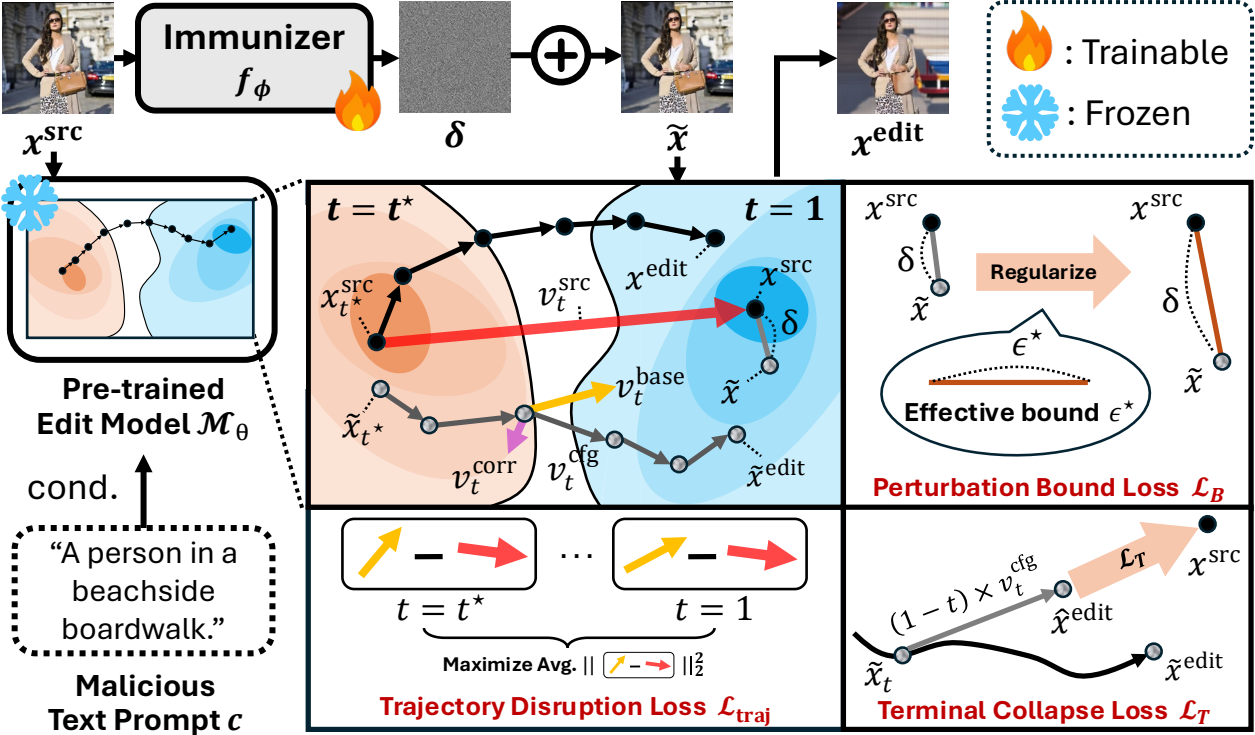
Immunizer
f_ϕ
x^{src}
δ
$\tilde{x}$
x^{edit}
: Trainable
: Frozen
Pre-trained Edit Model $\mathcal{M}_\theta$
cond.
"A person in a beachside boardwalk."
Malicious Text Prompt c
$t = t^\star$
$t = 1$
$x_{t^\star}^{\text{src}}$
v_t^{src}
x^{edit}
x^{src}
δ
$\tilde{x}$
v_t^{base}
$\tilde{x}_{t^\star}$
v_t^{corr}
v_t^{cfg}
$\tilde{x}^{\text{edit}}$
Regularize
$\epsilon^\star$
Effective bound $\epsilon^\star$
Perturbation Bound Loss $\mathcal{L}_B$
$t = t^\star$
$t = 1$
Maximize Avg. $\|\cdot\|_2^2$
Trajectory Disruption Loss $\mathcal{L}_{\text{traj}}$
$(1-t) \times v_t^{\text{cfg}}$
$\mathcal{L}_T$
$\hat{x}^{\text{edit}}$
$\tilde{x}_t$
Terminal Collapse Loss $\mathcal{L}_T$