TOWARDS EFFICIENT FAIRNESS IMAGE RETRIEVAL WITH DISENTANGLED INFORMATION SUPPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep hashing has emerged as an effective method for large-scale image retrieval, improving computational efficiency by converting high-dimensional data into compact binary codes. Despite its success, recent studies reveal that deep hashing methods may exhibit fairness issues, leading to biased or discriminatory retrieval results across demographic groups. To jointly improve retrieval accuracy and group fairness, we introduce **D**isentangled **I**nformation **S**uppressed **H**ashing (**DISH**), a framework that learns fair and discriminative representations. DISH employs a disentangled encoder to decompose each image into factor-specific representations. To encourage semantic concentration and interpretability, a disentangled consistency objective is introduced to enforce factor-level stability under augmentation and align semantic evidence with latent factors. Furthermore, an information suppression module is designed to mitigate sensitive information leakage through probability-driven channel masking, channel-wise adversarial learning, and conditional covariance regularization. These components work collaboratively to eliminate sensitive signals both within and between feature channels while preserving semantic discriminability. Extensive experiments on multiple benchmarks show that DISH substantially outperforms state-of-the-art deep hashing baselines in retrieval accuracy while achieving better fairness.

1 INTRODUCTION

Deep hashing has emerged as an effective approach for large-scale image retrieval tasks. By leveraging deep neural networks, deep hashing methods transform image data into compact binary codes, substantially reducing storage requirements and enabling rapid retrieval through efficient binary operations(Slaney & Casey, 2008; Gong et al., 2012; Liu et al., 2012; Hoe et al., 2021). Compared to traditional retrieval methods, hashing-based approaches offer remarkable advantages in retrieval speed and scalability(Yuan et al., 2020; Wang et al., 2023), making them especially valuable for applications such as search engines(Wang et al., 2012), recommendation systems(Luo et al., 2024).

Despite their success in retrieval accuracy, recent studies have revealed that deep hashing models, similar to many other representation learning systems, may inherit and even amplify societal biases present in training data, leading to systematically unfair retrieval results across demographic groups defined by sensitive attributes such as age, gender or ethnicity(Zhang et al., 2024). Although fairness has been extensively explored in general classification tasks(Berk et al., 2017; Hardt et al., 2016; Nabi & Shpitser, 2018; Zafar et al., 2017; Sattigeri et al., 2019), it remains under-explored in the context of deep hashing: most existing methods focus solely on maximizing retrieval accuracy(Li et al., 2015; 2017; Su et al., 2018; Li et al., 2019). Even when fairness is considered, interventions are often applied post-hoc or confined to the final hash space, where the extreme compression and discreteness of binary codes severely limit the capacity to disentangle and suppress sensitive signals without sacrificing semantic utility(Zhang et al., 2024).

To tackle these issues, we propose **D**isentangled **I**nformation **S**uppressed **H**ashing (**DISH**), a framework that learns fair and discriminative hash representations by intervening in the continuous feature space prior to binarization. DISH employs a disentangled encoder to decompose each image into factor-specific representations, these representations are regularized by a disentangled consistency objective, which promotes semantic concentration and factor-level stability under augmentations. Built upon this disentangled structure, DISH suppresses sensitive cues through a combination of

channel-wise adversarial learning and conditional covariance regularization, theoretically minimizing their recoverability both within and across feature dimensions. A final semantic alignment loss in Hamming space ensures the binarized codes retain strong discriminative power, striking an effective balance between fairness and performance.

In summary, our key contributions are as follows: (1) We propose DISH, a fairness-aware hashing framework that disentangles semantic and sensitive factors in the continuous feature space prior to binarization, enabling targeted suppression of bias while preserving retrieval semantics; (2) We introduce a theoretically grounded information-suppression mechanism that combines channel-wise adversarial learning with conditional covariance regularization, minimizing sensitive leakage both within and across latent factors with formal guarantees via mutual information bounds; (3) We conduct comprehensive evaluations on multiple benchmarks, showing that DISH establishes new state-of-the-art results in balancing retrieval performance and fairness.

2 Related work

Learning to Hash. Hashing has been widely adopted for large-scale image retrieval due to its computational and storage efficiency. Early work spans data-independent LSH (Slaney & Casey, 2008) and data-dependent schemes such as ITQ and supervised hashing (Gong et al., 2012; Liu et al., 2012; Shen et al., 2015). With the rise of deep learning, deep hashing methods have achieved substantial gains in retrieval accuracy. Deep hashing methods are broadly categorized by their use of data: pairwise methods,triplet-based methods, and pointwise methods. Pairwise methods that optimize hash codes to preserve pairwise similarity relationships (Wang et al., 2010; Li et al., 2015); Triplet-based methods that enforce relative ranking constraints among anchor-positive-negative triplets (Wang et al., 2017); and Pointwise methods that directly supervise hash codes using class labels or semantic prototypes (Su et al., 2018; Yuan et al., 2020; Hoe et al., 2021; Wang et al., 2023). These approaches have achieved remarkable gains in retrieval accuracy and efficiency, but they neglect the societal implications of biased or discriminatory outcomes. Recently, FATE (Zhang et al., 2024) made the first attempt to incorporate fairness into hashing. However, due to the extreme compression and discreteness of binary codes, interventions in Hamming space inherently limit the ability to disentangle and suppress sensitive signals without compromising semantic utility.

Disentangled Representation Learning. Disentangled representation learning(DRL) aims to encode underlying factors of variation into separate and interpretable dimensions of the latent space. Existing DRL methods can be broadly categorized into dimension-wise and vector-wise approaches based on their granularity of semantic alignment. Dimension-wise methods typically map one semantic factor to one latent dimension, e.g., β-VAE (Higgins et al., 2017), FactorVAE (Kim & Mnih, 2018), β-TCVAE (Chen et al., 2018), and GAN-based InfoGAN (Chen et al., 2016). By contrast, vector-wise methods represent a factor with a low-dimensional subspace, e.g., DR-GAN (Tran et al., 2017), DRNET (Denton et al., 2017), and MAP-IVR (Liu et al., 2021). Recent advancements integrate contrastive learning with disentanglement paradigms to enhance representation quality without relying heavily on labeled data (Li et al., 2021; Wang et al., 2024). In the context of fairness, disentanglement has been explored to isolate sensitive attributes from task-relevant features (Zhu et al., 2024; Zhang et al., 2025), yielding improved fairness–utility trade-offs.

Fairness in Machine Learning. Fairness in machine learning has been extensively studied, particularly in classification and recommendation systems (Agarwal & Deshpande, 2022; Padh et al., 2021; Li et al., 2023). Common fairness notions include demographic parity, equalized odds and individual fairness(Caton & Haas, 2024). Techniques to mitigate bias can be broadly categorized into pre-processing (e.g., data reweighting or transformation (Krasanakis et al., 2018; Gronowski et al., 2023)), in-processing (e.g., adversarial debiasing (Celis & Keswani, 2019; Xu et al., 2019) or fairness constraints (Donini et al., 2018)), and post-processing methods (e.g., calibration of outputs (Hébert-Johnson et al., 2018)). Despite this progress, fairness in retrieval systems—particularly hashing-based methods—remains under-explored. Recent work by Zhang et al. (2024) represents a notable step toward fair hashing by incorporating adversarial learning and contrastive objectives directly in the hash space. However, operating solely in the compressed and discrete hash space limits the flexibility and effectiveness of bias mitigation.

3 PROBLEM DEFINITION

Let $\mathcal{D} = \{I_i, y_i, s_i\}_{i=1}^{N}$ denote a dataset, where I_i represents an input image, $y_i \in \mathcal{Y}$ is the corresponding target attribute label, and $s_i \in \mathcal{S}$ denotes the associated sensitive attribute label. The objective is to learn a deep hashing function $\mathcal{H} : I \mapsto \boldsymbol{b} \in \{-1, 1\}^L$, which maps images to compact binary hash codes of length L. This function should satisfy two properties: **(1) High Retrieval Accuracy**: relevant images (i.e., those sharing label y_i) should be ranked above those irrelevant ones; **(2) Group Fairness**: retrieval outcomes should be equitable across sensitive groups $\mathcal{S}$.

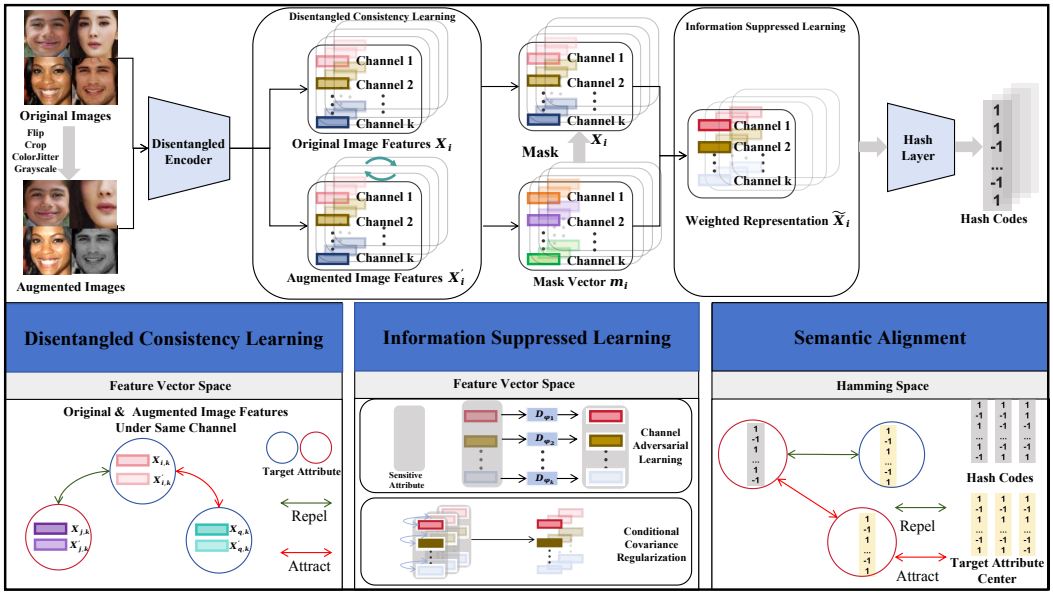

Figure 1: Overview of DISH. Here, $\mathbf{x}_{i,k}$ and $\mathbf{x'}_{i,k}$ are the k-th factor features for I_i and I_i'; $\mathbf{m}_i = [p_\theta(k \mid \mathbf{x}_i)]_{k=1}^{K}$ is the assignment mask and $\tilde{\mathbf{x}}_{i,k} = m_{i,k}\mathbf{x}_{i,k}$ the weighted feature; D_{ψ_k} (with GRL) is the channel discriminator; DISH learns fair and discriminative hash codes through disentangled consistency learning, information suppressed learning, and semantic alignment.

