$$

Because $W \perp S$, it follows that

$$R^* \perp S.$$

For minimality, let $\mathcal{M} \subseteq \mathbb{R}^d$ satisfy the stated equalization property. Then

$$P_{\mathcal{M}^\perp}((A_g - A_h)w + (b_g - b_h)) = 0 \quad \text{for every } g, h \in \{1, \dots, G\}, w \in \mathcal{H},$$

hence

$$(A_g - A_h)w + (b_g - b_h) \in \mathcal{M} \quad \text{for every } g, h, w.$$

Taking $w = 0$ gives $b_g - b_h \in \mathcal{M}$. Subtracting those vectors then gives

$$(A_g - A_h)w \in \mathcal{M} \quad \text{for every } g, h \in \{1, \dots, G\}, w \in \mathcal{H}.$$

Thus $\mathcal{M}$ contains all generators of $\mathcal{N}_{\text{aff}}$, so

$$\mathcal{N}_{\text{aff}} \subseteq \mathcal{M}.$$

□

Theorem C.3 identifies the canonical linear object whose removal makes all groups agree in the affine structural model. The next definition fixes the prototype space in which KINDER operates.

Definition C.4 (Prototype family and cleaned prototype representation). Fix a prototype family $\mathfrak{F}$, and let

$$\Psi_{\mathfrak{F}}(z) \in \mathcal{H}_{\mathfrak{F}}$$

denote the corresponding prototype map into a Hilbert space $\mathcal{H}_{\mathfrak{F}}$. For each group $g \in \{1, \dots, G\}$, define

$$\mu_g^{\mathfrak{F}} := \mathbb{E}[\Psi_{\mathfrak{F}}(Z) \mid S = g], \quad \bar{\mu}^{\mathfrak{F}} := \sum_{g=1}^G \pi_g \mu_g^{\mathfrak{F}}.$$

Let

$$\mathcal{N}_{\mathfrak{F}} := \text{span}\{\mu_g^{\mathfrak{F}} - \bar{\mu}^{\mathfrak{F}} : g = 1, \dots, G\},$$

let $P_{\perp, \mathfrak{F}}$ denote the orthogonal projector onto $\mathcal{N}_{\mathfrak{F}}^{\perp}$, and define

$$R := P_{\perp, \mathfrak{F}} \Psi_{\mathfrak{F}}(Z).$$

Write

$$\ell_{\mathfrak{F}}(r, r') := \langle r, r' \rangle_{\mathcal{H}_{\mathfrak{F}}}$$

for the linear kernel on the cleaned prototype space. We write

$$R \perp_{\mathfrak{F}} S$$

when

$$\text{MMD}_{\ell_{\mathfrak{F}}}(P_{R|S=g}, P_{R|S=h}) = 0 \quad \text{for every } g, h \in \{1, \dots, G\},$$

that is, independence at the level encoded by the prototype family $\mathfrak{F}$.

Theorem C.5 (KINDER kernel $\mathfrak{F}$ -independence theorem). *In the setting of Definitions C.1 and C.4,*

$$R \perp_{\mathfrak{F}} S.$$

Equivalently,

$$\text{MMD}_{\ell_{\mathfrak{F}}}(P_{R|S=g}, P_{R|S=h}) = 0 \quad \text{for every } g, h \in \{1, \dots, G\}.$$

Proof. For each $g \in \{1, \dots, G\}$,

$$\mathbb{E}[R \mid S = g] = \mathbb{E}[P_{\perp, \mathfrak{F}} \Psi_{\mathfrak{F}}(Z) \mid S = g] = P_{\perp, \mathfrak{F}} \mu_g^{\mathfrak{F}}.$$

Since $\mu_g^{\mathfrak{F}} - \bar{\mu}^{\mathfrak{F}} \in \mathcal{N}_{\mathfrak{F}}$, the projector $P_{\perp, \mathfrak{F}}$ annihilates this difference, so

$$P_{\perp, \mathfrak{F}} \mu_g^{\mathfrak{F}} = P_{\perp, \mathfrak{F}} \bar{\mu}^{\mathfrak{F}} \quad \text{for every } g \in \{1, \dots, G\}.$$

Hence

$$\mathbb{E}[R \mid S = g] = \mathbb{E}[R \mid S = h] \quad \text{for every } g, h \in \{1, \dots, G\}.$$

