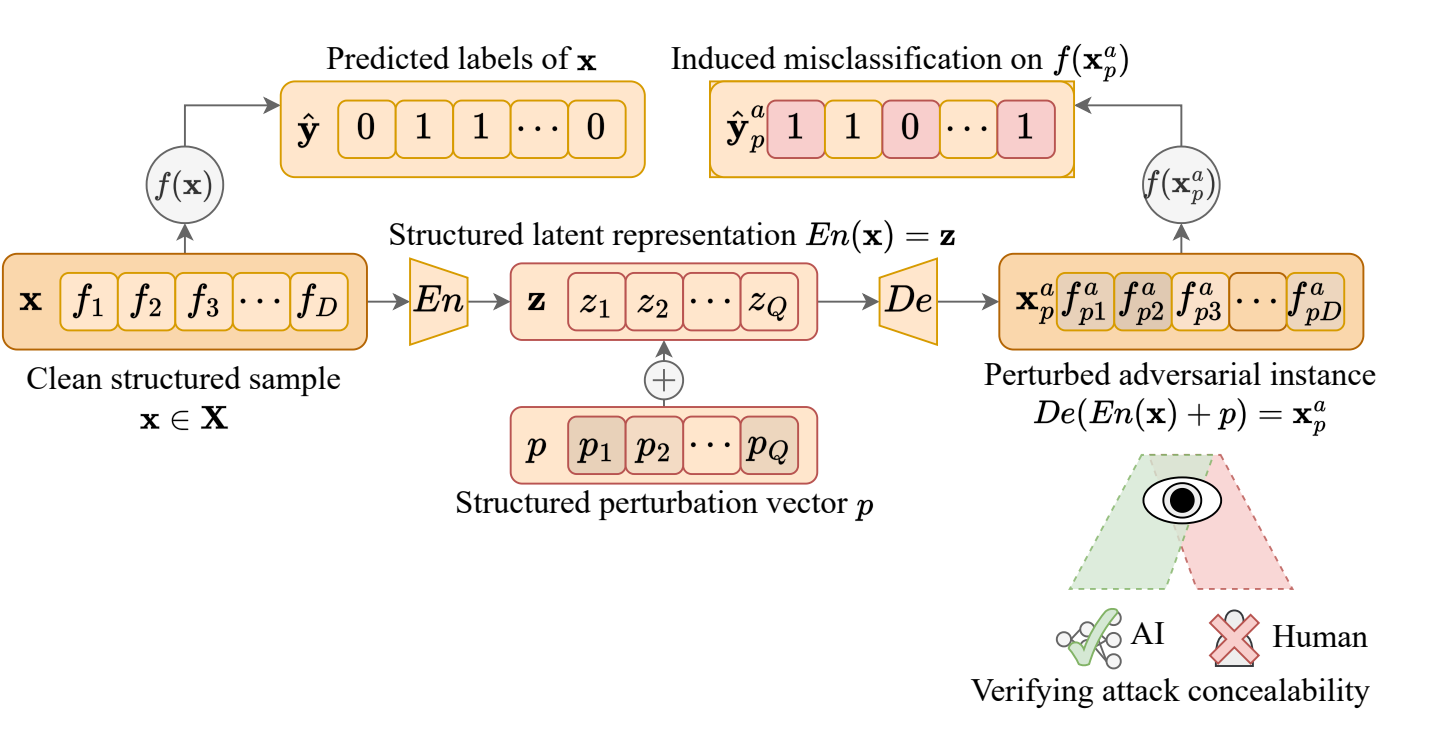
Predicted labels of $\mathbf{x}$
$\hat{\mathbf{y}}$ 0 1 1 $\cdots$ 0
Induced misclassification on $f(\mathbf{x}_p^a)$
$\hat{\mathbf{y}}_p^a$ 1 1 0 $\cdots$ 1
$f(\mathbf{x})$
$f(\mathbf{x}_p^a)$
Structured latent representation $En(\mathbf{x}) = \mathbf{z}$
$\mathbf{x}$ $f_1$ $f_2$ $f_3$ $\cdots$ $f_D$
$En$
$\mathbf{z}$ $z_1$ $z_2$ $\cdots$ $z_Q$
$De$
$\mathbf{x}_p^a$ $f_{p1}^a$ $f_{p2}^a$ $f_{p3}^a$ $\cdots$ $f_{pD}^a$
Clean structured sample
$\mathbf{x} \in \mathbf{X}$
$p$ $p_1$ $p_2$ $\cdots$ $p_Q$
Structured perturbation vector $p$
Perturbed adversarial instance
$De(En(\mathbf{x}) + p) = \mathbf{x}_p^a$
AI
Human
Verifying attack concealability