4 THE PROPOSED FRAMEWORK

This section presents the overall architecture of the proposed DISH framework, which comprises four primary components: (1) *Disentangled Encoder*, (2) *Disentangled Consistency Learning*, (3) *Information-Suppressed Learning*, and (4) *Semantic Alignment*. The first three components operate in the continuous representation space, whereas the fourth operates in the discrete Hamming space. A schematic of the full architecture is shown in Figure 1.

4.1 DISENTANGLED ENCODER

The Disentangled Encoder is designed to decompose each input image into K factor-specific representations, enabling the inference of underlying latent factors that contribute to the image content. The encoder comprises a pre-trained feature extractor (e.g., ResNet-50 (He et al., 2016)) to obtain high-level semantic features, followed by K parallel multilayer perceptron (MLP) branches. Each branch is responsible for modeling a distinct and approximately independent factor of variation. Given an input image I_i, the encoder produces a disentangled feature representation: $\mathbf{x}_i = [\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,K}] \in \mathbb{R}^d$, where $\mathbf{x}_{i,k} \in \mathbb{R}^{d/K}$ $(1 \le k \le K)$ corresponds to the k-th factor-specific component, and d denotes the total feature dimension. Additionally, we apply data augmentation to I_i (e.g., random cropping, flipping, and color jittering) to obtain an augmented view I_i'. The encoder processes I_i' through the same network, yielding: $\mathbf{x}_i' = [\mathbf{x}_{i,1}', \mathbf{x}_{i,2}', \ldots, \mathbf{x}_{i,K}'] \in \mathbb{R}^d$, which serves as a counterpart for disentangled consistency learning.

4.2 DISENTANGLED CONSISTENCY LEARNING

We define "consistency" at the factor level: the model's responses to the same image and its augmented view should remain stable; samples from the same class should cluster while different classes separate; and the data-driven factor assignment should align with label-based semantic evidence. To achieve this, we perform supervised contrastive learning within each factor channel, using labels to shape the similarity structure so that semantic information concentrates in a small number of interpretable factors rather than being mixed in a single representation space. The prototype-based assignments are then coupled with the contrastive signal, encouraging stable posterior preferences. Firstly, given $\mathbf{x}_i$, we compute the factor assignment probability $p_\theta(k \mid \mathbf{x}_i)$ using a prototype-based method. Specifically, we introduce K latent factor prototypes $\{\mathbf{c}_k\}_{k=1}^K$. The probability that the k^{th} latent factor is reflected in representation $\mathbf{x}_i$ is parameterized as:

$$p_\theta(k \mid \mathbf{x}_i) = \frac{\exp\big(\phi(\mathbf{x}_{i,k}, \mathbf{c}_k)\big)}{\sum_{k'=1}^K \exp\big(\phi(\mathbf{x}_{i,k'}, \mathbf{c}_{k'})\big)}. \tag{1}$$

where ϕ is the temperature-scaled cosine similarity and $\tau > 0$ is the temperature:

$$\phi(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2 \tau}. \tag{2}$$

Then, we define the supervised contrastive learning task under k-th latent factor. Given a minibatch $\mathcal{B}$, define $A(i) = \mathcal{B} \setminus \{i\}$ and $P(i) = \{p \in A(i) : y_p = y_i\}$. The contrastive softmax likelihood is:

$$p_\theta(y_i \mid \mathbf{x}_i, k) = \frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{\exp\big(\phi(\mathbf{x}_{i,k}, \mathbf{x}'_{p,k})\big)}{\sum_{a \in A(i)} \exp\big(\phi(\mathbf{x}_{i,k}, \mathbf{x}'_{a,k})\big)}. \tag{3}$$

We model the label evidence over latent factors as:

$$p_\theta(y_i \mid \mathbf{x}_i) = \sum_{k=1}^K p_\theta(k \mid \mathbf{x}_i)\, p_\theta(y_i \mid \mathbf{x}_i, k). \tag{4}$$

However, direct optimization is intractable due to the latent factors. Therefore, we instead optimize the evidence lower bound (ELBO) of the log-likelihood. For any distribution $q(k \mid \mathbf{x}_i, y_i)$:

$$\log p_\theta(y_i \mid \mathbf{x}_i) \;\geq\; \mathbb{E}_{k \sim q}[\log p_\theta(y_i \mid \mathbf{x}_i, k)] - D_{\mathrm{KL}}(q(\cdot \mid \mathbf{x}_i, y_i) \,\|\, p_\theta(\cdot \mid \mathbf{x}_i)). \tag{5}$$

Equality is attained by the variational posterior:

$$q_\theta(k \mid \mathbf{x}_i, y_i) = \frac{p_\theta(k \mid \mathbf{x}_i)\, p_\theta(y_i \mid \mathbf{x}_i, k)}{\sum_{k'=1}^K p_\theta(k' \mid \mathbf{x}_i)\, p_\theta(y_i \mid \mathbf{x}_i, k')}. \tag{6}$$

Apply Jensen's inequality to $\log \mathbb{E}_q[\cdot]$ to obtain equation 5. Maximization w.r.t. q yields equation 6.

Jensen Bound For Contrastive Likelihood. Let $s_{i,k}(a) = \exp\{\phi(\mathbf{x}_{i,k}, \mathbf{x}'_{a,k})\}$ and $u_p = \frac{s_{i,k}(p)}{\sum_{a \in A(i)} s_{i,k}(a)} \in (0, 1)$. Then, by concavity of log, $\log \frac{1}{|P(i)|} \sum_{p \in P(i)} u_p \geq \frac{1}{|P(i)|} \sum_{p \in P(i)} \log u_p$:

$$\log p_\theta(y_i \mid \mathbf{x}_i, k) \;\geq\; \frac{1}{|P(i)|} \sum_{p \in P(i)} \Big(\log s_{i,k}(p) - \log \sum_{a \in A(i)} s_{i,k}(a) \Big). \tag{7}$$

Computable Minibatch Lower Bound. Define the per-factor term with a partition function

$$Z_{i,k} := \sum_{a \in A(i)} \exp\big(\phi(\mathbf{x}_{i,k}, \mathbf{x}'_{a,k})\big), \ell_{i,k} := \frac{1}{|P(i)|} \sum_{p \in P(i)} \Big(\phi(\mathbf{x}_{i,k}, \mathbf{x}'_{p,k}) - \log Z_{i,k}\Big). \tag{8}$$

Combining equation 5 and equation 7, for any q, we get the following inequality:

$$\log p_\theta(y_i \mid \mathbf{x}_i) \;\geq\; \sum_{k=1}^K q(k \mid \mathbf{x}_i, y_i)\, \ell_{i,k} \;-\; D_{\mathrm{KL}}\big(q(\cdot \mid \mathbf{x}_i, y_i) \,\|\, p_\theta(\cdot \mid \mathbf{x}_i)\big). \tag{9}$$

Let q_θ be the variational posterior in equation 6. Define the batch lower bound (to maximize)

$$\mathcal{L}_{\mathrm{DCL}}^{\mathrm{lb}} = \sum_{i \in \mathcal{B}} \Big(\sum_{k=1}^K q_\theta(k \mid \mathbf{x}_i, y_i)\, \ell_{i,k} - D_{\mathrm{KL}}\big(q_\theta \,\|\, p_\theta\big) \Big). \tag{10}$$

In practice, we optimize the negative of the bound as the training loss:

$$\mathcal{L}_{\mathrm{DCL}} = -\,\mathcal{L}_{\mathrm{DCL}}^{\mathrm{lb}}. \tag{11}$$

4.3 INFORMATION-SUPPRESSED LEARNING

To promote invariance to the sensitive attribute without sacrificing semantic discriminability, we introduce an information-suppressed framework that combines probability-driven channel masking, channel-wise adversarial learning, and conditional covariance regularization to preserve task semantics while removing both per-channel and cross-channel sensitive cues.

Channel Masking. Building upon the factor assignment probabilities $p_\theta(k \mid \mathbf{x}_i)$ from equation 1, we design a probability-driven masking concentrates information into high-assignment factors and attenuates low-assignment ones, without requiring access to target labels or sensitive attributes. For each representation $\mathbf{x}_i$, we compute a channel-wise mask vector.

$$\mathbf{m}_i = \big[m_{i,1}, \ldots, m_{i,K}\big] = \big[p_\theta(1 \mid \mathbf{x}_i), \ldots, p_\theta(K \mid \mathbf{x}_i)\big]. \tag{12}$$

We then obtain an assignment weighted representation via channel-wise scaling:

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i \odot_c \mathbf{m}_i = \big[\mathbf{x}_{i,1}\, m_{i,1}, \ldots, \mathbf{x}_{i,K}\, m_{i,K}\big]. \tag{13}$$

where $\odot_c$ denotes multiplication by channel. This multiplication attenuates channels with lower $p_\theta(k \mid \mathbf{x}_i)$, concentrating semantics into more informative factors.