For the linear kernel $\ell_{\mathfrak{F}}$ on $\mathcal{H}_{\mathfrak{F}}$,

$$\text{MMD}_{\ell_{\mathfrak{F}}}^2(P_{R|S=g}, P_{R|S=h}) = \|\mathbb{E}[R \mid S = g] - \mathbb{E}[R \mid S = h]\|_{\mathcal{H}_{\mathfrak{F}}}^2.$$

The right-hand side is therefore zero for every pair g, h , which gives

$$\text{MMD}_{\ell_{\mathfrak{F}}}(P_{R|S=g}, P_{R|S=h}) = 0 \quad \text{for every } g, h \in \{1, \dots, G\}.$$

Thus $R \perp_{\mathfrak{F}} S$. □

Theorem C.5 is the prototype-level statement for KINDER. It says that the cleaned group-conditional laws become indistinguishable by the kernel generated by the chosen prototype family. Full probabilistic independence requires one further identification condition on the cleaned model class.

Corollary C.6 (Exact independence under law-identifying prototypes). *In the setting of Theorem C.5, assume that the prototype functionals in $\mathfrak{F}$ are rich enough to identify the relevant group laws in the cleaned representation space. More precisely, let $\mathcal{P}_{\mathfrak{F}}$ be a class of probability laws on $\mathcal{H}_{\mathfrak{F}}$ such that $P_{R|S=g} \in \mathcal{P}_{\mathfrak{F}}$ for every $g \in \{1, \dots, G\}$, and assume that for any $P, Q \in \mathcal{P}_{\mathfrak{F}}$, equality of all prototype expectations implies equality of laws:*

$$\mathbb{E}_{R \sim P}[\psi(R)] = \mathbb{E}_{R \sim Q}[\psi(R)] \quad \text{for every } \psi \in \mathfrak{F} \implies P = Q.$$

Then the cleaned representation is independent of the protected attribute:

$$R \perp S.$$

Proof. By Theorem C.5, the cleaned representation has zero prototype-space discrepancy across all protected groups:

$$\text{MMD}_{\ell_{\mathfrak{F}}}(P_{R|S=g}, P_{R|S=h}) = 0 \quad \text{for all } g, h \in \{1, \dots, G\}.$$

Zero MMD under the induced prototype-space linear kernel $\ell_{\mathfrak{F}}$ means that the corresponding prototype mean embeddings are equal. Therefore, for every prototype functional $\psi \in \mathfrak{F}$,

$$\mathbb{E}_{R \sim P_{R|S=g}}[\psi(R)] = \mathbb{E}_{R \sim P_{R|S=h}}[\psi(R)] \quad \text{for all } g, h \in \{1, \dots, G\}.$$

By the law-identifying assumption on $\mathfrak{F}$, these equal prototype expectations imply equality of the actual group-conditional laws:

$$P_{R|S=g} = P_{R|S=h} \quad \text{for all } g, h \in \{1, \dots, G\}.$$

Since S is discrete, equality of all group-conditional laws implies that the conditional distribution of R does not depend on S . Hence,

$$R \perp S.$$

□

D More on Experimental Settings

D.1 Datasets

Bias for Bios. We use BIOS for occupation classification from free-text biographies. BIOS was introduced to study bias in occupation prediction from semantic representations, and is particularly well suited for concept-erasure evaluation because the target task is strongly correlated with demographic cues such as gender. [15]

Moji. We use Moji for binary sentiment classification from tweets. The dataset is designed to study social bias in language representations by comparing sentiment prediction across dialect groups. The protected attribute is dialect, distinguishing African-American English from Standard American English. This setting is useful for evaluating whether a model’s sentiment predictions rely on dialectal cues rather than task-relevant semantic content. [4]

Harvard-FairVLMed. We use Harvard-FairVLMed for multimodal vision-language learning. The task is binary glaucoma classification from paired Scanning Laser Ophthalmoscopy fundus images and de-identified ophthalmology clinical notes. The dataset contains 10,000 image–text pairs with official training, validation, and test splits. For fairness evaluation, we use race, gender, age group, and ethnicity as protected attributes. [38]

MIMIC-IV. We use MIMIC-IV for both classification and regression experiments. MIMIC-IV is a large deidentified electronic health record database collected from Beth Israel Deaconess Medical Center and contains rich clinical information including demographics, diagnoses, procedures, laboratory measurements, treatments, and notes. Its breadth makes it a useful benchmark for studying fairness in realistic clinical prediction settings. [28]

FairSeg. We evaluate segmentation on FairSeg, a medical-image fairness benchmark built from 10,000 fundus images with pixel-level disc and cup masks and associated demographic attributes. FairSeg was introduced to support fairness research in medical segmentation, where both predictive quality and group-level equity are important. [62]

