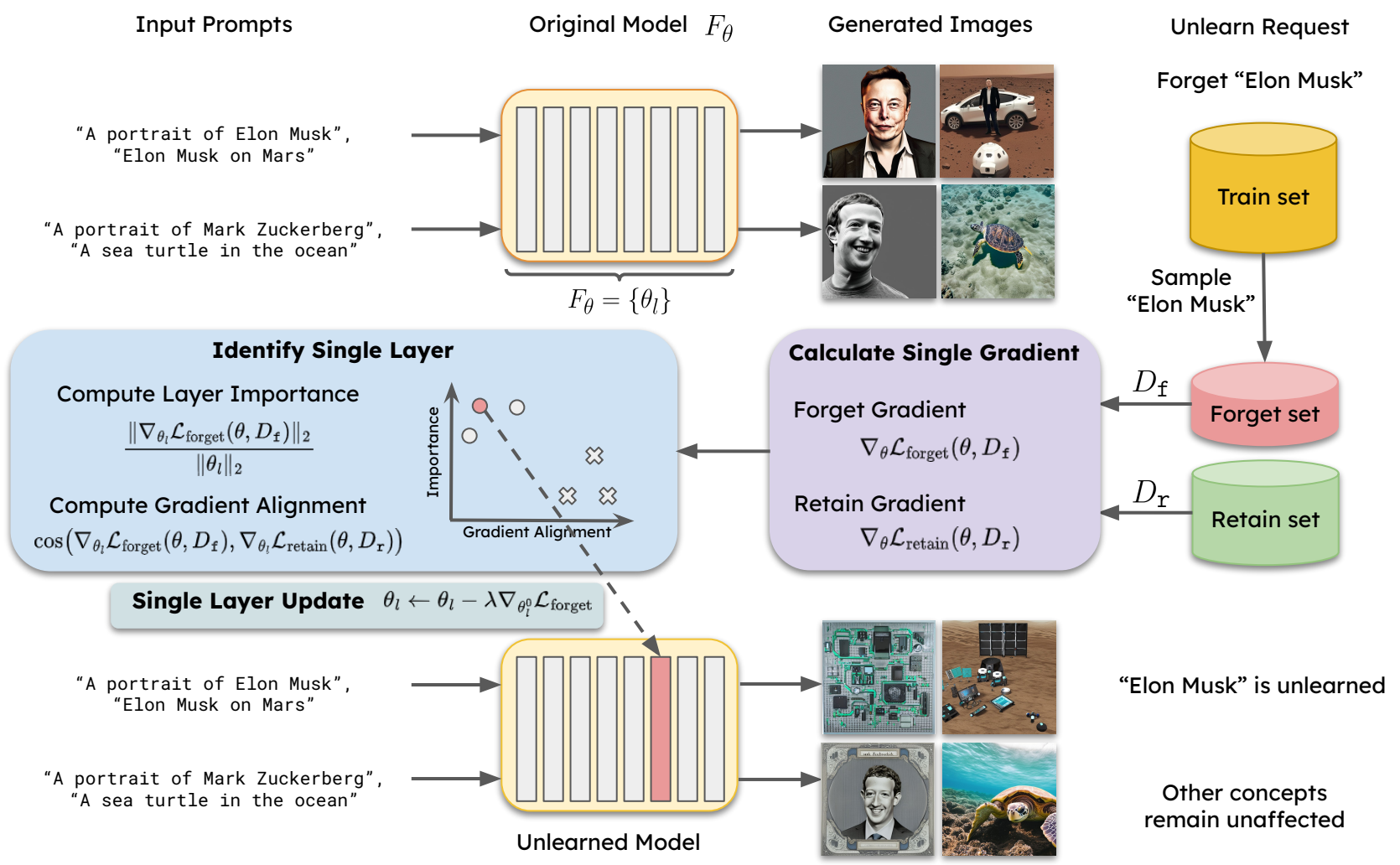
Input Prompts
Original Model F_θ
Generated Images
Unlearn Request
"A portrait of Elon Musk",
"Elon Musk on Mars"
"A portrait of Mark Zuckerberg",
"A sea turtle in the ocean"
$F_\theta = \{\theta_l\}$
Forget "Elon Musk"
Train set
Sample
"Elon Musk"
Forget set
Retain set
$D_\mathtt{f}$
$D_\mathtt{r}$
Identify Single Layer
Compute Layer Importance
$\frac{\|\nabla_{\theta_l}\mathcal{L}_{\text{forget}}(\theta, D_\mathtt{f})\|_2}{\|\theta_l\|_2}$
Compute Gradient Alignment
$\cos\left(\nabla_{\theta_l}\mathcal{L}_{\text{forget}}(\theta, D_\mathtt{f}), \nabla_{\theta_l}\mathcal{L}_{\text{retain}}(\theta, D_\mathtt{r})\right)$
Importance
Gradient Alignment
Calculate Single Gradient
Forget Gradient
$\nabla_\theta\mathcal{L}_{\text{forget}}(\theta, D_\mathtt{f})$
Retain Gradient
$\nabla_\theta\mathcal{L}_{\text{retain}}(\theta, D_\mathtt{r})$
Single Layer Update $\theta_l \leftarrow \theta_l - \lambda\nabla_{\theta_l^0}\mathcal{L}_{\text{forget}}$
"A portrait of Elon Musk",
"Elon Musk on Mars"
"A portrait of Mark Zuckerberg",
"A sea turtle in the ocean"
Unlearned Model
"Elon Musk" is unlearned
Other concepts
remain unaffected