Probe-Me-Not: Protecting Pre-trained Encoders from Malicious Probing

Ruyi Ding, Tong Zhou, Lili Su, Aidong Adam Ding, Xiaolin Xu, Yunsi Fei

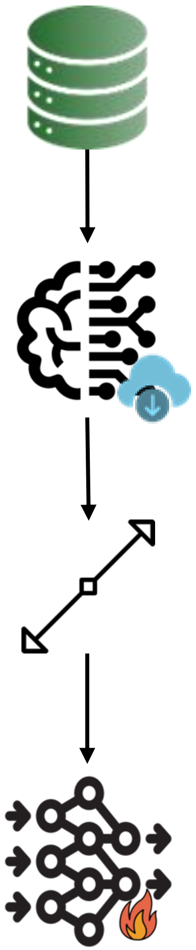
Northeastern University

Acknowledgement:

This work was supported in part by National Science Foundation under grants CNS-2212010, SaTC-1929300, IUCRC-1916762, CNS-2239672, CNS-2326597, and CCF-2340482.

Train DNNs with Pre-trained Models

Low data requirement



User's data

Encoder API

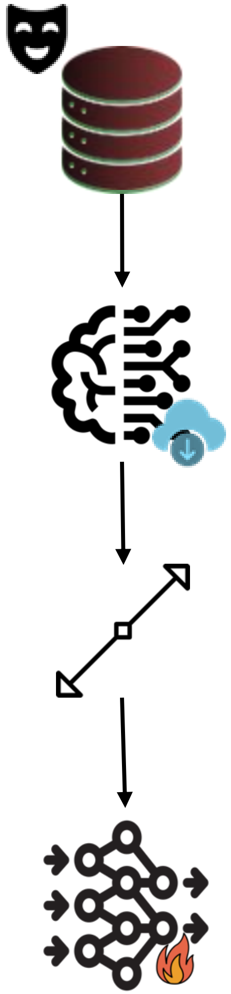
Embeddings

Downstream head

Low training cost

Train DNNs with Pre-trained Models

Easy to be misused



Unauthorized data

Encoder API

clarifai

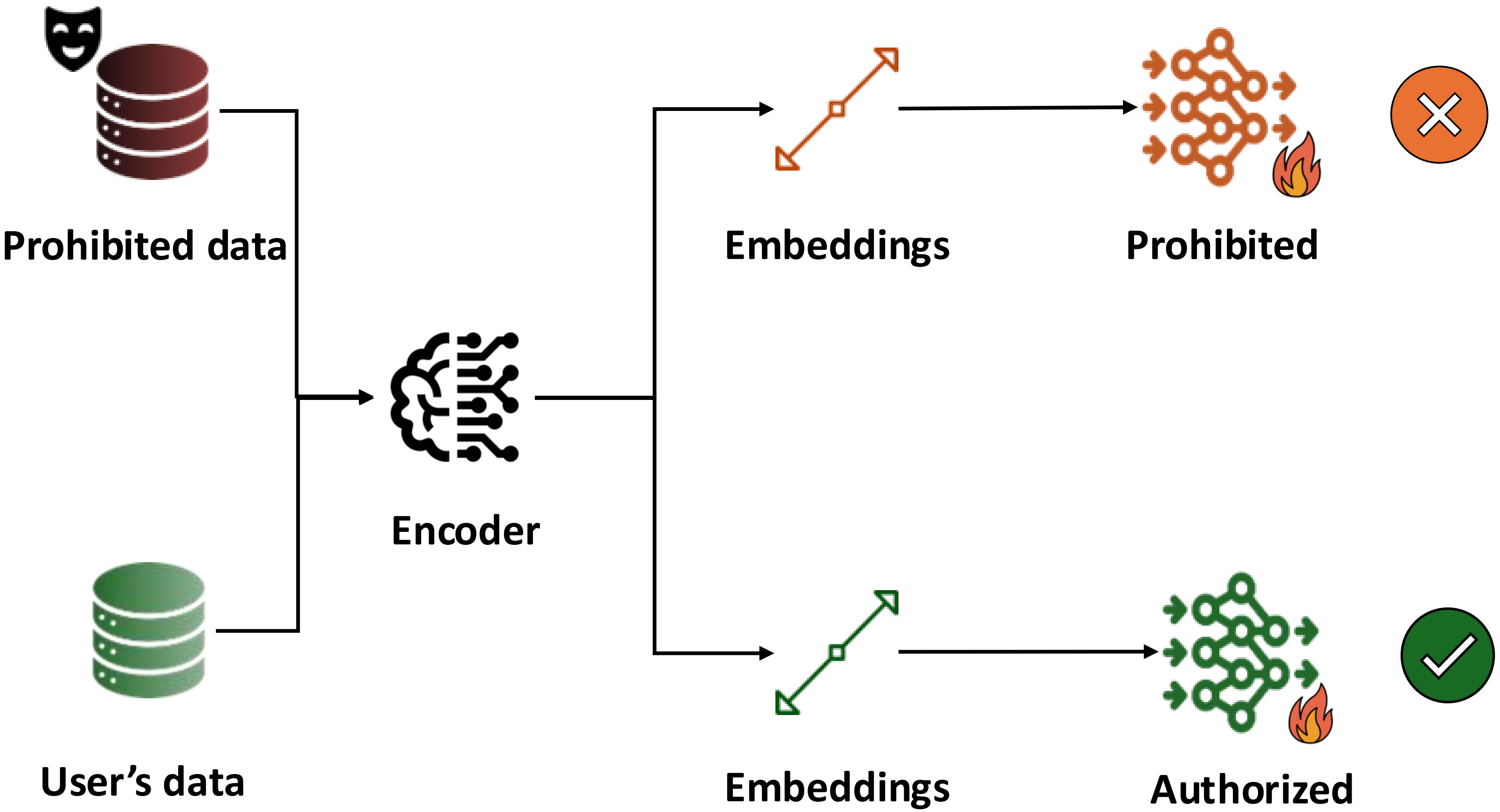
Embeddings

Low cost of unauthorized transfer

Downstream head

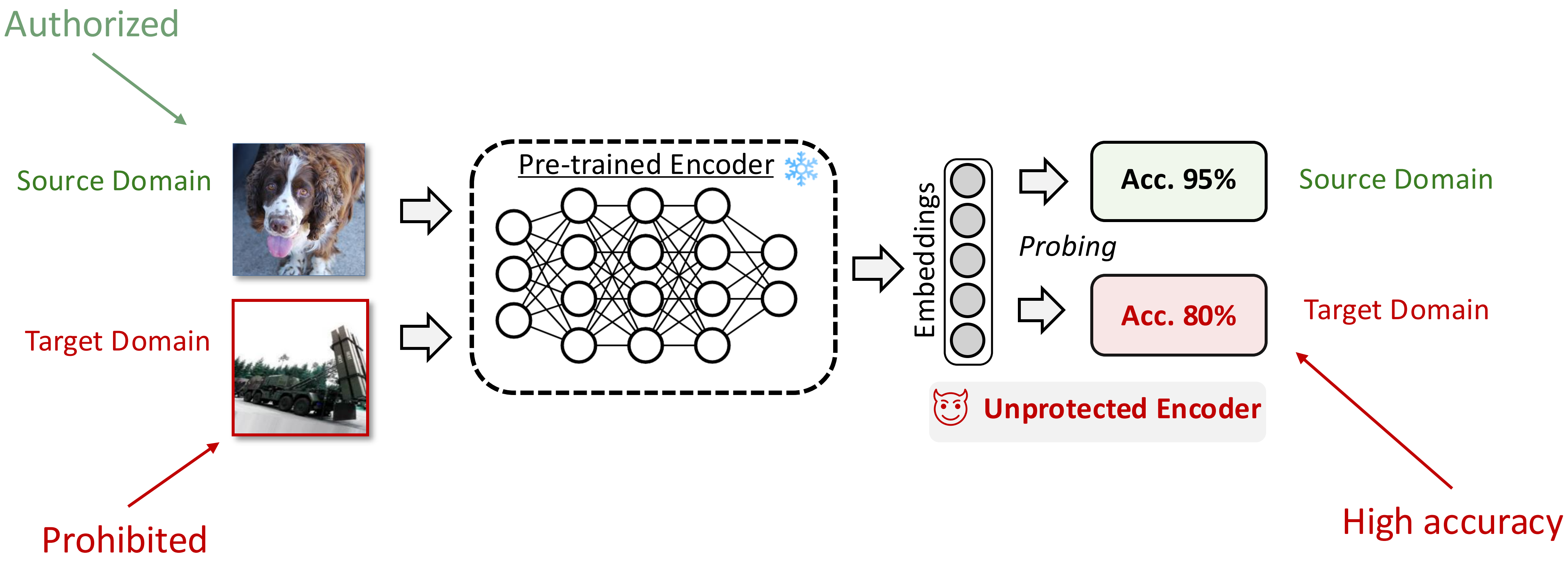
Applicability Authorization

- Prevent **pre-trained models** from being misused by **proactively restricting their transferability** for harmful tasks.



