Appendix A. Dual Permutation Mixture Formulation

Below is the detailed step-by-step transformation from the primal mixture formulation of the adversarial prediction task for bipartite matching (Eq. (2)) to the dual formulation (Eq. (3)):

$$\min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{x\sim\tilde{P};\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}} \left[\text{loss}(\hat{\pi},\check{\pi})\right] \quad \text{s.t.} \quad \mathbb{E}_{x\sim\tilde{P};\check{\pi}|x\sim\check{P}} \left[\sum_{i=1}^{n} \phi_i(x,\check{\pi}_i)\right] = \mathbb{E}_{(x,\pi)\sim\tilde{P}} \left[\sum_{i=1}^{n} \phi_i(x,\pi_i)\right] \tag{17}$$

$$\overset{(a)}{=} \max_{\check{P}(\check{\pi}|x)} \min_{\hat{P}(\hat{\pi}|x)} \mathbb{E}_{x\sim\tilde{P};\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}} \left[\text{loss}(\hat{\pi},\check{\pi})\right] \quad \text{s.t.} \quad \mathbb{E}_{x\sim\tilde{P};\check{\pi}|x\sim\check{P}} \left[\sum_{i=1}^{n} \phi_i(x,\check{\pi}_i)\right] = \mathbb{E}_{(x,\pi)\sim\tilde{P}} \left[\sum_{i=1}^{n} \phi_i(x,\pi_i)\right] \tag{18}$$

$$\overset{(b)}{=} \max_{\check{P}(\check{\pi}|x)} \min_{\theta} \min_{\hat{P}(\hat{\pi}|x)} \mathbb{E}_{(x,\pi)\sim\tilde{P};\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}} \left[\text{loss}(\hat{\pi},\check{\pi}) + \theta^{\mathrm{T}} \left(\sum_{i=1}^{n} \phi_i(x,\check{\pi}_i) - \sum_{i=1}^{n} \phi_i(x,\pi_i)\right)\right] \tag{19}$$

$$\overset{(c)}{=} \min_{\theta} \max_{\check{P}(\check{\pi}|x)} \min_{\hat{P}(\hat{\pi}|x)} \mathbb{E}_{(x,\pi)\sim\tilde{P};\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}} \left[\text{loss}(\hat{\pi},\check{\pi}) + \theta^{\mathrm{T}} \left(\sum_{i=1}^{n} \phi_i(x,\check{\pi}_i) - \sum_{i=1}^{n} \phi_i(x,\pi_i)\right)\right] \tag{20}$$

$$\overset{(d)}{=} \min_{\theta} \mathbb{E}_{(x,\pi)\sim\tilde{P}} \max_{\check{P}(\check{\pi}|x)} \min_{\hat{P}(\hat{\pi}|x)} \mathbb{E}_{\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}} \left[\text{loss}(\hat{\pi},\check{\pi}) + \theta \cdot \sum_{i=1}^{n} \left(\phi_i(x,\check{\pi}_i) - \phi_i(x,\pi_i)\right)\right] \tag{21}$$

$$\overset{(e)}{=} \min_{\theta} \mathbb{E}_{(x,\pi)\sim\tilde{P}} \min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}} \left[\text{loss}(\hat{\pi},\check{\pi}) + \theta \cdot \sum_{i=1}^{n} \left(\phi_i(x,\check{\pi}_i) - \phi_i(x,\pi_i)\right)\right]. \tag{22}$$

The transformation steps above are described in the following:

(a) Flipping the min and max order using the minimax duality (Von Neumann & Morgenstern, 1945).

(b) Introducing the Lagrange dual variable θ.

(c) The domain of $\check{P}(\check{\pi}|x)$ is a compact convex set (i.e., permutation mixture distribution), whereas the domain of θ is convex (i.e., $\mathbb{R}^d$ where d is the number of features). The objective is concave on $\check{P}(\check{\pi}|x)$ since a non-negative linear combination of minimums of affine function is concave, while it is convex on θ. Sion's minimax theorem (Sion, 1958) says that a strong duality holds. Therefore, we can flip the order of $\check{P}(\check{\pi}|x)$ and θ in the optimization.