Channel Adversarial Learning. Let $s_i \in \{1, \ldots, C_s\}$ denote the sensitive attribute of I_i. From equation 13, write $\tilde{\mathbf{x}}_i = \big[\tilde{\mathbf{x}}_{i,1}, \ldots, \tilde{\mathbf{x}}_{i,K}\big]$ with $\tilde{\mathbf{x}}_{i,k} = m_{i,k}\, \mathbf{x}_{i,k} \in \mathbb{R}^{d/K}$ the k-th channel feature. For each channel, we instantiate a discriminator $D_{\psi_k} : \mathbb{R}^{d/K} \to \Delta^{C_s - 1}$ that predicts sensitive label from $\tilde{\mathbf{x}}_{i,k}$. We define adversarial loss by averaging the per-channel cross-entropy over minibatch:

$$\mathcal{L}_{\mathrm{CAL}}(\theta, \{\psi_k\}_{k=1}^K) = \frac{1}{|\mathcal{B}|\, K} \sum_{i \in \mathcal{B}} \sum_{k=1}^{K} \mathrm{CE}\big(s_i,\ D_{\psi_k}\big(\tilde{\mathbf{x}}_{i,k}(\theta)\big)\big)\,. \tag{14}$$

where CE denotes cross-entropy. Fairness is enforced via the following saddle-point objective:

$$\min_{\{\psi_k\}_{k=1}^K} \ \max_{\theta} \ \ \mathcal{L}_{\mathrm{CAL}}(\theta, \{\psi_k\}_{k=1}^K). \tag{15}$$

In practice, we implement the saddle-point objective in equation 15 via a gradient-reversal layer (GRL), which preserves the theoretical min–max view while yielding a simple training procedure. The GRL encourages each channel feature $\tilde{\mathbf{x}}_{i,k}$ to be uninformative about s_i. This realization is theoretically consistent with the per-channel min–max scheme and empirically stable in training.

Conditional Covariance Regularization. While the adversarial objective suppresses direct leakage of S from each channel, it does not prevent channels from jointly encoding complementary sensitive information. To further mitigate this effect, we introduce a conditional covariance penalty that encourages approximate conditional independence between channels given S. In practice, for a minibatch $\mathcal{B}$ with representations $\{\tilde{\mathbf{x}}_i\}_{i \in \mathcal{B}}$ and corresponding sensitive labels $\{s_i\}_{i \in \mathcal{B}}$, we first group samples by their sensitive class. For each class s, let $\mathcal{B}_s = \{i \in \mathcal{B} : s_i = s\}$ denote the subset of samples with label s. For each pair of channels (k, ℓ), we compute the empirical covariance:

$$\mathrm{Cov}\big(\tilde{\mathbf{X}}_k, \tilde{\mathbf{X}}_\ell \mid S = s\big) = \frac{1}{|\mathcal{B}_s| - 1} \sum_{i \in \mathcal{B}_s} \big(\tilde{\mathbf{x}}_{i,k} - \bar{\mathbf{x}}_{k,s}\big)\big(\tilde{\mathbf{x}}_{i,\ell} - \bar{\mathbf{x}}_{\ell,s}\big)^\top. \tag{16}$$

where $\bar{\mathbf{x}}_{k,s} = \frac{1}{|\mathcal{B}_s|} \sum_{i \in \mathcal{B}_s} \tilde{\mathbf{x}}_{i,k}$. The conditional covariance regularization loss is then defined as

$$\mathcal{L}_{\mathrm{CCR}}(\theta) = \sum_{s=1}^{C_s} \sum_{k \neq \ell} \Big\| \mathrm{Cov}\Big(\tilde{\mathbf{X}}_k, \tilde{\mathbf{X}}_\ell \mid S = s\Big) \Big\|_F^2. \tag{17}$$

This conditional covariance regularizer promotes conditional decorrelation between channels, reducing the possibility that sensitive information is recoverable through higher-order interactions.

Theoretical Properties. We provide an information-theoretic characterization of the channel-wise adversarial objective in equation 14. Let S be the sensitive attribute with entropy $H(S)$, and let $\tilde{\mathbf{X}}_k$ denote the random variable corresponding to the k-th masked channel feature $\tilde{\mathbf{x}}_{i,k}$. For fixed encoder

parameters θ, minimizing the per-channel cross-entropy yields the Bayes discriminator $D^*_{\psi_k}(\tilde{\mathbf{x}}_{i,k}) = p_\theta(s_i \,|\, \tilde{\mathbf{x}}_{i,k})$, and the inner optimum equals the average conditional entropy:

$$\min_{\{\psi_k\}} \mathcal{L}_{\mathrm{CAL}}(\theta, \{\psi_k\}) \;=\; \frac{1}{K}\sum_{k=1}^{K} H\Big(S \,\Big|\, \tilde{\mathbf{X}}_k\Big). \tag{18}$$

The outer maximization over θ is equivalent to minimizing the average mutual information.

$$\max_\theta \frac{1}{K}\sum_k H(S \mid \tilde{\mathbf{X}}_k) \iff \min_\theta \frac{1}{K}\sum_k I_\theta\Big(S; \tilde{\mathbf{X}}_k\Big). \tag{19}$$

since $I_\theta(S; \tilde{\mathbf{X}}_k) = H(S) - H(S \mid \tilde{\mathbf{X}}_k)$. Let $\tilde{\mathbf{X}} = [\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_K]$ be the concatenated masked representation; by the chain rule of mutual information, $I_\theta(S; \tilde{\mathbf{X}}) = \sum_k I_\theta(S; \tilde{\mathbf{X}}_k \mid \tilde{\mathbf{X}}_{1:(k-1)}) \le \sum_k I_\theta(S; \tilde{\mathbf{X}}_k)$. Under the mild condition that $\{\tilde{\mathbf{X}}_k\}$ are approximately conditionally uncorrelated given S (enforced in practice by minimizing $\sum_{k\neq\ell} \|\mathrm{Cov}(\tilde{\mathbf{X}}_k, \tilde{\mathbf{X}}_\ell \mid S)\|_F^2$), the upper bound becomes tight, so minimizing $\frac{1}{K}\sum_k I_\theta(S; \tilde{\mathbf{X}}_k)$ effectively reduces global leakage $I_\theta(S; \tilde{\mathbf{X}})$. For any downstream hash mapping $\mathbf{B} = h(\tilde{\mathbf{X}}) \in \{-1, +1\}^L$, data processing further gives $I_\theta(S; \mathbf{B}) \le I_\theta(S; \tilde{\mathbf{X}}) \le \sum_k I_\theta(S; \tilde{\mathbf{X}}_k)$; moreover, for any bounded retrieval score $g : \{-1, +1\}^L \to [0, 1]$, Pinsker-type inequalities imply $\big|\mathbb{E}[g(\mathbf{B}) \mid S{=}a] - \mathbb{E}[g(\mathbf{B}) \mid S{=}b]\big| \le 2\,\mathrm{TV}(P_{\mathbf{B}|S=a}, P_{\mathbf{B}|S=b}) \le C\sqrt{I_\theta(S; \mathbf{B})} \le C\sqrt{\sum_k I_\theta(S; \tilde{\mathbf{X}}_k)}$ for a universal constant $C > 0$. Thus, reducing the channel-averaged leakage tightens an explicit, information-theoretic bound on downstream retrieval.

4.4 Semantic Alignment

We generate discrete hash codes in the Hamming space from masked representations. Let $f_\eta : \mathbb{R}^d \to \mathbb{R}^L$ be the hash head with parameters η. Given $\tilde{\mathbf{x}}_i$ from equation 13, we compute

$$\mathbf{u}_i = f_\eta(\tilde{\mathbf{x}}_i), \qquad \mathbf{b}_i = \mathrm{sign}(\mathbf{u}_i) \in \{-1, +1\}^L. \tag{20}$$

and use a differentiable relaxation $\mathbf{r}_i = \tanh(\mathbf{u}_i)$ for backpropagation during training. To inject class semantics, we maintain $C = |\mathcal{Y}|$ *semantic anchors* $\{\mathbf{z}_c\}_{c=1}^C$ in the Hamming space, each $\mathbf{z}_c \in \{-1, +1\}^L$. Anchors are initialized by i.i.d. Rademacher sampling (each bit ± 1 with probability $1/2$), which yields an expected inter-class Hamming distance of $L/2$ and thus large initial separation. The semantic alignment loss is defined as follows, where $\tau_s > 0$ is a temperature.

$$\mathcal{L}_{\mathrm{SA}}(\theta, \eta, \{\mathbf{z}_c\}) = -\sum_{i\in\mathcal{B}} \log \frac{\exp\big(\mathbf{z}_{y_i}^\top \mathbf{r}_i/\tau_s\big)}{\sum_{c=1}^C \exp\big(\mathbf{z}_c^\top \mathbf{r}_i/\tau_s\big)}. \tag{21}$$

4.5 Overall Objective

Let θ denote the encoder and prototype parameters, $\Psi = \{\psi_k\}_{k=1}^K$ the per-channel discriminators, and η the hash head. We adopt a single saddle-point formulation that unifies disentangled consistency, channel adversarial learning, conditional covariance regularization, and semantic alignment:

$$\min_{\theta,\eta,\{\mathbf{z}_c\}} \max_{\{\psi_k\}} \Big[\underbrace{\mathcal{L}_{\mathrm{DCL}}(\theta)}_{\substack{\text{disentangled consistency}\\ \text{(min)}}} - \lambda_1 \underbrace{\mathcal{L}_{\mathrm{CAL}}(\theta, \{\psi_k\})}_{\substack{\text{channel adversary}\\ \text{(max by } \theta\text{, min by } \psi_k)}} + \lambda_2 \underbrace{\mathcal{L}_{\mathrm{CCR}}(\theta)}_{\substack{\text{conditional covariance}\\ \text{(min)}}} + \underbrace{\mathcal{L}_{\mathrm{SA}}(\theta, \eta, \{\mathbf{z}_c\})}_{\substack{\text{semantic alignment}\\ \text{(min)}}}\Big]. \tag{22}$$

Here, $\lambda_1 > 0$ controls the strength of channel adversarial learning and $\lambda_2 > 0$ controls the strength of conditional covariance regularization. $\mathcal{L}_{\mathrm{DCL}}(\theta)$ is the *disentangled consistency* loss (Eq. 11), encouraging factor assignments. $\mathcal{L}_{\mathrm{CAL}}(\theta, \{\psi_k\})$ is the *channel adversarial* loss (Eq. 14), where each discriminator ψ_k is trained to minimize sensitive prediction error while the encoder maximizes it, thus suppressing recoverable sensitive information in masked factors. $\mathcal{L}_{\mathrm{CCR}}(\theta)$ is the *conditional covariance regularization* loss (Eq. 17), penalizing cross-channel covariance given sensitive labels to reduce joint sensitive leakage. Finally, $\mathcal{L}_{\mathrm{SA}}(\theta, \eta, \{\mathbf{z}_c\})$ is the *semantic alignment* loss (Eq. 21), pulling hash logits in Hamming space. We provide an algorithm procedure in the appendix B.