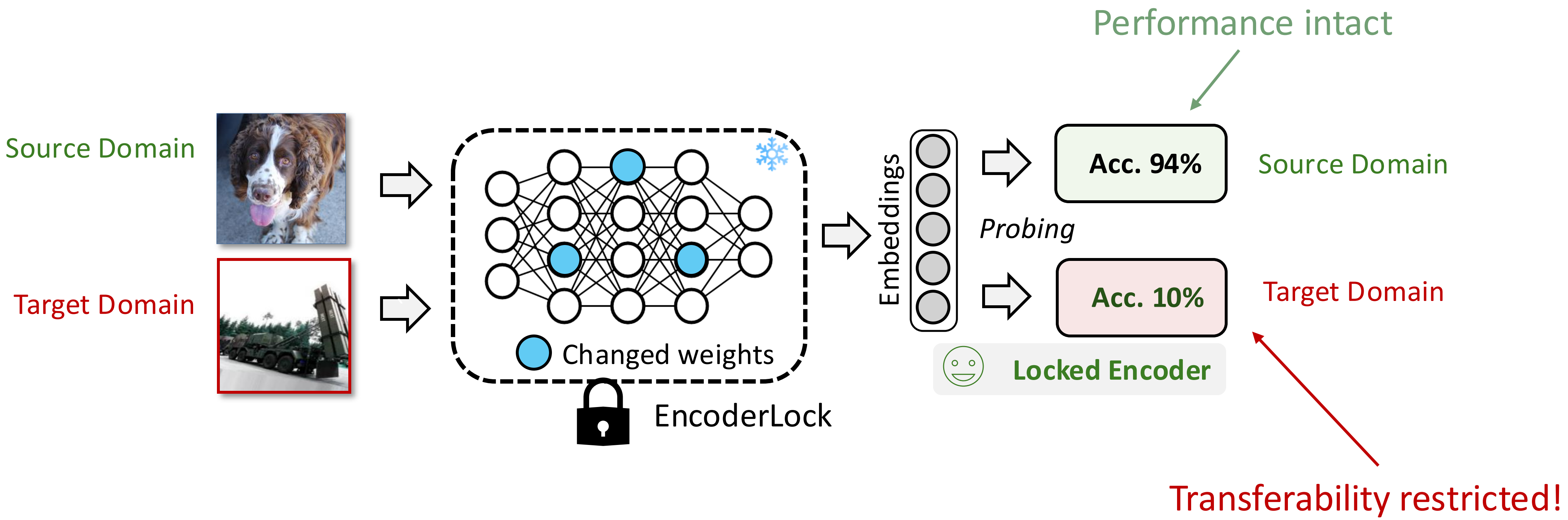
Malicious Probing

- Unauthorized transfer learning on pre-trained encoders

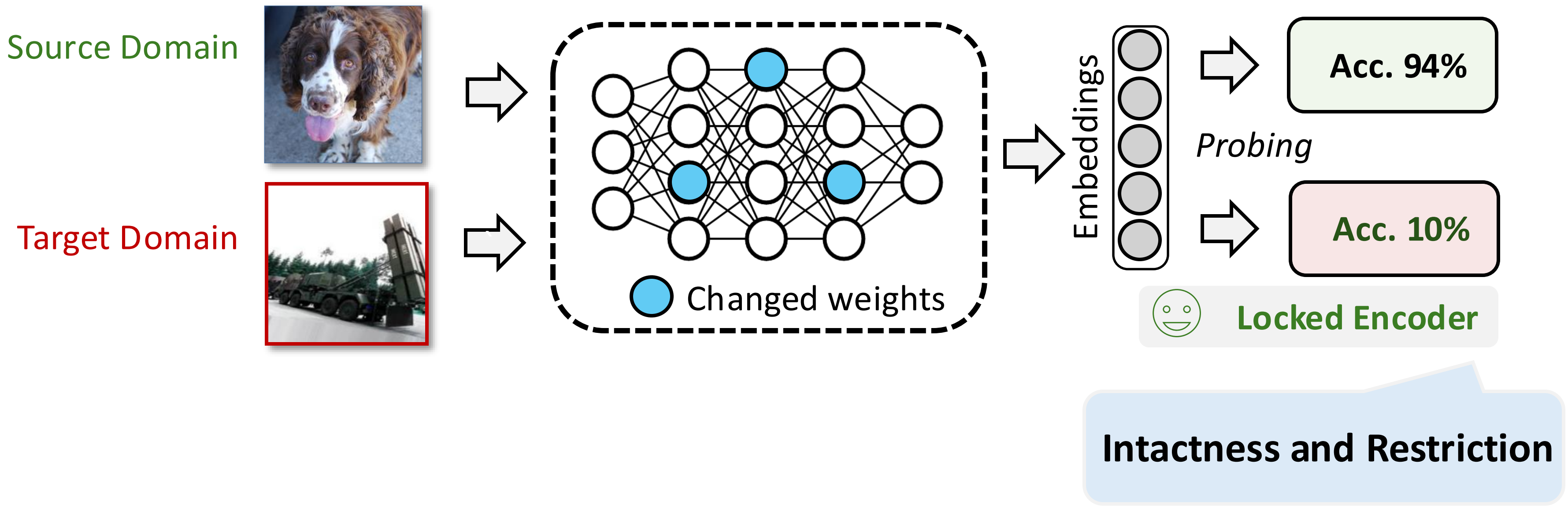


EncoderLock for Applicability Authorization

- Protecting the model's applicability

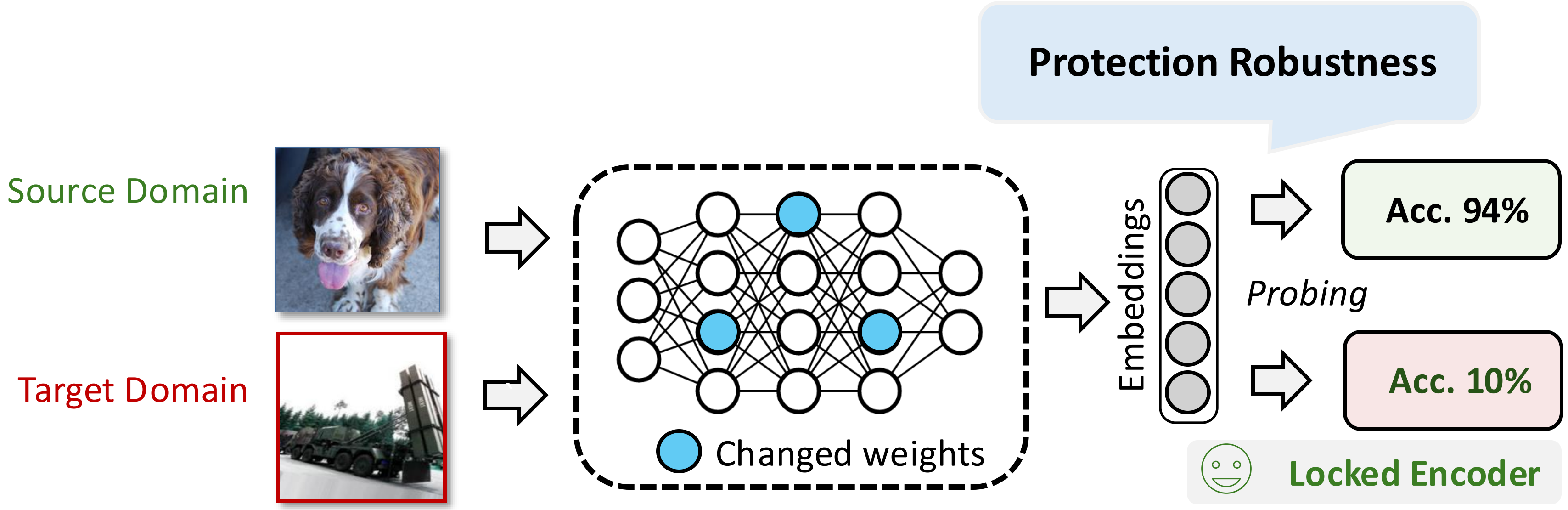


Protection Objectives



Protection Objectives

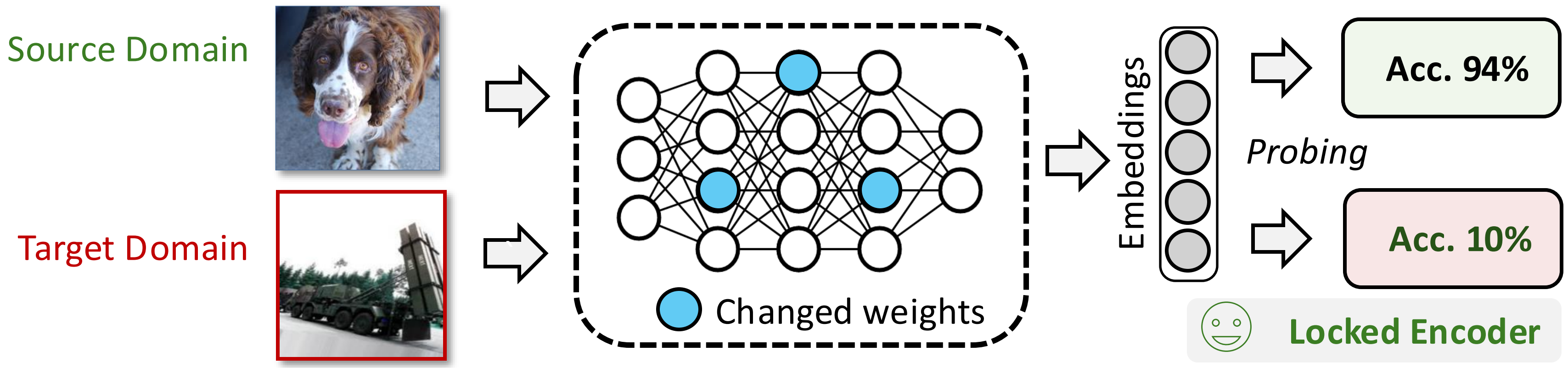
Intactness and Restriction



Protection Objectives

Intactness and Restriction

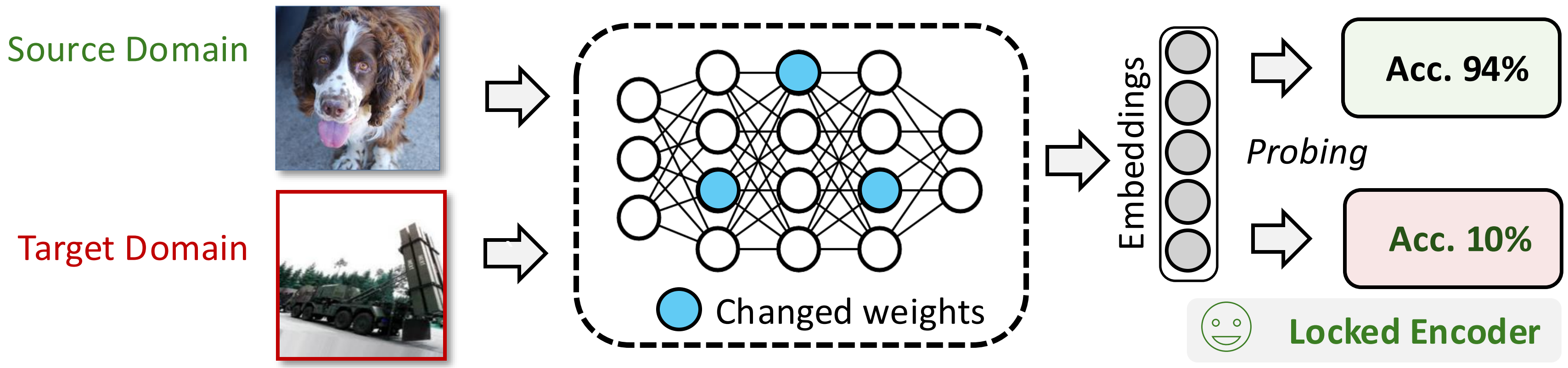
Protection Robustness



Protection Objectives

Intactness and Restriction

Protection Robustness



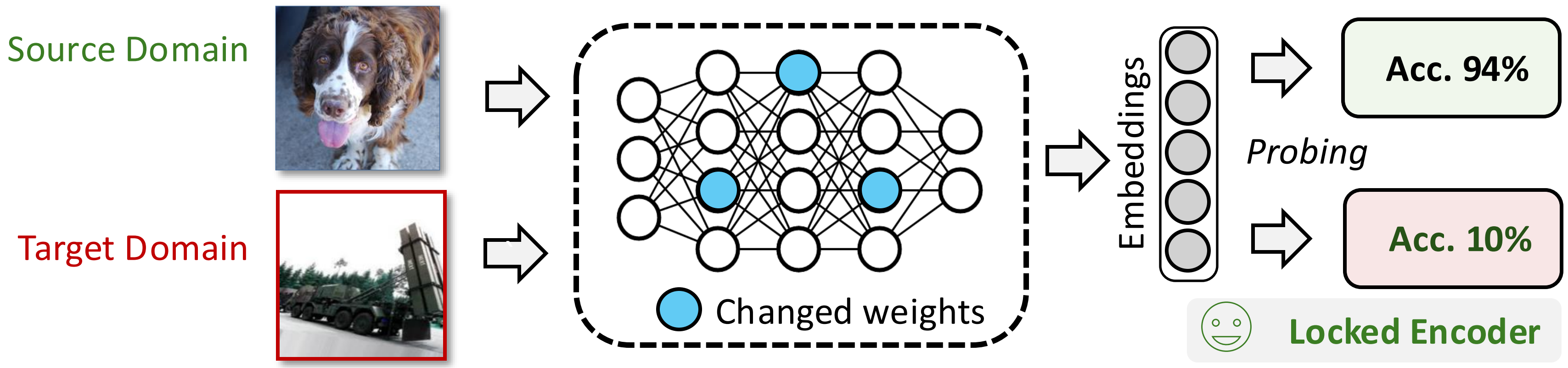
Prohibited Data Accessibility

Protection Objectives