CelebA. We use CelebA for unimodal face-attribute prediction. The task is binary smile classification from face images. Images and attribute labels are taken from the original CelebA dataset, using its standard train/validation/test partition. For fairness evaluation, we use age group and gender as protected attributes.[35]

Communities and Crime. We use Communities and Crime for tabular regression. This dataset combines community-level socioeconomic variables, law-enforcement statistics, and crime outcomes across U.S. communities, making it a standard benchmark for studying regression under sensitive demographic structure. [41]

Dataset split imbalance. Table D.8 summarizes the imbalance structure of the evaluation splits under the same protected attributes used in the reported fairness metrics. The table separates marginal protected-group imbalance from the more difficult label-conditioned or target-conditioned subgroup imbalance. This distinction is important because several datasets that appear moderate under a single protected attribute become substantially more skewed after conditioning on the task label or on protected-attribute intersections. The table should be read as a characterization of the statistical stress imposed by the benchmark: classification settings test whether errors and true-positive behavior remain stable across label-conditioned protected groups; segmentation settings test whether spatial prediction quality remains stable across protected groups and their intersections; and regression settings test whether prediction error remains stable across groups with different target distributions.

These split properties are closely related to the design of KINDER. Because KINDER estimates protected-group structure through group prototypes in the representation space, it is intended to remove demographic variation without tying the erasure operation to a particular prediction head or task loss. The imbalance summary therefore helps explain why the experiments emphasize not only marginal protected attributes, but also multi-attribute and intersectional settings. In these regimes, fairness failures can remain hidden when only aggregate performance or single-attribute summaries are considered. The empirical results are broadly consistent with this motivation, especially in the text, MIMIC-IV classification, FairSeg, and FairVLMed settings, where improvements are observed on fairness metrics that directly depend on class-conditioned, multi-attribute, or intersectional subgroup behavior. At the same time, the regression results indicate that severe intersectional sparsity remains a difficult case.

Dataset and asset licenses. All datasets and baseline assets are used according to their original research-use terms. MIMIC-IV is used under the PhysioNet credentialed-access data use agreement. CelebA is used under its non-commercial research-use terms. BIOS, Moji, FairSeg, Harvard-FairVLMed, and Communities and Crime are public research datasets used according to their original release terms. We do not redistribute any dataset.

D.2 Implementation Details

Software and optimization. Unless specified otherwise by a baseline paper, all models were implemented in PyTorch. Models were optimized with Adam [32] or AdamW [36], matching the optimizer used by the corresponding baseline trainer when applicable. Hyperparameters were selected according to validation task utility, using AUC for classification, MAE for regression, and Dice score for segmentation.

For all experiments, the main hyperparameter tuned for KINDER was the Gamma parameter, searched over $\Gamma \in \{0.1, 0.5, 1.0, 1.5, 2.0, 10, 15, 50\}$. We also performed a smaller grid search over the random Fourier feature dimension, $m_{\text{RFF}} \in \{128, 256, 512, 1024\}$. All remaining hyperparameters were kept identical to those of the corresponding baseline model.

Compute resources. All experiments were run on a cluster using a single NVIDIA A100 80GB GPU per run. Individual training runs usually took up to 3 GPU-hours depending on the task, dataset,

Table D.8: Dataset split imbalance summary. R denotes the max/min ratio over nonzero groups or cells. For classification datasets, the final column reports the smallest nonzero label-by-intersectional protected cell. For segmentation, it reports the smallest intersectional protected group. For regression, it reports the largest target-mean shift across protected groups or intersections. Non-specific protected labels such as unknown/other are excluded.