5 EXPERIMENT

5.1 DATASET

We evaluated on two facial attribute datasets, UTKFace (Zhang et al., 2017) and CelebA (Liu et al., 2015). UTKFace ($\approx$ 20K images with age/gender/ethnicity annotations) is used in two configurations. In the first, *Ethnicity* is the target label (five categories) and *Age* serves as the sensitive attribute, binarized as $<$35 vs. $\geq$35. In the second, *Age* becomes the target (five bins: 0–20, 20–40, 40–60, 60–80, 80+) and *Ethnicity* is the sensitive attribute, binarized as European American vs. non–European American. CelebA ($\approx$ 200K images with 40 binary attributes) is used with *Attractive* as target and *Male* as the sensitive attribute. For retrieval evaluation, we randomly draw 100 query images on UTKFace and 500 on CelebA; all remaining images are used for training and retrieval.

5.2 EVALUATION METRICS

We evaluate two aspects: retrieval accuracy and group fairness.

Retrieval accuracy. *Mean Average Precision*: $\mathrm{MAP} = \frac{1}{|Q|}\sum_{q\in Q}\frac{1}{m_q}\sum_{k=1}^{n_q}\mathrm{Precision@}k(q)\cdot \mathrm{rel}_q(k)$, where Q is the query set, m_q is the number of relevant items for query q, n_q is the list length, and $\mathrm{rel}_q(k)\in\{0,1\}$ indicates relevance at rank k.

Fairness Metrics. We use DP, EOP, and EOD as our fairness measures.

Demographic Parity (DP): $\mathrm{DP} = \left|P(\hat{Q}_i \mid S_i{=}1) - P(\hat{Q} \mid S_i{=}0)\right|$

Equal Opportunity (EOP): $\mathrm{EOP} = \left|P(\hat{Q}_i \mid Y_i{=}1, S_i{=}1) - P(\hat{Q}_i \mid Y_i{=}1, S_i{=}0)\right|$

Equalized Odds (EOD): $\mathrm{EOD} = \left|P(\hat{Q}_i \mid Y_i{=}y, S_i{=}1) - P(\hat{Q}_i \mid Y_i{=}y, S_i{=}0)\right|_{y\in\{0,1\}}$

5.3 PERFORMANCE COMPARISON

We conduct extensive experiments on three benchmark datasets (UTKFace with two target–sensitive configurations and CelebA) under hash code lengths ranging from 16 to 128 bits, and compare against a diverse set of competitive baselines; the full list is provided in appendix C. Each experiment is repeated five times with different random seeds, and we report the mean and standard deviation of all metrics to ensure statistical robustness. The results consistently show that DISH achieves the best performance in terms of both retrieval accuracy (MAP) and fairness measures (DP, EOP, and EOD), thereby representing a clear Pareto improvement over existing baselines. For example, on UTKFace with ethnicity as the target attribute, DISH reaches a MAP of 72.99 at 16 bits while simultaneously reducing EOP and EOD to 2.08 and 4.22, respectively. Similar trends hold across longer code lengths, the alternative UTKFace setting with age as target, and the CelebA dataset. These consistent gains can be attributed to the design of our framework: disentangled consistency learning ensures that semantic information is stably concentrated within factor-specific channels; probability-driven channel masking together with channel-wise adversarial learning effectively suppresses sensitive leakage at the per-channel level; and conditional covariance regularization further mitigates cross-channel correlations that could reintroduce bias. Finally, semantic alignment in the Hamming space preserves inter-class discriminability after binarization.

Table 1: Performance comparison (%) with the state-of-the-art methods on UTKFace with code lengths varying from 16 to 128. Target Attribute: ethnicity, Sensitive Attribute: age.

Method	16 bits				32 bits				64 bits				128 bits			
	MAP ↑	EOD ↓	EOP ↓	DP ↓	MAP ↑	EOD ↓	EOP ↓	DP ↓	MAP ↑	EOD ↓	EOP ↓	DP ↓	MAP ↑	EOD ↓	EOP ↓	DP ↓
OrthoHash	57.92 ± 1.04	11.03 ± 0.60	6.83 ± 0.22	7.43 ± 0.95	59.34 ± 1.30	11.30 ± 1.45	7.08 ± 0.31	7.81 ± 1.28	61.90 ± 2.45	10.95 ± 1.57	6.92 ± 0.47	7.89 ± 1.29	62.79 ± 3.10	10.57 ± 1.01	6.50 ± 0.46	8.18 ± 0.41
Bihalf	52.35 ± 1.63	13.88 ± 2.27	9.20 ± 0.68	7.24 ± 1.19	55.26 ± 1.10	11.33 ± 1.76	7.36 ± 1.04	7.28 ± 1.33	56.93 ± 1.49	11.00 ± 1.76	6.87 ± 0.71	7.50 ± 1.93	55.94 ± 1.10	11.28 ± 1.82	7.17 ± 0.86	7.44 ± 1.48
CE	55.62 ± 2.37	10.48 ± 0.80	7.03 ± 0.42	6.01 ± 0.81	58.58 ± 1.38	10.02 ± 1.32	6.76 ± 0.58	6.76 ± 1.00	59.35 ± 1.33	11.19 ± 0.68	7.55 ± 0.34	7.37 ± 0.68	60.05 ± 0.87	10.57 ± 1.06	7.02 ± 0.55	7.12 ± 0.42
CSQ	63.14 ± 1.25	9.41 ± 1.50	5.35 ± 0.37	7.92 ± 0.44	64.23 ± 1.14	9.01 ± 0.42	5.19 ± 0.08	7.91 ± 0.44	65.27 ± 1.02	8.68 ± 0.82	5.16 ± 0.29	7.75 ± 0.21	62.14 ± 3.89	9.15 ± 1.38	5.51 ± 0.38	7.50 ± 1.03
DFH	46.74 ± 2.75	14.99 ± 1.79	8.23 ± 0.77	9.27 ± 1.44	55.07 ± 2.59	13.75 ± 3.01	7.97 ± 1.23	8.79 ± 1.95	59.20 ± 3.07	14.12 ± 1.23	7.92 ± 0.39	9.12 ± 1.17	60.64 ± 2.87	16.40 ± 4.21	10.32 ± 1.66	9.60 ± 0.85
DPSH	62.30 ± 1.13	9.47 ± 0.87	5.41 ± 0.17	8.11 ± 0.56	62.99 ± 0.59	10.27 ± 0.66	6.03 ± 0.25	8.48 ± 0.25	64.12 ± 0.46	10.50 ± 0.69	6.33 ± 0.11	8.59 ± 0.42	63.71 ± 1.24	10.19 ± 1.00	6.09 ± 0.23	8.43 ± 0.61
DTSH	60.83 ± 1.63	9.09 ± 0.40	5.37 ± 0.23	7.74 ± 0.28	61.53 ± 1.97	11.91 ± 1.33	7.30 ± 0.28	8.89 ± 1.02	60.83 ± 0.38	11.93 ± 1.45	7.63 ± 0.19	8.69 ± 0.98	60.61 ± 1.74	10.94 ± 1.12	7.06 ± 0.27	8.01 ± 0.40
GreedyHash	64.62 ± 1.23	10.08 ± 1.39	6.51 ± 0.50	7.87 ± 0.62	64.63 ± 1.41	10.42 ± 0.60	7.22 ± 0.40	7.70 ± 0.41	63.78 ± 0.78	10.63 ± 0.66	7.30 ± 0.40	7.74 ± 0.46	55.92 ± 9.19	10.60 ± 1.22	6.08 ± 0.87	8.26 ± 0.94
SDH-C	57.48 ± 1.51	10.15 ± 2.17	6.13 ± 0.13	6.73 ± 1.72	62.77 ± 0.52	11.78 ± 1.69	7.81 ± 0.28	8.29 ± 0.78	63.48 ± 0.32	11.82 ± 0.38	8.18 ± 0.23	7.75 ± 0.35	63.98 ± 1.31	12.25 ± 0.83	8.65 ± 0.22	7.57 ± 0.12
DLBD	29.17 ± 1.26	17.39 ± 1.29	10.16 ± 0.33	7.79 ± 0.51	29.22 ± 1.26	17.47 ± 1.29	10.35 ± 0.33	7.73 ± 0.41	30.59 ± 1.18	16.51 ± 1.17	9.79 ± 0.19	7.31 ± 0.44	31.44 ± 1.13	17.50 ± 1.94	10.31 ± 0.93	7.71 ± 1.00
MDSHC	62.88 ± 1.66	8.64 ± 1.13	5.96 ± 0.67	7.04 ±0.32	65.75 ± 1.65	7.77 ± 0.53	5.16 ± 0.28	6.69 ± 0.31	64.11 ± 1.33	8.84 ± 0.33	6.69 ± 0.89	6.90 ± 0.15	62.78 ± 0.68	8.62 ± 1.14	5.39 ± 0.25	7.76 ± 0.41
FATE	70.01 ± 1.27	4.47 ± 1.26	2.57 ± 0.52	5.09 ±0.32	69.98 ± 1.18	5.37 ± 0.64	3.96 ± 0.28	6.60 ± 0.33	72.22 ± 1.46	6.12 ± 0.44	3.39 ± 0.42	6.03 ± 0.75	72.62 ± 1.18	6.51 ± 0.54	3.68 ± 0.35	6.42 ± 0.31
DISH	**72.99 ± 1.32**	**4.22 ± 0.75**	**2.08 ± 0.75**	**4.79 ± 0.15**	**73.20 ± 1.64**	**4.12 ± 0.57**	**2.55 ± 0.49**	**6.54 ± 0.28**	**73.67 ± 1.83**	**4.28 ± 0.98**	**2.96 ± 0.84**	**5.63 ± 0.37**	**73.33 ± 0.91**	**3.66 ± 0.64**	**2.37 ± 0.45**	**6.10 ± 0.26**

Table 2: Performance comparison (%) with the state-of-the-art methods on UTKFace with code lengths varying from 16 to 128. Target Attribute: age, Sensitive Attribute: ethnicity.