Intactness and Restriction

Protection Robustness

Prohibited Data Accessibility

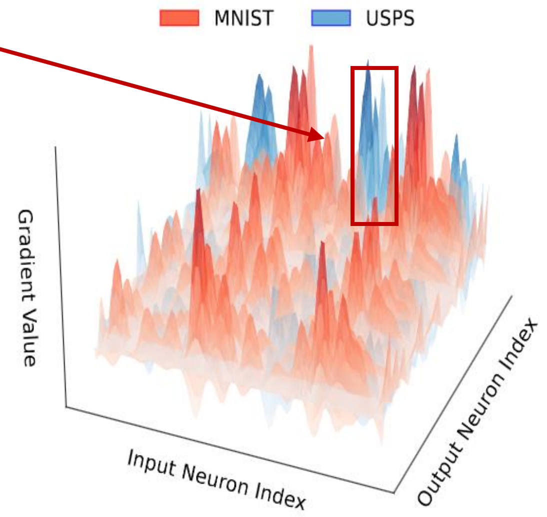


Domain-aware Weight Optimization

O1: Intactness and Restriction

- What to optimize:
  - critical to the target domain
  - not important to the source domain

Weight importance—measured by gradient



Domain-aware Weight Optimization

O1: Intactness and Restriction

- What to optimize:
  - critical to the target domain
  - not important to the source domain

- How to optimize:
  - EncoderLock loss

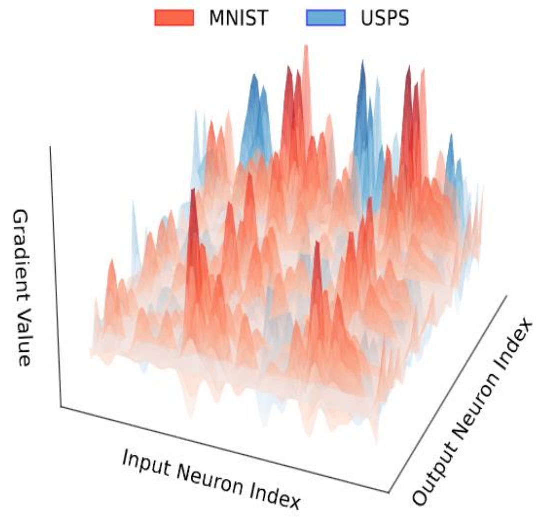
For optimization continuity

$$L_{el} = L_S + R_T, \text{ where } R_T = \log(1 + \alpha \frac{L_S}{L_T})$$

Source domain loss

