

Exposed Parts
(# in SD1.4)

sexual

self-harm

violence

harassment

hate

Total

SteerDiff (Ours)

ESD

Methods

-75

-50

-25

0

% Change

-100

-50

0

% Change

-75

-50

-25

0

% Change

-75

-50

-25

0

% Change

-75

-50

-25

0

% Change

-75

-50

-25

0

% Change