Method	16 bits				32 bits				64 bits				128 bits			
	MAP ↑	EOD ↓	EOP ↓	DP ↓	MAP ↑	EOD ↓	EOP ↓	DP ↓	MAP ↑	EOD ↓	EOP ↓	DP ↓	MAP ↑	EOD ↓	EOP ↓	DP ↓
OrthoHash	56.41 ± 1.48	12.97 ± 1.48	4.87 ± 0.27	11.08 ± 0.65	58.48 ± 0.89	14.91 ± 2.19	5.76 ± 0.14	12.45 ± 1.55	59.35 ± 0.19	15.20 ± 1.43	5.93 ± 0.09	12.98 ± 0.70	60.64 ± 0.86	14.80 ± 0.84	5.60 ± 0.3	11.60 ± 0.41
Bihalf	52.55 ± 2.2	19.97 ± 4.56	7.56 ± 0.37	15.07 ± 2.86	53.30 ± 2.33	18.88 ± 3.14	6.78 ± 1.45	14.42 ± 1.31	54.46 ± 1.2	18.84 ± 0.7	6.59 ± 0.35	14.46 ± 0.55	53.58 ± 1.15	17.53 ± 0.92	6.60 ± 0.38	12.57 ± 0.82
CE	57.87 ± 1.5	15.66 ± 0.58	5.62 ± 0.67	12.90 ± 0.79	57.08 ± 0.86	15.15 ± 0.58	5.28 ± 0.22	12.29 ± 0.39	57.93 ± 1.6	16.02 ± 1.15	5.79 ± 0.56	12.74 ± 0.55	57.67 ± 0.61	15.03 ± 0.88	5.23 ± 0.28	12.37 ± 0.51
CSQ	62.11 ± 1.37	12.44 ± 1.35	4.18 ± 0.79	11.75 ± 1.54	64.16 ± 1.06	12.87 ± 0.78	4.82 ± 0.48	11.89 ± 0.68	64.61 ± 1.56	11.82 ± 0.97	4.48 ± 0.23	10.93 ± 0.53	64.73 ± 1.22	12.18 ± 0.5	4.28 ± 0.3	13.46 ± 0.56
DFH	56.62 ± 1.24	13.72 ± 0.94	4.35 ± 0.61	12.43 ± 0.99	55.91 ± 1.66	14.98 ± 0.55	5.60 ± 0.36	12.38 ± 0.47	57.72 ± 1.17	15.81 ± 2.13	6.12 ± 0.54	12.81 ± 1.15	57.60 ± 1.39	16.75 ± 1.04	6.43 ± 0.62	15.29 ± 1.00
DPSH	57.64 ± 1.51	16.37 ± 0.48	5.93 ± 0.55	14.18 ± 0.43	56.84 ± 1.11	17.27 ± 1.4	6.32 ± 0.36	14.78 ± 0.92	57.46 ± 1.01	17.15 ± 0.44	6.35 ± 0.42	14.40 ± 0.28	58.15 ± 0.41	17.01 ± 0.55	6.11 ± 0.41	15.17 ± 0.64
DTSH	58.04 ± 0.87	15.96 ± 1.44	5.75 ± 0.27	13.83 ± 0.74	58.08 ± 1.15	16.33 ± 0.73	5.79 ± 0.26	14.20 ± 0.52	56.73 ± 1.17	16.37 ± 0.52	5.81 ± 0.17	14.14 ± 0.33	56.76 ± 1.27	16.80 ± 1.55	5.84 ± 0.29	14.22 ± 0.72
GreedyHash	61.35 ± 1.89	14.99 ± 0.7	5.32 ± 0.12	12.40 ± 0.2	61.59 ± 1.05	15.64 ± 1.04	5.91 ± 0.4	12.71 ± 0.69	60.68 ± 0.56	14.90 ± 0.43	5.51 ± 0.36	12.15 ± 0.28	58.96 ± 0.64	14.95 ± 0.68	5.77 ± 0.15	11.74 ± 0.63
SDH-C	50.36 ± 1.13	13.77 ± 2.93	4.74 ± 0.92	11.28 ± 1.35	59.75 ± 1.61	16.17 ± 0.53	6.04 ± 0.12	12.58 ± 0.47	60.44 ± 0.87	16.02 ± 0.28	6.10 ± 0.21	12.22 ± 0.25	60.69 ± 0.75	15.86 ± 0.85	6.31 ± 0.32	11.87 ± 0.41
DLBD	35.62 ± 1.92	20.99 ± 1.20	11.12 ± 1.96	10.70 ± 0.51	39.05 ± 1.52	20.83 ± 2.12	11.19 ± 0.05	10.98 ± 2.25	39.00 ± 1.21	21.90 ± 1.61	13.15 ± 0.22	9.93 ± 1.53	40.24 ± 0.92	22.69 ± 1.44	12.87 ± 0.43	11.24 ± 2.19
MDSHC	64.59 ± 1.28	14.45 ± 0.63	5.48 ± 0.57	9.18 ±0.35	60.47 ± 1.19	13.10 ± 0.19	5.92 ± 0.40	9.34 ± 0.29	61.08 ± 0.41	11.64 ± 0.51	5.35 ± 0.42	11.55 ± 0.13	61.73 ± 0.46	11.85 ± 0.45	5.11 ± 0.37	11.26 ± 0.59
FATE	69.78 ± 1.76	7.74 ± 0.73	2.06 ± 0.76	8.18 ±0.18	69.52 ± 1.36	7.26 ± 0.33	2.54 ± 0.31	8.43 ± 0.25	68.61 ± 1.74	9.11 ± 0.56	2.86 ± 0.14	10.31 ± 0.35	70.17 ± 1.13	7.92 ± 0.32	2.32 ± 0.25	9.35 ± 0.41
DISH	**72.08 ± 0.71**	**6.46 ± 0.49**	**1.41 ± 0.38**	**7.08 ± 0.39**	**71.85 ± 1.82**	**6.76 ± 0.24**	**1.59 ± 0.22**	**8.14 ± 0.26**	**71.59 ± 0.51**	**6.03 ± 0.35**	**1.29 ± 0.20**	**8.83 ± 0.38**	**72.40 ± 1.35**	**6.34 ± 0.68**	**2.09 ± 0.32**	**8.51 ± 0.52**

Table 3: Performance comparison (%) with the state-of-the-art methods on CelebA with code lengths varying from 16 to 128. Target Attribute: attractiveness, Sensitive Attribute: male.