Target domain loss

Weight importance—measured by gradient

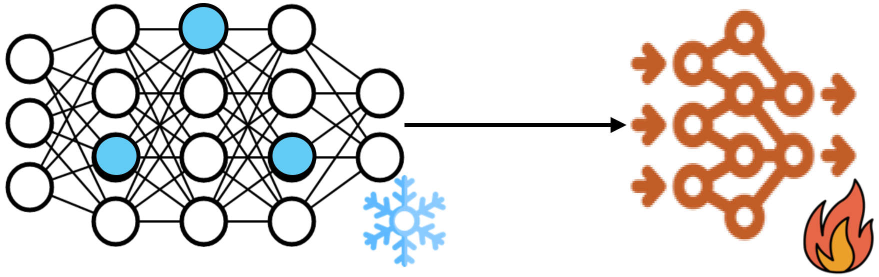


Self-Challenging Training Scheme

O2: Protection Robustness

$$L_{el} = L_{\mathcal{S}} + R_{\mathcal{T}}, \text{ where } R_{\mathcal{T}} = \log(1 + \alpha \frac{L_{\mathcal{S}}}{L_{\mathcal{T}}})$$

- How to ensure EncoderLock's protection on different downstream heads?

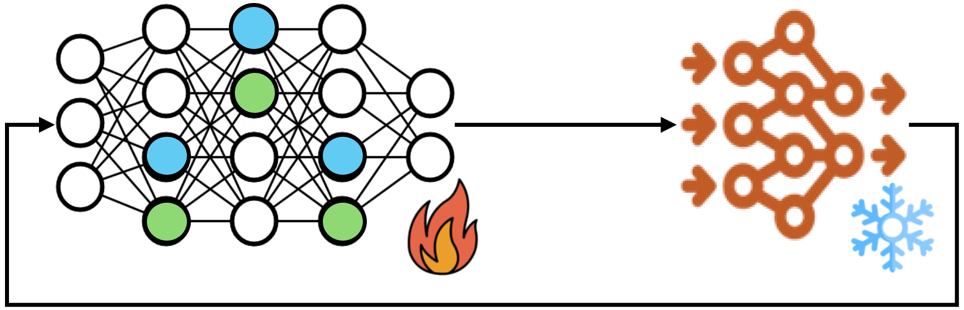


Self-Challenging Training Scheme

O2: Protection Robustness

$$L_{el} = L_{\mathcal{S}} + R_{\mathcal{T}}, \text{ where } R_{\mathcal{T}} = \log(1 + \alpha \frac{L_{\mathcal{S}}}{L_{\mathcal{T}}})$$

- How to ensure EncoderLock's protection on different downstream heads?