Dataset	Task / target	N	Outcome	Smallest group per attribute	Worst split imbalance
FairSeg	Segmentation disc/cup mask	9,432	mask target	race: Asian, 9.2%, $R=8.3\times$ gender: male, 41.5%, $R=1.4\times$ ethnicity: Hispanic, 3.9%, $R=24.4\times$	Asian $\times$ male $\times$ Hispanic $< 0.1\%$, $R=4051.0\times$
FairVLMed	Binary classification glaucoma	9,460	positive 50.6%	race: Asian, 8.3%, $R=9.3\times$ gender: male, 43.6%, $R=1.3\times$ ethnicity: Hispanic, 4.2%, $R=22.8\times$	positive $\times$ Asian $\times$ male $\times$ Hispanic $< 0.1\%$, $R=2089.0\times$
MIMIC-IV ICU mortal- ity	Binary classification hospital mortality	72,114	positive 10.1%	gender: female, 45.0%, $R=1.2\times$ race: NHPI, 0.2%, $R=483.3\times$ age: ≤ 50 , 16.4%, $R=5.1\times$	positive $\times$ female $\times$ AIAN $\times \leq 50$ $< 0.1\%$, $R=24464.0\times$
MIMIC-IV ICU LOS	Regression ICU length of stay	56,527	median 2.36 IQR 1.55–4.29	gender: female, 44.2%, $R=1.3\times$ race: NHPI, 0.2%, $R=524.8\times$ age: 18–39, 7.9%, $R=4.5\times$	intersectional target-mean shift 5.8
Bias for Bios	Multiclass classification profession	396,189	28 classes class $R=84.2\times$	gender: group 1, 46.1%, $R=1.2\times$	profession 23 $\times$ group 1 $< 0.1\%$, $R=476.6\times$
Moji 90%	Binary classification sentiment	100,000	positive 50.0%	dialect: protected=0, 50.0%, $R=1.0\times$	negative $\times$ protected=1 5.0%, $R=9.0\times$
Communities and Crime	Regression violent crime rate	1,994	median 0.15 IQR 0.07–0.33	race: Asian, 4.4%, $R=17.9\times$	race target-mean shift 0.392

and backbone. Hyperparameter grid searches required approximately 1–48 GPU-hours depending on the size of the search space and the task.

Self-supervised learning experiments. For SSL methods, we swept a 3×3 grid over learning rates $\{10^{-5}, 10^{-4}, 10^{-3}\}$ and weight-decay coefficients $\{10^{-6.5}, 10^{-5}, 10^{-3.5}\}$. These values correspond to three log-spaced points spanning the same decade ranges as in prior work [8]. Objective-specific architectural hyperparameters were fixed to their default values for each loss and architecture. Within this grid, SSL models were trained for 10 epochs. We used a batch size of 512 for multimodal runs and 1024 for unimodal runs.

Text classification experiments. For text classification, we used a batch size of 512, a learning rate of 2×10^{-5} , a weight decay of 10^{-2} , and early stopping patience of 3. These hyperparameters were kept fixed across the corresponding baseline and KINDER experiments.

Regression experiments. For regression tasks, we used a batch size of 1024, an evaluation batch size of 4096, and a maximum training length of 50 epochs. Model selection was performed using validation MAE, and the same optimization and model-selection protocol was used for both the baseline and KINDER experiments.

Segmentation experiments. For segmentation tasks, we used a batch size of 64, a learning rate of 10^{-3} , a weight decay of 10^{-4} , and early stopping patience of 3. These hyperparameters were kept fixed across all segmentation experiments. Importantly, prior concept-erasure methods were not directly compatible with segmentation tasks, since they typically require removing or retraining the higher-level learned weights after the layer to which the erasure method is applied. To adapt these methods to segmentation, we used the following protocol: first, we trained the baseline model; second, we applied the concept-erasure method to the encoder; and third, we retrained the decoder on the cleaned representations while keeping the encoder frozen.

Representation learning backbones. For Harvard-FairVLMed multimodal experiments, we report CLIP and FactorCL backbones from the FairVLMed/FairCLIP setting [38, 47]. For MIMIC-IV tabular *classification*, the backbone is TabM [21]. For tabular *regression*, we report ResNet [25] and periodic-embedding MLPs (MLP-PLR) [22] on MIMIC-IV, and MLP-PLR on Communities and Crime. For Bias in Bios and Moji, we use the BERT Transformer encoder from HuggingFace

[65]. For FairSeg, we use U-Net [53] following the reported FairSeg benchmark configuration [62]. Protected attributes are used only in fairness evaluation and in the prototype-indexed KINDER variants.

D.3 Compared Method Details

We compare KINDER-augmented training against standard post-training *concept erasure* and projection baselines applied to frozen features before probing. The compared baselines are INLP [49], RLACE [50], KCE [51], LEACE [3], and SUP [57]. For each method, we follow the implementation and optimization details reported in the corresponding paper whenever applicable, and tune only the main erasure-strength hyperparameter used by that method: the projection count for INLP, the removed dimension for RLACE, the kernel family and kernel bandwidth γ for KCE, the sufficient-projection dimension q for SUP, and the removed rank for LEACE. All baseline hyperparameters are selected by validation task utility rather than fairness or erasure scores, using the task-specific validation criteria described in Appendix D.2. This matches our research question on how much protected-attribute information can be removed while preserving strong downstream performance.