Method	16 bits				32 bits				64 bits				128 bits			
	MAP ↑	EOD ↓	EOP ↓	DP ↓	MAP ↑	EOD ↓	EOP ↓	DP ↓	MAP ↑	EOD ↓	EOP ↓	DP ↓	MAP ↑	EOD ↓	EOP ↓	DP ↓
OrthoHash	76.82 ± 0.52	4.09 ± 0.05	2.87 ± 0.23	2.99 ± 0.22	77.91 ± 0.06	4.97 ± 0.40	3.59 ± 0.38	3.35 ± 0.15	78.25 ± 0.90	4.69 ± 0.33	3.45 ± 0.28	3.38 ± 0.14	78.90 ± 0.66	3.52 ± 1.38	2.69 ± 0.96	3.03 ± 0.42
Bihalf	77.87 ± 0.49	3.19 ± 1.09	2.35 ± 0.78	2.66 ± 0.11	78.69 ± 0.23	2.91 ± 0.63	2.25 ± 0.54	2.71 ± 0.21	78.09 ± 0.37	3.51 ± 0.22	3.40 ± 0.15	3.66 ± 0.15	78.34 ± 0.76	3.19 ± 0.28	2.38 ± 0.18	2.54 ± 0.15
CE	77.33 ± 0.42	2.98 ± 0.68	2.70 ± 0.56	2.40 ± 0.13	78.10 ± 0.93	3.88 ± 0.74	2.88 ± 0.56	2.90 ± 0.33	77.27 ± 0.13	4.56 ± 0.35	3.35 ± 0.28	3.13 ± 0.11	77.97 ± 0.58	4.76 ± 0.14	3.54 ± 0.18	3.18 ± 0.13
CSQ	78.66 ± 1.02	2.93 ± 0.49	2.45 ± 0.34	2.70 ± 0.18	78.27 ± 1.59	2.98 ± 0.55	2.48 ± 0.35	2.71 ± 0.23	77.92 ± 0.18	3.76 ± 0.45	3.37 ± 0.35	2.62 ± 0.16	77.92 ± 2.31	2.91 ± 0.39	2.46 ± 0.22	2.70 ± 0.16
DFH	78.71 ± 0.88	2.88 ± 0.11	2.42 ± 0.11	2.66 ± 0.05	78.89 ± 0.73	2.94 ± 0.11	2.56 ± 0.21	2.74 ± 0.11	77.29 ± 1.04	3.83 ± 0.24	2.92 ± 0.32	2.63 ± 0.22	77.02 ± 0.18	3.47 ± 0.19	2.46 ± 0.19	2.31 ± 0.12
DPSH	78.37 ± 0.96	4.33 ± 1.25	3.22 ± 0.79	2.85 ± 0.67	79.01 ± 0.86	2.92 ± 0.98	2.21 ± 0.74	2.46 ± 0.16	77.47 ± 0.04	5.09 ± 1.94	3.80 ± 1.23	3.15 ± 0.72	78.37 ± 2.95	4.84 ± 1.57	3.59 ± 1.42	2.94 ± 0.74
DTSH	78.69 ± 0.51	3.35 ± 0.12	2.52 ± 0.21	2.35 ± 0.04	78.28 ± 0.88	3.25 ± 0.50	2.46 ± 0.39	2.48 ± 0.15	77.23 ± 1.05	3.31 ± 0.59	2.52 ± 0.58	2.42 ± 0.31	77.70 ± 2.01	3.58 ± 0.22	2.58 ± 0.01	2.46 ± 0.06
GreedyHash	78.62 ± 0.89	3.18 ± 0.31	2.83 ± 0.11	2.53 ± 0.14	77.74 ± 1.24	3.52 ± 0.26	2.57 ± 0.11	2.37 ± 0.12	77.57 ± 1.19	3.73 ± 0.65	2.70 ± 0.37	2.54 ± 0.16	77.35 ± 0.78	4.06 ± 0.43	2.94 ± 0.28	2.71 ± 0.06
SDH-C	78.79 ± 0.79	3.09 ± 0.19	2.33 ± 0.17	2.54 ± 0.17	78.37 ± 0.15	3.58 ± 0.24	2.55 ± 0.22	2.54 ± 0.32	78.05 ± 0.26	3.34 ± 0.11	2.54 ± 0.12	2.41 ± 0.14	77.23 ± 0.81	3.37 ± 0.39	2.43 ± 0.23	2.40 ± 0.18
DLBD	66.91 ± 0.43	5.98 ± 0.16	3.61 ± 0.10	3.16 ± 0.07	68.10 ± 0.43	6.68 ± 0.29	4.04 ± 0.19	3.54 ± 0.13	68.76 ± 0.46	6.84 ± 0.33	4.16 ± 0.20	3.64 ± 0.15	68.53 ± 0.30	7.06 ± 0.36	4.28 ± 0.21	3.74 ± 0.16
MDSHC	77.14 ± 2.65	3.40 ± 0.52	2.42 ± 0.30	2.38 ± 0.21	76.30 ± 2.21	3.78 ± 0.76	2.63 ± 0.46	2.52 ± 0.28	75.39 ± 1.70	4.35 ± 0.86	3.04 ± 0.56	2.60 ± 0.39	75.61 ± 1.16	5.37 ± 0.48	3.73 ± 0.30	3.12 ± 0.21
FATE	76.99 ± 0.76	2.40 ± 0.25	2.26 ± 0.20	2.07 ± 0.07	76.07 ± 2.43	2.52 ± 0.51	1.75 ± 0.35	2.01 ± 0.14	74.61 ± 1.74	3.71 ± 0.56	2.86 ± 0.14	2.31 ± 0.35	75.95 ± 0.83	5.31 ± 0.27	3.93 ± 0.23	3.42 ± 0.12
DISH	**78.92 ± 0.48**	**2.08 ± 0.22**	**2.06 ± 0.17**	**1.91 ± 0.06**	**79.26 ± 1.30**	**2.08 ± 0.21**	**1.68 ± 0.12**	**1.82 ± 0.11**	**78.29 ± 0.23**	**3.23 ± 0.18**	**2.48 ± 0.16**	**1.91 ± 0.04**	**79.10 ± 0.29**	**2.82 ± 0.25**	**2.23 ± 0.18**	**2.16 ± 0.13**

5.4 Ablation Study

To further validate the effectiveness of different components in the proposed DISH framework, we conduct a series of ablation studies on the UTKFace dataset with 128-bit hash codes. Each experiment is repeated five times under different random seeds, and we report both mean and standard deviation to ensure statistical reliability. Specifically, we examine the contribution of each loss function, compare different masking strategies, and evaluate several model variants to better understand how the architectural choices affect retrieval accuracy and fairness.

Table 4: Performance comparison (%) with different masking methods with 128bits on UTKFace

$\mathcal{L}_{\text{DCL}}$	$\mathcal{L}_{\text{CAL}}$	$\mathcal{L}_{\text{CCR}}$	TA: ethnicity, SA: age				TA: age, SA: ethnicity			
			MAP ↑	EOD ↓	EOP ↓	DP ↓	MAP ↑	EOD ↓	EOP ↓	DP ↓
-	-	-	61.57 ± 1.25	10.73 ± 1.43	7.71 ± 0.64	7.84 ± 0.83	59.10 ± 1.27	14.01 ± 0.67	5.89 ± 0.44	11.06 ± 0.97
✓			62.14 ± 1.21	9.93 ± 0.87	7.70 ± 0.61	7.06 ± 0.54	61.39 ± 0.73	12.82 ± 1.13	5.36 ± 0.63	10.64 ± 0.72
	✓		64.37 ± 1.78	8.95 ± 0.61	6.62 ± 0.54	6.77 ± 0.44	64.63 ± 1.00	10.77 ± 0.93	4.70 ± 0.68	9.41 ± 0.50
		✓	64.58 ± 2.38	9.24 ± 1.84	5.74 ± 1.25	8.53 ± 0.90	67.82 ± 2.37	11.78 ± 1.54	3.76 ± 0.90	11.59 ± 0.67
✓	✓		68.07 ± 0.66	7.23 ± 0.18	3.10 ± 0.15	8.43 ± 0.13	70.52 ± 1.03	6.84 ± 0.06	2.36 ± 0.11	8.92 ± 0.14
✓		✓	67.76 ± 1.02	6.24 ± 0.97	3.53 ± 0.75	6.39 ± 0.69	68.04 ± 1.79	8.01 ± 1.27	2.58 ± 0.68	9.01 ± 0.68
	✓	✓	68.41 ± 1.03	5.99 ± 0.57	3.45 ± 0.43	6.49 ± 0.43	67.43 ± 0.30	7.37 ± 0.24	2.41 ± 0.21	9.18 ± 0.06
✓	✓	✓	**73.33 ± 0.91**	**3.66 ± 0.64**	**2.37 ± 0.45**	**6.10 ± 0.26**	**72.40 ± 1.35**	**6.34 ± 0.68**	**2.09 ± 0.32**	**8.51 ± 0.52**

Table 5: Performance comparison (%) with different masking methods with 128 bits on UTKFace

Method	TA: ethnicity, SA: age				TA: age, SA: ethnicity			
	MAP ↑	EOD ↓	EOP ↓	DP ↓	MAP ↑	EOD ↓	EOP ↓	DP ↓
No-Mask	72.10 ± 1.43	4.35 ± 0.90	2.78 ± 0.61	6.45 ± 0.35	71.70 ± 1.62	6.82 ± 0.80	2.36 ± 0.38	8.74 ± 0.60
Random	71.34 ± 1.53	4.41 ± 0.95	2.83 ± 0.62	6.79 ± 0.46	70.60 ± 1.55	6.90 ± 0.82	3.01 ± 0.47	8.80 ± 0.62
1/K	72.95 ± 1.10	3.88 ± 0.70	2.49 ± 0.50	6.22 ± 0.30	72.10 ± 1.20	6.55 ± 0.75	2.88 ± 0.33	8.60 ± 0.45
DISH	**73.33 ± 0.91**	**3.66 ± 0.64**	**2.37 ± 0.45**	**6.10 ± 0.26**	**72.40 ± 1.35**	**6.34 ± 0.68**	**2.09 ± 0.32**	**8.51 ± 0.52**

Table 6: Performance comparison (%) with different model variants with 128 bits on UTKFace

Method	TA: ethnicity, SA: age				TA: age, SA: ethnicity			
	MAP ↑	EOD ↓	EOP ↓	DP ↓	MAP ↑	EOD ↓	EOP ↓	DP ↓
Variant 1	70.40 ± 2.95	6.35 ± 2.10	3.28 ± 1.60	6.90 ± 0.95	70.70 ± 2.20	7.25 ± 1.12	2.38 ± 0.55	8.85 ± 1.00
Variant 2	71.10 ± 1.35	4.92 ± 0.75	2.70 ± 0.50	6.60 ± 0.38	71.50 ± 1.18	7.05 ± 0.72	2.25 ± 0.42	8.88 ± 0.56
Variant 3	72.05 ± 1.12	4.50 ± 0.70	2.62 ± 0.50	6.35 ± 0.34	71.90 ± 1.22	6.80 ± 0.68	2.20 ± 0.36	8.65 ± 0.52
Variant 4	70.95 ± 1.55	6.15 ± 0.82	3.18 ± 0.62	6.95 ± 0.42	71.05 ± 1.48	7.50 ± 0.80	2.45 ± 0.46	9.00 ± 0.60
DISH	$\mathbf{73.33 \pm 0.91}$	$\mathbf{3.66 \pm 0.64}$	$\mathbf{2.37 \pm 0.45}$	$\mathbf{6.10 \pm 0.26}$	$\mathbf{72.40 \pm 1.35}$	$\mathbf{6.34 \pm 0.68}$	$\mathbf{2.09 \pm 0.32}$	$\mathbf{8.51 \pm 0.52}$

Effect Of Individual Loss Functions. Table 4 shows that removing all fairness-related objectives leads to a substantial drop in both MAP and fairness metrics. Adding the disentangled consistency loss alone yields modest MAP gains and small yet consistent reductions in disparities, as it mainly promotes semantic concentration rather than directly suppressing sensitive leakage. Introducing channel adversarial loss or conditional covariance regularization independently improves fairness and also increases MAP, though the improvements are smaller and less balanced than the full model. Combined with disentangled consistency, each component contributes complementary benefits, and the full objective achieves the best overall trade-off between accuracy and fairness.

Comparison Of Masking Strategies. Table 5 presents a comparison between the proposed probability-driven channel masking and several alternative masking schemes. The no-mask and random baselines achieve reasonable MAP but exhibit higher group disparities, as they fail to leverage the learned factor assignment structure and therefore cannot effectively concentrate semantic evidence. A naive $1/K$ uniform weighting offers marginal improvements by balancing channel contributions, but it still neglects the semantic concentration property learned through disentanglement. In contrast, our probability-driven masking adaptively emphasizes channels with higher assignment probabilities while attenuating those with weaker semantic relevance. Importantly, this process does not rely on target labels or sensitive attributes; instead, it exploits the intrinsic factor probabilities to preserve discriminative semantics. As a result, DISH achieves both the highest MAP and the lowest group disparities, demonstrating that adaptive, data-driven channel weighting is critical to simultaneously maintaining retrieval accuracy and enhancing fairness.