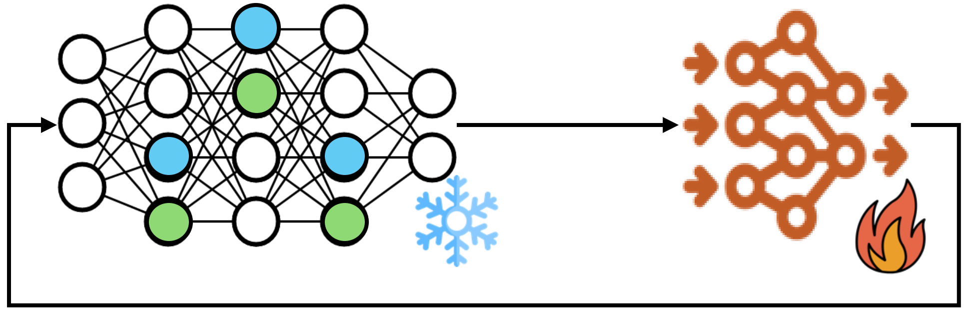


Self-Challenging Training Scheme

O2: Protection Robustness

$$L_{el} = L_{\mathcal{S}} + R_{\mathcal{T}}, \text{ where } R_{\mathcal{T}} = \log(1 + \alpha \frac{L_{\mathcal{S}}}{L_{\mathcal{T}}})$$

- How to ensure EncoderLock's protection on different downstream heads?



$\phi \sim encoder$;
$\theta_T \sim downstream\ head\ T$;
$\theta_S \sim downstream\ head\ S$;
$M \sim \max \#\ modified\ weights$;

target head

$$\phi^* = \arg\min_{\phi} \max_{\theta_{\mathcal{T}}} L_{el}(\phi, \theta_{\mathcal{S}}, \theta_{\mathcal{T}}) \quad \text{s.t.} \quad \|\phi^* - \phi\|_0 \leq M$$

Encoder

Adapting Learning Methods to Data Accessibility

O3: Prohibited Data Accessibility

$$L_{el} = L_S + R_T, \text{ where } R_T = \log(1 + \alpha \frac{L_S}{L_T})$$

Supervised EncoderLock

Data + Label from prohibited domain

- Supervised EncoderLock: cross entropy loss

$$L_S \triangleq l_{ce}(x_S, y_S) \qquad L_T \triangleq l_{ce}(x_T, y_T)$$

Adapting Learning Methods to Data Accessibility

O3: Prohibited Data Accessibility

$$L_{el} = L_S + R_T, \text{ where } R_T = \log(1 + \alpha \frac{L_S}{L_T})$$

Supervised EncoderLock

Data + Label from prohibited domain

- Supervised EncoderLock: cross entropy loss

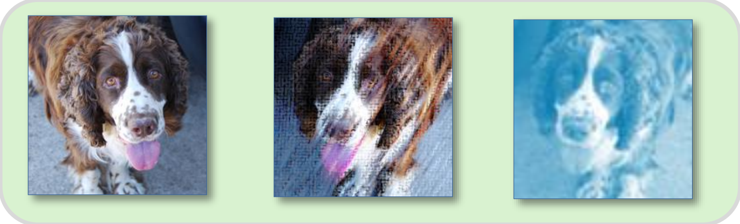
$$L_S \triangleq l_{ce}(x_S, y_S) \qquad L_T \triangleq l_{ce}(x_T, y_T)$$

- Unsupervised EncoderLock:

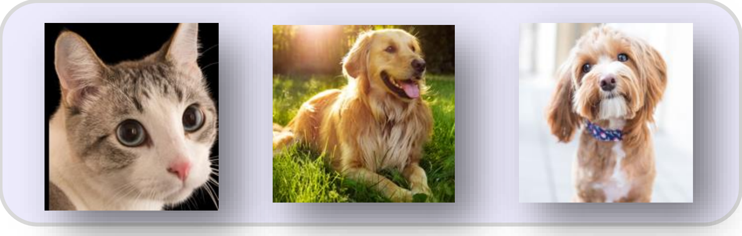
Unsupervised EncoderLock

Only Data from prohibited domain

Sample i



Positive pairs with i



Negative pairs with i

Adapting Learning Methods to Data Accessibility

O3: Prohibited Data Accessibility

$$L_{el} = L_{\mathcal{S}} + R_{\mathcal{T}}, \text{ where } R_{\mathcal{T}} = \log(1 + \alpha \frac{L_{\mathcal{S}}}{L_{\mathcal{T}}})$$

- Supervised EncoderLock: cross entropy loss

$$L_S \triangleq l_{ce}(x_S, y_S) \qquad L_T \triangleq l_{ce}(x_T, y_T)$$

- Unsupervised EncoderLock: contrastive loss

$$L^{\text{cont}} := -\frac{1}{N_B}\sum_{i=1}^{N_B} \log(\frac{\text{sim}(z_i, \tilde{z}_i)}{\sum_{j=1}^{N_B} \text{sim}(z_i, \tilde{z}_j)})$$

Supervised EncoderLock

Data + Label from prohibited domain

Unsupervised EncoderLock

Only Data from prohibited domain

Adapting Learning Methods to Data Accessibility

O3: Prohibited Data Accessibility

$$L_{el} = L_S + R_T, \text{ where } R_T = \log(1 + \alpha \frac{L_S}{L_T})$$

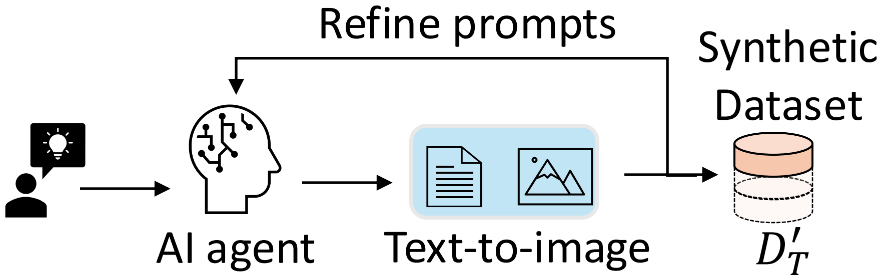
- **Supervised EncoderLock:** cross entropy loss

$$L_S \triangleq l_{ce}(x_S, y_S) \qquad L_T \triangleq l_{ce}(x_T, y_T)$$

- **Unsupervised EncoderLock:** contrastive loss

$$L^{\text{cont}} := -\frac{1}{N_B} \sum_{i=1}^{N_B} \log\left(\frac{\text{sim}(z_i, \tilde{z}_i)}{\sum_{j=1}^{N_B} \text{sim}(z_i, \tilde{z}_j)}\right)$$