(d) Pushing the expectation over the empirical distribution outside the inner maximin, and changing the vector multiplication notation into a vector dot product.

(e) Applying the minimax duality (Von Neumann & Morgenstern, 1945) again to flip the optimization order of the inner minimax, resulting in Eq. (3).

Appendix B. Proofs for the Consistency Analysis

B.1 Proof of Theorem 1

Theorem 1. *Suppose loss$(\pi,\bar{\pi}) =$ loss$(\bar{\pi},\pi)$ (symmetry) and loss$(\pi,\pi) <$ loss$(\bar{\pi},\pi)$ for all $\bar{\pi} \neq \pi$. Then the adversarial permutation loss AL_f^{perm} is Fisher consistent if f is over all measurable functions and $\Pi^\diamond$ is a singleton.*

Proof. Denote $\mathbf{p}$ as the probability mass given by the predictor player $\hat{P}(\hat{\pi}|x)$, $\mathbf{q}$ as the probability mass given by the adversary player $\check{P}(\check{\pi}|x)$, and $\mathbf{d}$ as the probability mass of the true distribution $P(\pi|x)$. So, all $\mathbf{p}$, $\mathbf{q}$, and $\mathbf{d}$ lie in the $n!$ dimensional probability simplex Δ. Let C be an $n!$-by-$n!$ matrix whose $(\pi,\bar{\pi})$-th entry is loss$(\pi,\bar{\pi})$. Let $\mathbf{f} \in \mathbb{R}^{n!}$ the vector encoding of the value of f at all permutations. The definition of f^* in Eq. (16) now becomes:

$$\mathbf{f}^* \in \operatorname*{argmin}_{\mathbf{f}} \max_{\mathbf{q}\in\Delta} \min_{\mathbf{p}\in\Delta} \left\{\mathbf{f}^\top\mathbf{q} + \mathbf{p}^\top C\mathbf{q} - \mathbf{d}^\top\mathbf{f}\right\} \tag{23}$$

$$= \operatorname*{argmin}_{\mathbf{f}} \max_{\mathbf{q}\in\Delta} \left\{\mathbf{f}^\top\mathbf{q} + \min_{\pi}(C\mathbf{q})_\pi - \mathbf{d}^\top\mathbf{f}\right\}. \tag{24}$$

Let $\Pi^\diamond = \operatorname{argmin}_\pi \mathbb{E}_{\bar{\pi}|x\sim P}[\text{loss}(\pi, \bar{\pi})]$ (or equivalently $\operatorname{argmin}_\pi (C\mathbf{d})_\pi$) contains only a singleton which we denote as $\pi^\diamond$. We are to show that $\operatorname{argmax}_\pi f^*(x, \pi)$ is a singleton, and its only element π^* is exactly $\pi^\diamond$. Since $\mathbf{f}^*$ is an optimal solution, the objective function must have a zero subgradient at $\mathbf{f}^*$. That means $\mathbf{0} = \mathbf{q}^* - \mathbf{d}$, where $\mathbf{q}^*$ is an optimal solution in Eq. (24) under $\mathbf{f}^*$. As a result:

$$\mathbf{d} \in \operatorname*{argmax}_{\mathbf{q}\in\Delta} \left\{ \mathbf{q}^\top \mathbf{f}^* + \min_\pi (C\mathbf{q})_\pi \right\}. \tag{25}$$

By the first order optimality condition of constrained convex optimization (see Eq. (4.21) of (Boyd & Vandenberghe, 2004)), this means (let $C_{:,\pi^\diamond}$ be the $\pi^\diamond$-th column of C):

$$(\mathbf{f}^* + C_{:,\pi^\diamond})^\top (\mathbf{u} - \mathbf{d}) \le 0 \quad \forall \mathbf{u} \in \Delta, \tag{26}$$