D.4 Definitions of Metrics Used

Let $\mathcal{D} = \{(x_i, y_i, \hat{y}_i, s_i)\}_{i=1}^n$ denote the test set, where y_i is the ground-truth target, $\hat{y}_i$ is the model prediction, and $s_i \in \mathcal{G}$ denotes the protected-group membership. The group variable s_i may denote either a single protected attribute or an intersectional group obtained by taking the Cartesian product of multiple protected attributes. Thus, all group-wise metrics below are computed in the same way for both marginal and intersectional fairness evaluations.

Let Y denote the ground-truth target, $\hat{Y}$ the model prediction, and $S \in \mathcal{G}$ the protected-group variable. The group variable S can represent either a single protected attribute or an intersectional attribute formed from multiple protected attributes; all group-wise metrics are computed with respect to the resulting group set $\mathcal{G}$.

Classification. For non-text classification tasks, we follow prior work [10] and report an aggregate fairness metrics based on statistical parity, equal opportunity, equalized odds, and equal accuracy. Let

$$\text{SR}(g) = \mathbb{P}(\hat{Y} = 1 \mid S = g), \quad (\text{D.12})$$

$$\text{TPR}(g) = \mathbb{P}(\hat{Y} = 1 \mid Y = 1, S = g), \quad (\text{D.13})$$

$$\text{FPR}(g) = \mathbb{P}(\hat{Y} = 1 \mid Y = 0, S = g), \quad (\text{D.14})$$

$$\text{Acc}(g) = \mathbb{P}(\hat{Y} = Y \mid S = g). \quad (\text{D.15})$$

For a group-wise quantity $m(g)$, define its group disparity as

$$\Delta_{\mathcal{G}}(m) = \max_{g \in \mathcal{G}} m(g) - \min_{g \in \mathcal{G}} m(g). \quad (\text{D.16})$$

We then compute

$$\text{SP}_{\text{gap}} = \Delta_{\mathcal{G}}(\text{SR}), \quad (\text{D.17})$$

$$\text{EOpp}_{\text{gap}} = \Delta_{\mathcal{G}}(\text{TPR}), \quad (\text{D.18})$$

$$\text{EOdd}_{\text{gap}} = \max \{ \Delta_{\mathcal{G}}(\text{TPR}), \Delta_{\mathcal{G}}(\text{FPR}) \}, \quad (\text{D.19})$$

$$\text{EAcc}_{\text{gap}} = \Delta_{\mathcal{G}}(\text{Acc}). \quad (\text{D.20})$$

The aggregate fairness gap is

$$\text{FairGap} = \frac{1}{4} (\text{SP}_{\text{gap}} + \text{EOpp}_{\text{gap}} + \text{EOdd}_{\text{gap}} + \text{EAcc}_{\text{gap}}). \quad (\text{D.21})$$

Lower values indicate smaller group disparity. When reporting worst-group performance, we use

$$\text{WGA} = \min_{g \in \mathcal{G}} \text{Acc}(g), \quad (\text{D.22})$$

where higher values indicate better worst-group utility.

Text classification. For text classification benchmarks, we follow the text fairness literature and report worst-group macro-F1 [7], worst-class TPR gap [26, 52], and RMS class TPR gap [49, 50]. These metrics are used because text benchmarks such as occupation and sentiment classification are commonly evaluated through class-wise group disparities rather than only aggregate binary fairness gaps.

Let $\mathcal{C}$ be the label set. For each group g and class c , let $F1_{g,c}$ denote the F1 score restricted to samples from group g and class c . The group macro-F1 is

$$\text{MacroF1}(g) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} F1_{g,c}. \quad (\text{D.23})$$

Worst-group macro-F1 is

$$\text{WgF1} = \min_{g \in \mathcal{G}} \text{MacroF1}(g). \quad (\text{D.24})$$

Higher values indicate better worst-group text classification performance.

For class-wise group recall disparity, define

$$\text{TPR}_{g,c} = \mathbb{P}(\hat{Y} = c \mid Y = c, S = g). \quad (\text{D.25})$$

The TPR gap for class c is

$$\text{TPRGap}_c = \max_{g \in \mathcal{G}} \text{TPR}_{g,c} - \min_{g \in \mathcal{G}} \text{TPR}_{g,c}. \quad (\text{D.26})$$

The worst-class TPR gap is

$$\text{WCTG} = \max_{c \in \mathcal{C}} \text{TPRGap}_c. \quad (\text{D.27})$$

The RMS class TPR gap is

$$\text{RCTG} = \sqrt{\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} (\text{TPRGap}_c)^2}. \quad (\text{D.28})$$

Lower WCTG and RCTG indicate smaller class-wise group disparity.

