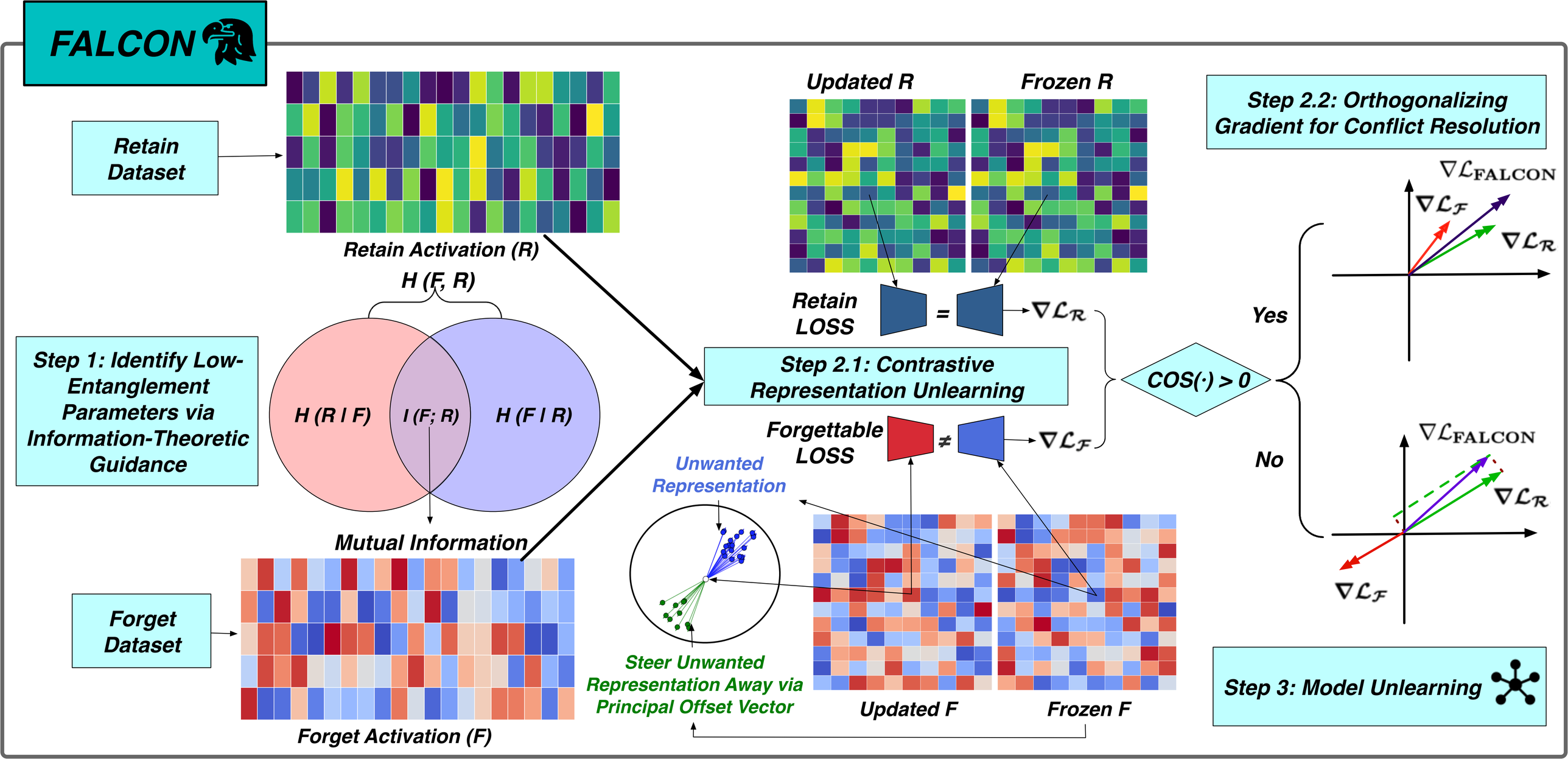
FALCON
Retain Dataset
Retain Activation (R)
Updated R
Frozen R
Retain LOSS
=
$\nabla\mathcal{L}_{\mathcal{R}}$
Step 2.2: Orthogonalizing Gradient for Conflict Resolution
$\nabla\mathcal{L}_{\text{FALCON}}$
$\nabla\mathcal{L}_{\mathcal{F}}$
$\nabla\mathcal{L}_{\mathcal{R}}$
Yes
No
COS(·) > 0
Step 1: Identify Low-Entanglement Parameters via Information-Theoretic Guidance
H (F, R)
H (R | F)
I (F; R)
H (F | R)
Mutual Information
Step 2.1: Contrastive Representation Unlearning
Forgettable LOSS
≠
Unwanted Representation
Steer Unwanted Representation Away via Principal Offset Vector
Forget Dataset
Forget Activation (F)
Updated F
Frozen F
Step 3: Model Unlearning