where $\mathbf{f}^* + C_{:,\pi^\diamond}$ is the gradient of the objective in Eq. (25) with respect to $\mathbf{q}$ evaluated at $\mathbf{q} = \mathbf{d}$. Here we used the definition of $\pi^\diamond$. However, this inequality can hold for some $\mathbf{d} \in \Delta \cap \mathbb{R}^{n!}_{++}$ only if $\mathbf{f}^* + C_{:,\pi^\diamond}$ is a uniform vector, i.e., $f^*_\pi + \text{loss}(\pi, \pi^\diamond)$ is a constant in π. To see this, let's assume the contrary that $\mathbf{v} \triangleq \mathbf{f}^* + C_{:,\pi^\diamond}$ is not a uniform vector, and let j be the index of its maximum element. Let $\mathbf{u}$ be a vector whose values are 1 for index j and 0 otherwise. It is clear that for any $\mathbf{d} \in \Delta \cap \mathbb{R}^{n!}_{++}$, $\mathbf{v}^\top \mathbf{u} > \mathbf{v}^\top \mathbf{d}$, and hence $(\mathbf{f}^* + C_{:,\pi^\diamond})^\top (\mathbf{u} - \mathbf{d}) > 0$.

Finally, using the assumption that $\text{loss}(\pi, \pi) < \text{loss}(\bar{\pi}, \pi)$ for all $\bar{\pi} \neq \pi$, it follows that $\pi^* = \pi^\diamond$, since $\operatorname{argmax}_\pi f^*(x, \pi) = \operatorname{argmin}_\pi (C_{:,\pi^\diamond})_\pi$. □

B.2 Proof of Theorem 2

Theorem 2. *Suppose loss$(\pi, \bar{\pi})$ = loss$(\bar{\pi}, \pi)$ (symmetry) and loss(π, π) < loss$(\bar{\pi}, \pi)$ for all $\bar{\pi} \neq \pi$. Furthermore if f is over all measurable functions, then:*

- *(a) there exists $f^* \in \mathcal{F}^*$ such that $\operatorname{argmax}_\pi f^*(x, \pi) \subseteq \Pi^\diamond$ (i.e., satisfies the Fisher consistency requirement). In fact, all elements in $\Pi^\diamond$ can be recovered by some $f^* \in \mathcal{F}^*$.*
- *(b) if $\operatorname{argmin}_\pi \sum_{\pi'\in\Pi^\diamond} \alpha_{\pi'} \text{loss}(\pi', \pi) \subseteq \Pi^\diamond$ for all $\alpha_{(\cdot)} \ge 0$; $\sum_{\pi'\in\Pi^\diamond} \alpha_{\pi'} = 1$, then $\operatorname{argmax}_\pi f^*(x, \pi) \subseteq \Pi^\diamond$ for **all** $f^* \in \mathcal{F}^*$. In this case, all $f^* \in \mathcal{F}^*$ satisfy the Fisher consistency requirement.*

Proof. Let $\Pi^\diamond$ be the set containing all of the solution of $\operatorname{argmin}_\pi (C\mathbf{d})_\pi$, i.e., $\Pi^\diamond = \{\pi^\diamond \mid (C\mathbf{d})_{\pi^\diamond} = \min_\pi (C\mathbf{d})_\pi\}$. The analyses in the proof of Theorem 1 still apply to this case, except for the Eq. (26). Denote $h(\mathbf{q}) \triangleq \mathbf{q}^\top \mathbf{f}^* + \min_\pi (C\mathbf{q})_\pi$. The sub-differential of $h(\mathbf{q})$ evaluated at $\mathbf{q} = \mathbf{d}$ is the set:

$$\partial h(\mathbf{d}) = \{\mathbf{f}^* + \mathbf{v} \mid \mathbf{v} \in \textbf{conv}\{C_{:,\pi^\diamond} \mid \pi^\diamond \in \Pi^\diamond\}\}, \tag{27}$$

where **conv** denotes the convex hull of a finite point set. By extending the first order optimality condition to the subgradient case, this means that there is a subgradient $\mathbf{g} \in \partial h(\mathbf{d})$ such that:

$$\mathbf{g}^\top (\mathbf{u} - \mathbf{d}) \le 0 \quad \forall \mathbf{u} \in \Delta. \tag{28}$$