Evaluation Of Model Variants. Finally, we investigate four structural variants of DISH, as shown in Table 6. Replacing supervised disentangled consistency learning with unsupervised contrastive learning (Variant 1) weakens semantic concentration, resulting in significantly lower MAP and unstable fairness. Substituting the channel-wise adversary with a single global adversary (Variant 2) yields better results than Variant 1 but still lags behind DISH, highlighting the importance of suppressing sensitive leakage at a finer granularity. Using a simple decorrelation instead of conditional covariance regularization (Variant 3) provides modest improvements, but it lacks sensitivity-specific conditioning and fails to achieve optimal fairness. Constraining fairness only in the Hamming space (Variant 4) further underperforms, confirming that interventions after binarization are insufficient due to the loss of representational flexibility. By contrast, the full DISH framework consistently achieves a better trade-off between accuracy and fairness, validating the necessity of its disentangled, multi-level suppression strategy.

Effects Of Hyperparameters. We assess three factors: ranked samples, channel count K, and loss weights (λ_1, λ_2). For ranked list depth, we compare DISH with FATE, DFH, and CE(Figure 2). For K and (λ_1, λ_2), we conduct ablations of DISH: an intermediate K and moderate loss weights provide better fairness-utility trade-off (Figures 3 and 4). Full results and details are in appendix D.

6 Conclusion

We introduce DISH, a fairness-aware deep hashing framework that combines disentangled consistency, probability-driven masking, channel-wise adversaries, and conditional covariance regularization. Extensive experiments on multiple benchmarks demonstrate that DISH consistently outperforms existing methods, achieving state-of-the-art retrieval accuracy together with substantially improved fairness. The ablation studies further validate the necessity of each component and confirm that their integration yields a clear Pareto improvement.

7 ETHICS STATEMENT

This research addresses a critical ethical concern in large-scale image retrieval systems: the potential for algorithmic bias that may lead to discriminatory outcomes across demographic groups defined by sensitive attributes such as age, gender, and ethnicity. The development and deployment of biased retrieval systems can perpetuate societal inequalities, reinforce stereotypes, and cause harm to marginalized communities through unequal access to information and opportunities. In this work, we explicitly acknowledge the ethical implications of biased representation learning and commit to developing solutions that promote fairness without compromising utility. Our dataset selection and annotation processes were conducted with careful consideration of demographic representation, ensuring balanced coverage across sensitive attribute categories where possible. We strictly adhered to the ethical guidelines established by the original dataset creators (UTKFace and CelebA), which include informed consent protocols for image collection and usage. Importantly, our framework DISH actively works to disentangle and suppress sensitive information while preserving semantic utility, thereby reducing the risk of discriminatory retrieval results. We recognize that fairness is a complex, context-dependent concept, and our approach focuses on group fairness metrics (demographic parity, equal opportunity, and equalized odds) as measurable objectives. While our method significantly improves fairness outcomes, we acknowledge that no technical solution can fully resolve deeply rooted societal biases, and we advocate for continued interdisciplinary collaboration between computer scientists, social scientists, and policymakers to address these challenges holistically.

8 REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our research findings, we have implemented rigorous experimental protocols and documentation practices throughout this study. All experiments were conducted using publicly available datasets (UTKFace and CelebA) with clearly specified preprocessing procedures, including detailed descriptions of how target and sensitive attributes were defined and binarized for each experimental configuration. For each experiment, we report results averaged over five independent runs with different random seeds, presenting both mean values and standard deviations to demonstrate statistical reliability. Our complete experimental setup including network architectures (ResNet-50 backbone with specified modifications), optimization parameters (learning rates, batch sizes, optimizer configurations), and hyperparameter settings (λ_1, λ_2 are explicitly documented in the methodology section and supplementary materials. The ablation studies systematically evaluate the contribution of each component in our framework, providing comprehensive evidence for our design choices. To facilitate comparison with existing methods, we have reimplemented all baseline approaches following their original publications and verified their performance against reported results where possible. We welcome replication attempts and are committed to providing support to researchers seeking to reproduce or build upon our work.

REFERENCES

Sushant Agarwal and Amit Deshpande. On the power of randomization in fair classification and representation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1542–1551, 2022.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, April 2024. ISSN 1557-7341. doi: 10.1145/3616865. URL `http://dx.doi.org/10.1145/3616865`.

L Elisa Celis and Vijay Keswani. Improved adversarial learning for fair classification. *arXiv preprint arXiv:1901.10443*, 2019.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Emily L Denton et al. Unsupervised learning of disentangled representations from video. *Advances in neural information processing systems*, 30, 2017.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2916–2929, 2012.

Adam Gronowski, William Paul, Fady Alajaji, Bahman Gharesifard, and Philippe Burlina. Classification utility, fairness, and compactness via tunable information bottleneck and rényi measures. *IEEE Transactions on Information Forensics and Security*, 19:1630–1645, 2023.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Jiun Tian Hoe, Kam Woh Ng, Tianyu Zhang, Chee Seng Chan, Yi-Zhe Song, and Tao Xiang. One loss for all: Deep hashing with a single cosine similarity based learning objective. *Advances in Neural Information Processing Systems*, 34:24286–24298, 2021.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pp. 2649–2658. PMLR, 2018.

Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pp. 853–862, 2018.

Haoyang Li, Xin Wang, Ziwei Zhang, Zehuan Yuan, Hang Li, and Wenwu Zhu. Disentangled contrastive learning on graphs. *Advances in Neural Information Processing Systems*, 34:21872–21884, 2021.

Qi Li, Zhenan Sun, Ran He, and Tieniu Tan. Deep supervised discrete hashing. *Advances in neural information processing systems*, 30, 2017.

Wu-Jun Li, Sheng Wang, and Wang-Cheng Kang. Feature learning based deep supervised hashing with pairwise labels. *arXiv preprint arXiv:1511.03855*, 2015.

Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–48, 2023.

Yunqiang Li and Jan van Gemert. Deep unsupervised image hashing by maximizing bit entropy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2002–2010, 2021.

Yunqiang Li, Wenjie Pei, Jan van Gemert, et al. Push for quantization: Deep fisher hashing. *arXiv preprint arXiv:1909.00206*, 2019.

Liu Liu, Jiangtong Li, Li Niu, Ruicong Xu, and Liqing Zhang. Activity image-to-video retrieval by disentangling appearance and motion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2145–2153, 2021.

Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang. Supervised hashing with kernels. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2074–2081. IEEE, 2012.

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*, pp. 3730–3738, 2015.

Fangyuan Luo, Honglei Zhang, Tong Li, and Jun Wu. Learning to hash for recommendation: A survey. *arXiv preprint arXiv:2412.03875*, 2024.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Kirtan Padh, Diego Antognini, Emma Lejal-Glaude, Boi Faltings, and Claudiu Musat. Addressing fairness in classification with a model-agnostic multi-objective algorithm. In *Uncertainty in artificial intelligence*, pp. 600–609. PMLR, 2021.

Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.

Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen. Supervised discrete hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 37–45, 2015.

Malcolm Slaney and Michael Casey. Locality-sensitive hashing for finding nearest neighbors [lecture notes]. *IEEE Signal processing magazine*, 25(2):128–131, 2008.

Shupeng Su, Chao Zhang, Kai Han, and Yonghong Tian. Greedy hash: Towards fast optimization for accurate hash coding in cnn. *Advances in neural information processing systems*, 31, 2018.

Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1415–1424, 2017.

Jinfeng Wang, Sifan Song, Jionglong Su, and S Kevin Zhou. Distortion-disentangled contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 75–85, 2024.

Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Semi-supervised hashing for scalable image retrieval. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3424–3431. IEEE, 2010.

Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Semi-supervised hashing for large-scale search. *IEEE transactions on pattern analysis and machine intelligence*, 34(12):2393–2406, 2012.

Liangdao Wang, Yan Pan, Cong Liu, Hanjiang Lai, Jian Yin, and Ye Liu. Deep hashing with minimal-distance-separated hash centers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23455–23464, 2023.

Xiaofang Wang, Yi Shi, and Kris M Kitani. Deep supervised hashing with triplet labels. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*, pp. 70–84. Springer, 2017.

Bin Xiao, Yang Hu, Bo Liu, Xiuli Bi, Weisheng Li, and Xinbo Gao. Dlbd: A self-supervised direct-learned binary descriptor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15846–15855, 2023.

Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, and Jiashi Feng. Central similarity quantization for efficient image and video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3083–3092, 2020.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017.

Fan Zhang, Chong Chen, Xian-Sheng Hua, and Xiao Luo. Fate: Learning effective binary descriptors with group fairness. *IEEE Transactions on Image Processing*, 2024.

Guixian Zhang, Guan Yuan, Debo Cheng, Lin Liu, Jiuyong Li, and Shichao Zhang. Disentangled contrastive learning for fair graph representations. *Neural Networks*, 181:106781, 2025.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5810–5818, 2017.

Yuchang Zhu, Jintang Li, Zibin Zheng, and Liang Chen. Fair graph representation learning via sensitive attribute disentanglement. In *Proceedings of the ACM Web Conference 2024*, pp. 1182–1192, 2024.

A THE USE OF LARGE LANGUAGE MODELS

We used a large language model–based assistant solely for linguistic polishing (e.g., grammar, style, clarity, and minor LaTeX formatting). All technical content, including problem formulation, algorithms, analyses, and conclusions, was conceived and verified by the authors. All references were selected and checked by the authors. Outputs from the assistant were reviewed and edited for accuracy; any remaining errors are the authors' responsibility.