Regression. For regression, we report mean absolute error (MAE) as the main utility metric and MAE gap for group error disparity [24]. The overall MAE is

$$\text{MAE} = \mathbb{E} [|Y - \hat{Y}|]. \quad (\text{D.29})$$

The group-specific MAE is

$$\text{MAE}(g) = \mathbb{E} [|Y - \hat{Y}| \mid S = g]. \quad (\text{D.30})$$

Worst-group MAE and MAE gap are defined as

$$\text{WG-MAE} = \max_{g \in \mathcal{G}} \text{MAE}(g), \quad (\text{D.31})$$

$$\text{MAE}_{\text{gap}} = \max_{g \in \mathcal{G}} \text{MAE}(g) - \min_{g \in \mathcal{G}} \text{MAE}(g). \quad (\text{D.32})$$

Lower values indicate smaller prediction error and smaller group error disparity. We also report R^2 as a regression utility metric:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}. \quad (\text{D.33})$$

Segmentation. For segmentation, we report Dice and HD95 as utility metrics. Dice measures region overlap between the predicted and ground-truth masks, while HD95 measures boundary error using the 95th-percentile Hausdorff distance. We additionally report worst-group Dice for overlap fairness [43] and HD95 gap for boundary fairness [6].

Let P denote the predicted segmentation mask and T the ground-truth mask. The Dice score is

$$\text{Dice}(P, T) = \frac{2|P \cap T|}{|P| + |T|}. \quad (\text{D.34})$$

Table E.9: **Feature-level KINDER order ablation on MOJI using BERT.** The RFF representation $\Phi(z)$ is used as the pre-projection reference. Higher prototype orders add residual nuisance directions beyond the lower-order subspace. Projection shift is $\|R - \Phi(z)\|_F / \|\Phi(z)\|_F$.

Representation	Projection Shift	RBF-HSIC(R, S) ↓	Mean Gap ↓	Cov. Gap ↓	3rd-Moment Gap ↓	Removed Rank
Raw encoder z	—	—	1.164	1.457	1.793	0
Raw RFF $\Phi(z)$	0.000	1.000	1.000	1.000	1.000	0
KINDER-order 1	0.155	0.975	0.998	0.980	0.931	1
KINDER-order 2	0.362	0.687	0.679	0.279	0.263	3
KINDER-order 3	0.540	0.433	0.563	0.124	0.065	4

For each group g , let $\text{Dice}(g)$ denote the average Dice score over samples with $S = g$. Worst-group Dice is

$$\text{WG-Dice} = \min_{g \in \mathcal{G}} \text{Dice}(g). \quad (\text{D.35})$$

Higher values indicate better worst-group overlap quality.

For HD95, let ∂P and ∂T denote the boundaries of the predicted and ground-truth masks. For point sets A and B , define the directed 95th-percentile boundary distance as

$$h_{95}(A, B) = Q_{0.95} \left(\left\{ \min_{b \in B} \|a - b\|_2 : a \in A \right\} \right), \quad (\text{D.36})$$

where $Q_{0.95}$ denotes the 95th percentile. The symmetric HD95 is

$$\text{HD95}(P, T) = \max \{h_{95}(\partial P, \partial T), h_{95}(\partial T, \partial P)\}. \quad (\text{D.37})$$

For each group g , let $\text{HD95}(g)$ denote the average HD95 over samples with $S = g$. The HD95 gap is

$$\text{HD95}_{\text{gap}} = \max_{g \in \mathcal{G}} \text{HD95}(g) - \min_{g \in \mathcal{G}} \text{HD95}(g). \quad (\text{D.38})$$

Lower HD95 and lower HD95 gap indicate better boundary quality and smaller boundary disparity.

E Additional Results

E.1 Prototype-order ablation

This ablation studies how the order of the prototype family affects feature-level sensitive-attribute removal on MOJI. The raw RFF representation is used as the reference, and higher-order KINDER variants add residual nuisance directions beyond the lower-order prototype subspace. As shown in Table E.9, first-order prototypes induce only a small projection shift but mostly remove mean-level variation. Second- and third-order prototypes remove substantially more dependence, reducing HSIC as well as covariance and third-moment gaps, at the cost of larger representation shifts and higher removed rank. This supports the use of prototype order as a controllable erasure-strength parameter.