Similar to the singleton $\Pi^\diamond$ case, this inequality can hold for some $\mathbf{d} \in \Delta \cap \mathbb{R}^{n!}_{++}$ only if $\mathbf{g}$ is a uniform vector. Based on Eq. (27), $\mathbf{g} - \mathbf{f}^*$ can be written as a convex combination of the elements in $\Pi^\diamond$, and thus:

$$\mathbf{f}^* = k\mathbf{1} - \sum_{\pi'\in\Pi^\diamond} \alpha_{\pi'} C_{:,\pi'}, \tag{29}$$

for some set of $\alpha_{(\cdot)} \ge 0$, $\sum_{\pi'\in\Pi^\diamond} \alpha_{\pi'} = 1$ and some constant k. This means that multiple solutions of $\mathbf{f}^*$ are possible. Let us denote the set of containing all solutions as $\mathcal{F}^*$. For each element $\pi^\diamond$ in $\Pi^\diamond$, we can recover a $f^*_{\pi^\diamond}$ in which the $\operatorname{argmax}_\pi f^*_{\pi^\diamond}(x, \pi)$ contains a singleton element $\pi^\diamond$ by using Eq. (29) with $\alpha_{\pi^\diamond} = 1$ and $\alpha_{\pi'\in\{\Pi^\diamond \setminus \pi^\diamond\}} = 0$. This is implied by our loss assumption that $\text{loss}(\pi, \pi) < \text{loss}(\bar{\pi}, \pi)$ for all $\bar{\pi} \neq \pi$, and hence $\operatorname{argmax}_\pi f^*_{\pi^\diamond}(x, \pi) = \operatorname{argmin}_\pi (C_{:,\pi^\diamond})_\pi$.

Furthermore, if we add another assumption on the loss function such that $\operatorname{argmin}_\pi \sum_{\pi'\in\Pi^\diamond} \alpha_{\pi'} \text{loss}(\pi', \pi) \subseteq \Pi^\diamond$ for all $\alpha_{(\cdot)} \ge 0$, $\sum_{\pi'\in\Pi^\diamond} \alpha_{\pi'} = 1$, then it follows that $\operatorname{argmax}_\pi f^*(x, \pi) \subseteq \Pi^\diamond$ for all $f^* \in \mathcal{F}^*$, since for any loss function that satisfy the assumption, $\operatorname{argmin}_\pi \left(\sum_{\pi'\in\Pi^\diamond} a_{\pi'} C_{:,\pi'}\right)_\pi \subseteq \Pi^\diamond$ for all $\alpha_{(\cdot)} \ge 0$, $\sum_{\pi'\in\Pi^\diamond} \alpha_{\pi'} = 1$. □

Appendix C. Feature Representation

We explain local binary pattern (LBP), color histogram (RGB), and optical flow within our feature representation in more details in this section. LBP is one of the best and most widely used texture descriptors in different applications like face detection. It assigns a label to every pixel of an image by thresholding the 3×3 neighborhood of each pixel with the center pixel value and reporting a binary number as the result. It is discriminative and invariant to monotonic gray-level changes (Ahonen et al., 2006). An important element in content-based image retrieval is the image color. Global histogram is one of the most popular color information representations. It presents the joint distribution of intensities of three-color (Red, Green, and Blue) channels. Its robustness to background complications and object distortion provides helpful hints for the subsequent expression of similarity between images (Wang et al., 2010).

To extract LBP and color of histograms features, we first divide the object regions to 7×3 blocks based on the aspect ratio of the detected pedestrians in the dataset, which is also 7:3. For each block, we calculate the distribution of LBP and then employ the Bhattacharyya coefficient to compute the affinity of a pair of distributions. Bhattacharyya coefficient (BC) measures the amount of overlap between two distributions. It returns 21 features for LBP. For color histogram, we represent the color information of each block using a 3D RGB color histogram of $8\times8\times8$ dimension. Then BC is applied and 21 RGB features are extracted.

Optical flow is an image motion representation and defined as the projection of velocities of 2D/3D surface points. It is based on correspondences between image features, correlations, or properties of intensity structures (Beauchemin & Barron, 1995). We compute the histogram of optical flow (HOF) for every detected box and employ BC to calculate the motion distribution relation. It returns one feature as optical flow.

We also consider three binary features (entering, leaving, and staying invisible) to indicate the status of each object between two consecutive frames.

Together, each feature vector, $\phi_i(x, j)$, has 48 values.