- **Zero-shot EncoderLock:** no data, no label



Supervised EncoderLock

Data + Label from prohibited domain

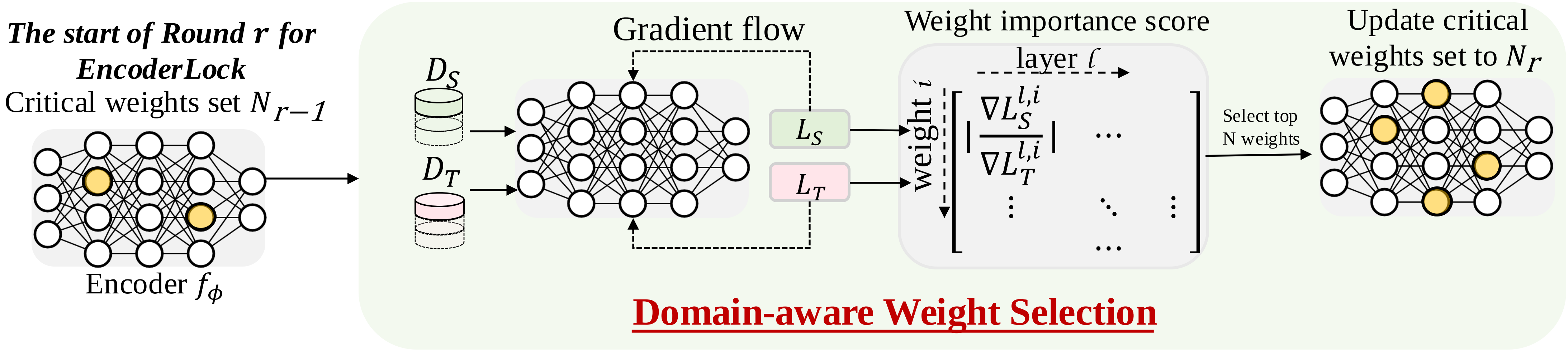
Unsupervised EncoderLock

Only Data from prohibited domain

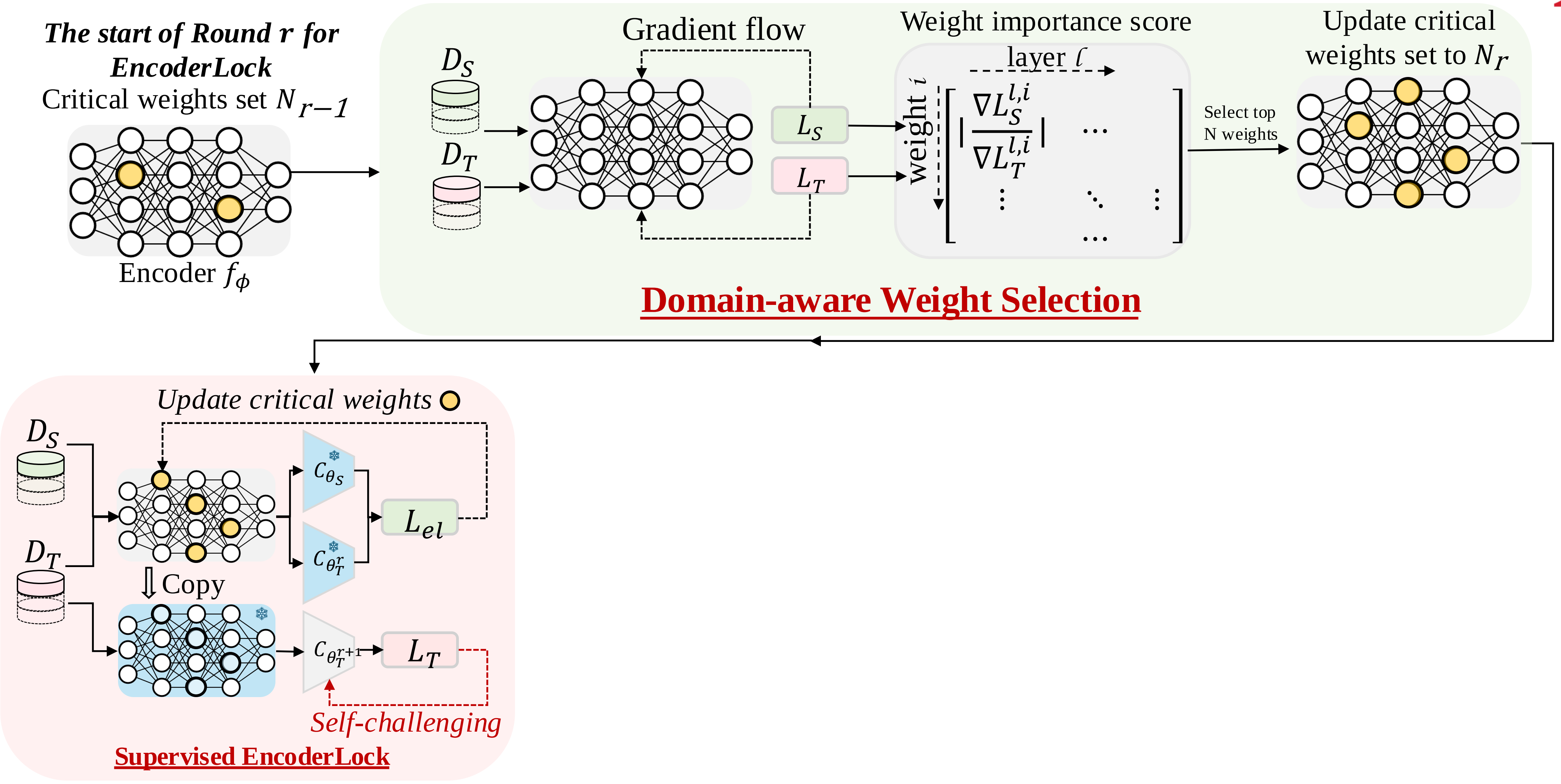
Zero-shot EncoderLock

NO data but descriptions of prohibited domain

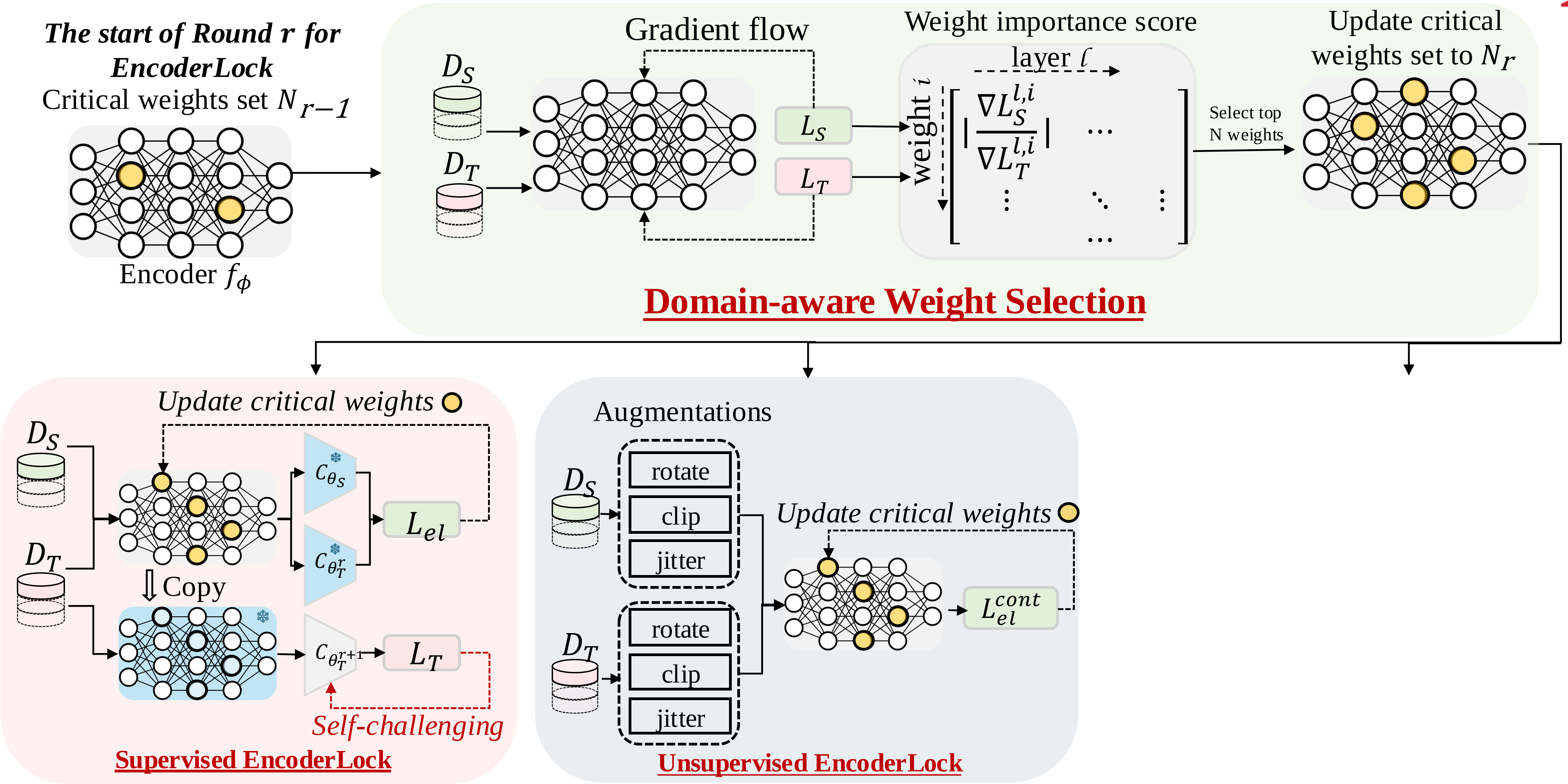
EncoderLock: Summary