E.2 Feature-level bootstrap significance tests

To assess whether the feature-level differences in Fig. 2(c–d) are significant, we performed paired bootstrap tests for KINDER vs the reported baselines on Moji dataset in Table E.10. For each bootstrap sample, we recompute the normalized RBF-HSIC and average moment-gap metrics. The improvement is defined as the baseline value minus the KINDER value, so positive values favor KINDER. We report one-sided bootstrap p -values for positive improvement and Benjamini–Hochberg corrected q -values across the tested comparisons. We observe that KINDER’s shown improvements in Fig. 2(c–d) are mostly significant, with the exception of KINDER vs KCE average gaps where despite the 95% CI being largely positive, it still holds zero inside it.

E.3 Sequential MIMIC-IV classification ablation

Table E.11 reports a sequential single-attribute removal ablation for MIMIC-IV classification. In the main experiment, Table 3, KINDER removes the joint protected-attribute structure directly. Here, we instead apply each erasure method repeatedly, removing one protected attribute at a time

Table E.10: Paired bootstrap tests for the two feature-level summary metrics used in Fig. 2(c–d). Lower values are better. Improvement is baseline minus KINDER, so positive values favor KINDER.

Baseline	Metric	Baseline value	KINDER value	Improvement	95% CI	p one-sided	q BH	Signif.
SUP	RBF-HSIC	0.89	0.46	0.43	[0.33, 0.70]	0.003	0.005	**
SUP	Average gap	0.96	0.27	0.69	[0.64, 0.80]	0.003	0.005	**
INLP	RBF-HSIC	0.56	0.46	0.09	[0.05, 0.46]	0.010	0.013	*
INLP	Average gap	0.89	0.27	0.62	[0.56, 0.75]	0.003	0.005	**
RLACE	RBF-HSIC	1.01	0.46	0.55	[0.42, 0.77]	0.003	0.005	**
RLACE	Average gap	0.96	0.27	0.69	[0.63, 0.79]	0.003	0.005	**
KCE	RBF-HSIC	1.10	0.46	0.64	[0.50, 0.85]	0.003	0.005	**
KCE	Average gap	0.30	0.27	0.03	[-0.03, 0.13]	0.113	0.123	

Table E.11: **Sequential post-hoc MIMIC-IV classification ablation** using ResNet. Higher is better for accuracy, F1, and worst-group accuracy; lower is better for gap measures. Best values in **bold** and second-best values are underlined.

Method	Acc $\uparrow$	F1 $\uparrow$	Gender Gap $\downarrow$	Race Gap $\downarrow$	Age Gap $\downarrow$	Inter. Gap $\downarrow$	WorstGroupAcc $\uparrow$
ResNet	0.92	0.40	0.02	0.23	0.03	0.47	<u>0.66</u>
+ INLP-seq	0.78	0.43	0.02	0.22	0.06	0.53	0.57
+ RLACE-seq	0.78	0.43	0.03	0.23	0.07	0.53	0.58
+ LEACE-seq	<u>0.81</u>	<u>0.44</u>	0.01	0.25	0.05	0.73	0.38
+ KCE-seq	0.80	<u>0.20</u>	0.02	<u>0.19</u>	<u>0.04</u>	0.76	0.00
+ KINDER-seq	0.79	0.43	0.02	0.16	0.08	<u>0.46</u>	0.50
+ KINDER-seq (NH)	0.92	0.46	0.01	0.31	0.03	0.44	0.69

in sequence, and then evaluate the final representation using the same marginal and intersectional fairness metrics. This tests whether methods originally designed around a single protected attribute remain stable when extended to multiple protected attributes through repeated erasure. The results show that repeated post-hoc removal can reduce some marginal gaps, but often introduces large utility loss or unstable intersectional behavior. In contrast, KINDER-seq preserves the baseline accuracy, gives the best F1 and worst-group accuracy, and obtains the lowest intersectional gap. The KINDER-seq vs KINDER-seq (NH) variants further shows the head-mismatch effect. Replacing the native task head with a linear probe improves some gap values but substantially reduces utility.