B ALGORITHM PROCEDURE

Algorithm 1 Algorithm procedure of DISH

Input: Dataset $\mathcal{D} = \{I_i, y_i, s_i\}_{i=1}^{N}$, code length L, channel number K, weights (λ_1, λ_2)
Output: Encoder & prototypes θ (incl. $\{\mathbf{c}_k\}_{k=1}^{K}$), hash head η, channel discriminators $\Psi = \{\psi_k\}_{k=1}^{K}$, semantic anchors $\{\mathbf{z}_c\}_{c=1}^{C}$

Init: Randomly initialize θ (encoder and prototypes $\{\mathbf{c}_k\}$), hash head f_η, and discriminators $\{D_{\psi_k}\}_{k=1}^{K}$; set anchors $\{\mathbf{z}_c\}$ by Rademacher.
Main training
for $e = 1$ **to** E **do**
  for minibatches $\mathcal{B} \subset \mathcal{D}$ **do**
    Augment: For each $I_i \in \mathcal{B}$, sample I_i'.
    Encode: $\mathbf{x}_{i,1:K}$, $\mathbf{x}'_{i,1:K} \leftarrow$ encoder θ.
    Factor assignments: $p_\theta(k \,|\, \mathbf{x}_i)$ by Eq. equation 1.
    DCL: compute per-factor $\ell_{i,k}$ and $q_\theta(k \,|\, \mathbf{x}_i, y_i)$ by Eqs. 7–6; get $\mathcal{L}_{\mathrm{DCL}}$ by Eq. 11.
    Masking: $\mathbf{m}_i = [p_\theta(k \,|\, \mathbf{x}_i)]_{k=1}^{K}$ and $\tilde{\mathbf{x}}_i = \mathbf{x}_i \odot_c \mathbf{m}_i$ by Eqs. 12–13.
    Channel adversary: for each k, predict $D_{\psi_k}(\tilde{\mathbf{x}}_{i,k})$ and compute $\mathcal{L}_{\mathrm{CAL}}$ by Eq. 14.
    Conditional covariance: compute $\mathcal{L}_{\mathrm{CCR}}$ on $\{\tilde{\mathbf{x}}_{i,k}\}$ grouped by s_i via Eq. 17.
    Semantic alignment: $\mathbf{u}_i = f_\eta(\tilde{\mathbf{x}}_i)$; compute $\mathcal{L}_{\mathrm{SA}}$ by Eq. 21.
    Joint objective (saddle point): $\mathcal{L} = \mathcal{L}_{\mathrm{DCL}} - \lambda_1 \mathcal{L}_{\mathrm{CAL}} + \lambda_2 \mathcal{L}_{\mathrm{CCR}} + \mathcal{L}_{\mathrm{SA}}$ (Eq. 22)
    Update Ψ: minimize $\mathcal{L}_{\mathrm{CAL}}$ w.r.t. $\{\psi_k\}$ (standard gradient).
    Update $\theta, \eta, \{\mathbf{z}_c\}$: minimize $\mathcal{L}$ (GRL implements the "$-\lambda_1 \mathcal{L}_{\mathrm{CAL}}$" effect on θ).
  end for
end for
Inference: given I, compute $\tilde{\mathbf{x}}$ and $\mathbf{u} = f_\eta(\tilde{\mathbf{x}})$; output $\mathbf{b} = \mathrm{sign}(\mathbf{u})$.

C BASELINES

To validate the effectiveness of our method, we conduct comprehensive comparisons with state-of-the-art deep hashing approaches. The selected baselines include: OrthoHash (Hoe et al., 2021), Bihalf (Li & van Gemert, 2021), CE (Li et al., 2017), CSQ (Yuan et al., 2020), DFH (Li et al., 2019), DPSH (Li et al., 2015) , DTSH (Wang et al., 2017) , GreedyHash (Su et al., 2018), SDH-C (Shen et al., 2015) , DLBD (Xiao et al., 2023), MDSHC (Wang et al., 2023) and FATE (Zhang et al., 2024).

D PARAMETER SENSITIVITY ANALYSIS

We study the sensitivity of DISH to three factors: the number of ranked samples, the channel count K, and the loss weights λ_1 (channel adversary) and λ_2 (conditional covariance). Varying the list depth shows that DISH sustains its fairness advantages over competitive baselines (FATE, DFH, CE) across a wide range of ranked samples without compromising retrieval accuracy, and the relative ordering among methods remains stable (Figure 2). Adjusting K yields a robust U-shaped trend (Figure 3): too few channels under-disentangle factors, while too many fragment semantics and weaken masking/regularization; an intermediate K delivers the best fairness–utility balance across code lengths. The (λ_1, λ_2) landscape (Figure 4) shows a similar concavity: a moderate-to-strong λ_1 paired with a moderate λ_2 forms a stable ridge that maximizes the Pareto frontier.

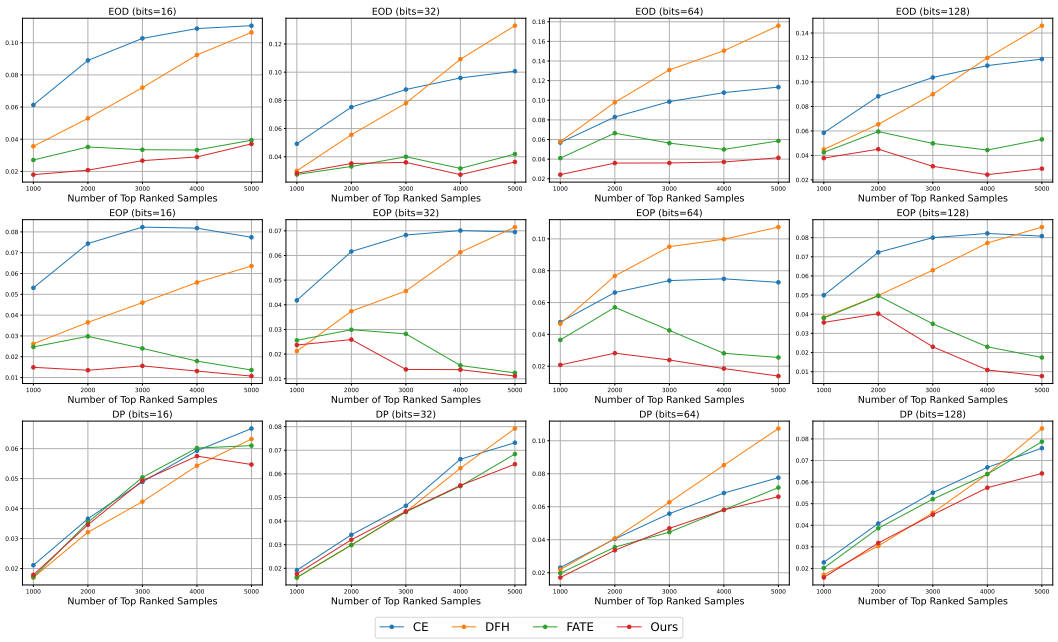

Figure 2: Sensitivity analysis of ranked samples with code lengths 128 on UTKFace. Target Attribute: age , Sensitive Attribute: ethnicity.

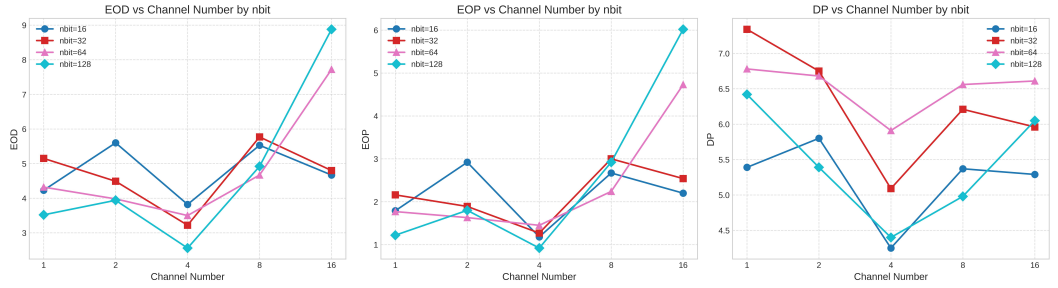

Figure 3: Sensitivity analysis of channel number K with code lengths 128 on UTKFace. Target Attribute: ethnicity, Sensitive Attribute: age.

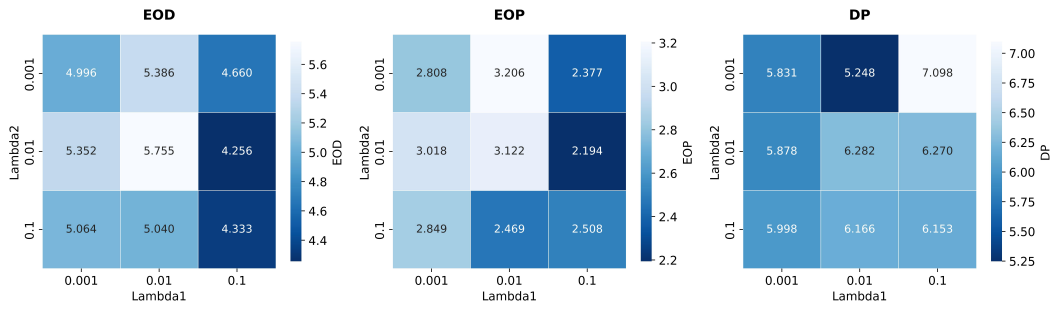

Figure 4: Sensitivity analysis of λ_1 and λ_2 with code lengths 128 on UTKFace. Target Attribute: ethnicity, Sensitive Attribute: age.

E Implementation Details

All experiments were conducted on a single NVIDIA RTX 4090 GPU (24 GB) and, unless otherwise noted, each configuration was trained for 100 epochs. To ensure strict comparability, DISH and all baselines use the same ResNet-50 backbone with identical input resolution, normalization, and augmentation; only method-specific heads and loss terms differ. Data splits and the query/gallery partitions are fixed and reused across methods. For each dataset and configuration, we run five independent trials and report mean $\pm$ std. Within each dataset, all methods share the same training recipe (batch size, optimizer, learning-rate schedule, weight decay, and augmentation pipeline). We control randomness with an identical set of random seeds for every method and trial, applied consistently to Python, NumPy, and PyTorch (including dataloader workers and CUDA determinism);