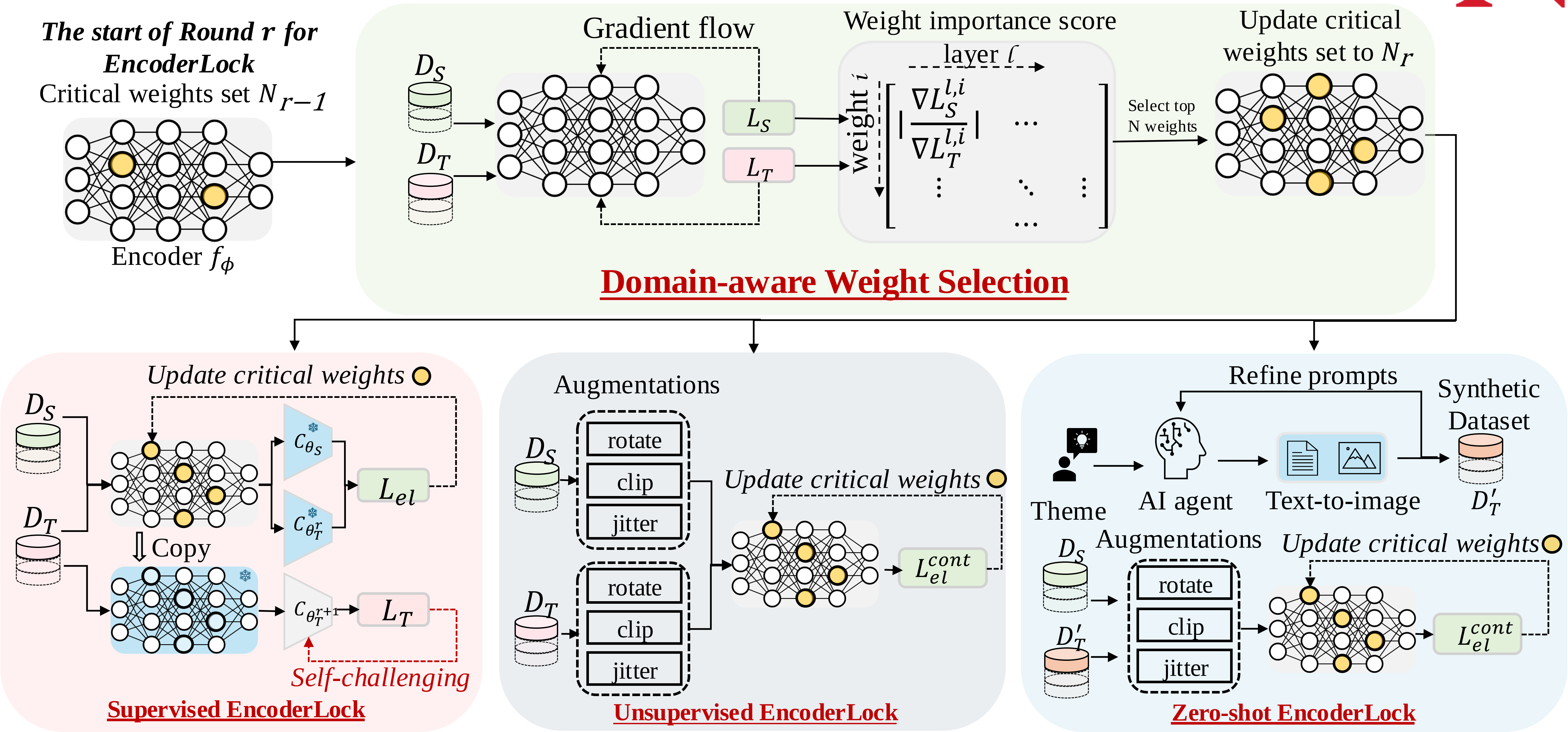
EncoderLock: Summary



EncoderLock: Summary

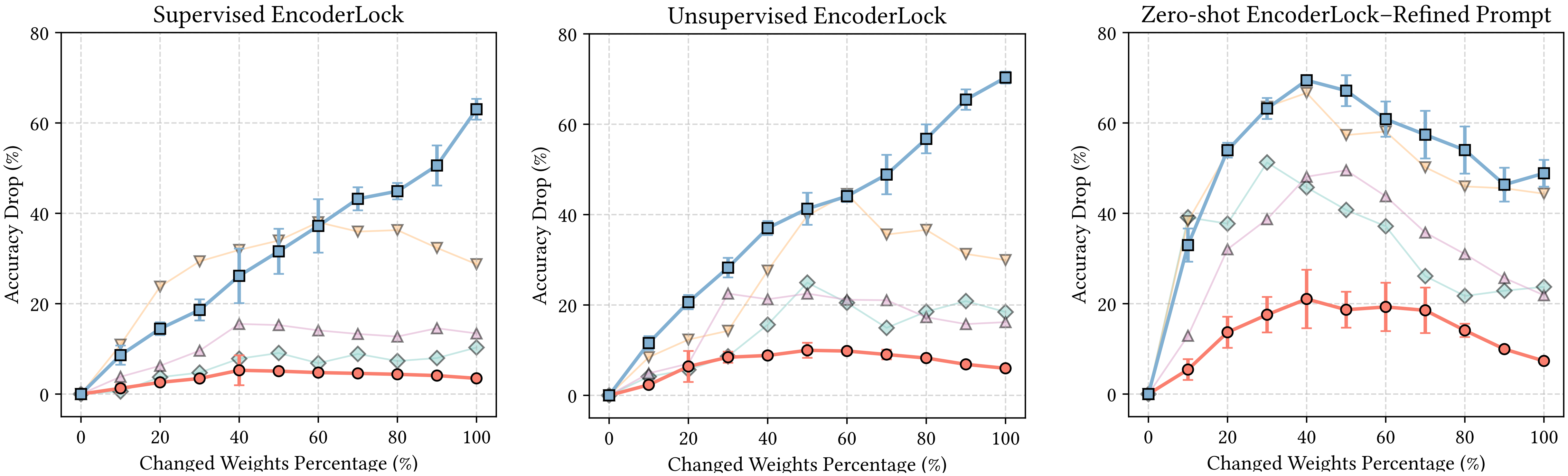


EncoderLock: Summary



EncoderLock Evaluation: Accuracy Drop

- Red: drop on source (imagenette)
- Blue: drop on target (military vehicle)
- Other: military weapon, ordinary vehicle, animal

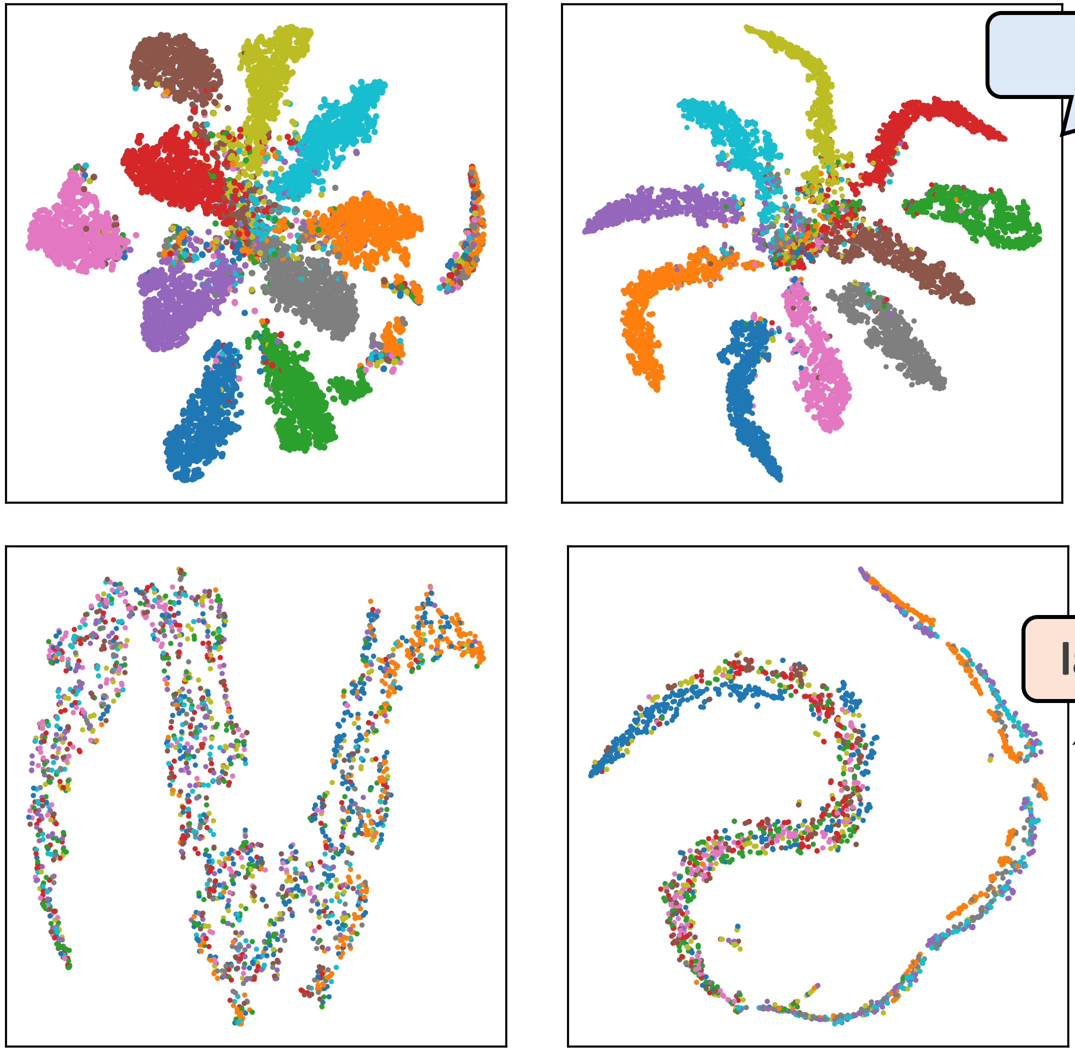


EncoderLock Evaluation: Latent Space

Supervised **Unsupervised**

Source

Target



well-separated

latent space collapse

Visualization the Focus on Encoder

Tank (Prohibited)

Unprotected

Focus on the tank gun for prohibited domain

Out of focus

Supervised

Unsupervised

Zero-shot

Thank you for listening!

Q&A

Ruyi Ding

Northeastern University

ding.ruy@northeastern.edu

rollinding.github.io