E.4 Communities and Crime Dataset

Table E.12 reports the Communities and Crime regression results using MLP-PLR. This dataset tests whether concept erasure remains useful in a low-dimensional tabular regression setting where protected-group structure is tied to community-level socioeconomic variables. KINDER achieves the best overall utility, with the lowest MAE and highest R^2 , and also gives the best race worst-group error. SUP and RLACE obtain slightly lower race-gap values, but KINDER remains close while preserving stronger predictive performance. This result is consistent with the main regression findings. Aggressive post-hoc removal can reduce some fairness gaps, but often loses substantial task information, while KINDER gives a stronger utility-preserving tradeoff.

F Limitations

KINDER assumes access to protected-attribute labels during training or post-hoc prototype estimation, which is standard in fairness evaluation but may not hold in all deployment settings. Its effectiveness also depends on the chosen prototype functionals and RFF approximation. Richer prototypes can capture more group-conditioned structure, but may increase computation and remove more task-relevant variation if the protected attributes are strongly entangled with the target. In addition, while our theory gives prototype-space independence, the full probabilistic independence requires the much stronger condition that the prototype family is defined in a way that identifies the relevant group laws. Future work should explore additional prototype families, settings with missing or noisy sensitive attributes, and larger-scale evaluations under severe intersectional sparsity.

Table E.12: **Communities and Crime regression results using MLP-PLR**. Lower is better for MAE, worst-group MAE, and MAE gap; higher is better for R^2 . Best values in **bold** and second-best values are underlined.

Method	MAE ↓	R^2 ↑	Race WG ↓	Race Gap ↓
MLP-PLR	<u>0.093</u>	<u>63.8</u>	0.154	0.090
+ INLP	0.148	18.2	0.288	0.164
+ LEACE	0.157	11.7	0.359	0.225
+ RLACE	0.094	63.8	0.152	<u>0.071</u>
+ KCE	0.096	60.2	0.190	0.113
+ SUP	0.094	63.8	<u>0.149</u>	0.070
+ KINDER	0.092	65.3	0.145	0.079

G Impact Statement

Our work explicitly recognises that ML models can cause negative societal impacts if not addressed. Thus, we present a method, KINDER to promote fairer machine learning (ML) methods that can be fully generalizable across a range of tasks, data types and training settings. By developing KINDER, we aim to reduce the ML model performance gaps across different demographic groups at both the group and intersectional level, thus generating a positive societal impact. We further acknowledge that the sensitive attributes may not capture all dimensions of identities and future work should take these into consideration.

NeurIPS Paper Checklist

1. Claims

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

Answer: [\[Yes\]](#)

Justification: The abstract and introduction stated four contributions. Our first key contribution is outlined in Section 4.3, our second contribution is outlined in Section 4.2, 4.3. The claim that KINDER is generalizable and works across diverse settings, tasks and modalities in Section 6.

Guidelines:

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

2. Limitations

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

Answer: [\[Yes\]](#)

Justification: Discussion and Conclusion Section (Section 7) has a dedicated coverage of the limitations of our work which we further expand on in Appendix F.

Guidelines:

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

3. Theory assumptions and proofs

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

Answer: [Yes]

Justification: The theoretical discussions in the paper list the assumptions made.

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [Yes]

Justification: All experimental details are included. Furthermore, an anonymized zip file with reproduction codes is provided in the supplementary material.

Guidelines:

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

5. Open access to data and code

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

Answer: [Yes]

Justification: All datasets used are publicly available. The code that can be used to reproduce our results are shared inside an anonymized zip provided in the supplementary material.

Guidelines:

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

6. Experimental setting/details

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

Answer: [Yes]

Justification: All the training setup are included inside the appendix.

Guidelines:

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

7. Experiment statistical significance

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

Answer: [Yes]

Justification: Significant test results are included inside the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

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

Answer: [Yes]

Justification: Details are included inside the appendix.

Guidelines:

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

9. Code of ethics

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

Answer: [Yes]

Justification: The research conducted in this paper conforms with the NeurIPS Code of Ethics.

Guidelines:

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

10. Broader impacts

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

Answer: [Yes]

Justification: A dedicated impact section has been included under Section G.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

Answer: [N/A]

Justification: We do not release data or generative models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

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

Answer: [Yes]

Justification: All datasets used are publicly available for research purposes. All datasets and baseline models have been cited properly.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [Yes]

Justification: We are releasing a detailed Readme file accompanying the code and trained models released with our paper.

Guidelines:

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

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

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

Answer: [N/A]

Justification: Not applicable.

Guidelines:

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

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

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

Answer: [N/A]

Justification: Not Applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: We have used an LLM only for paraphrasing text to improve the overall clarity.

Guidelines:

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