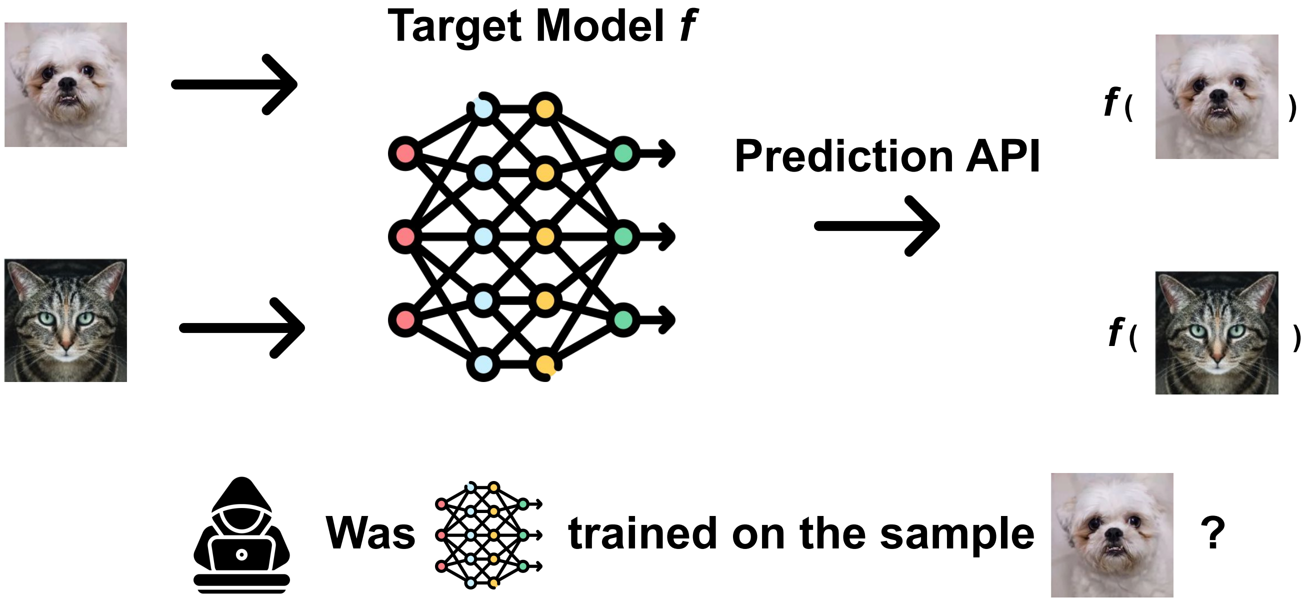
UMassAmherst | Manning College of Information & Computer Sciences

DIFFENCE: Fencing Membership Privacy With Diffusion Models

Artifact Evaluated
NDSS SYMPOSIUM
Available
Functional
Reproduced

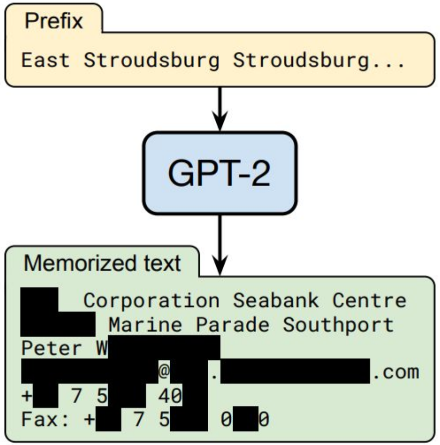
Yuefeng Peng, Ali Naseh, Amir Houmansadr

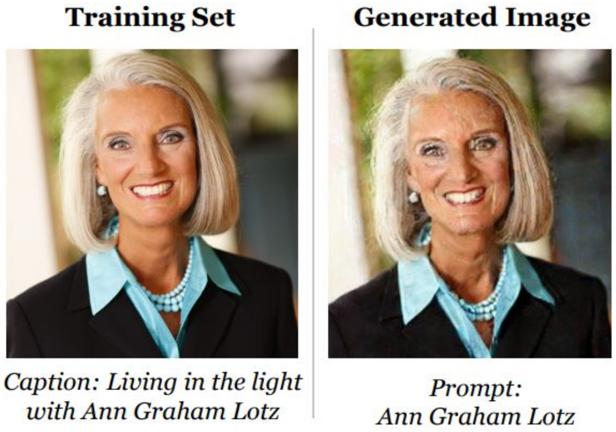
Membership Inference Attacks (MIAs)



Why Do MIAs Matter?

- **Privacy leakage**
- **Stepping stone for stronger attacks[1][2]**
- **Privacy auditing**
- **…**



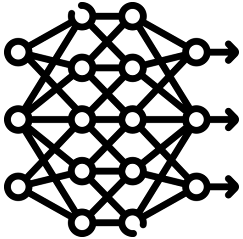


MIA-based Data Extraction Attacks

[1] Carlini et al. Extracting training data from large language models, USENIX Security 2021.
[2] Carlini et al. Extracting Training Data from Diffusion Models,, USENIX Security 2023.

Review: ML Pipelines

Initialized Model

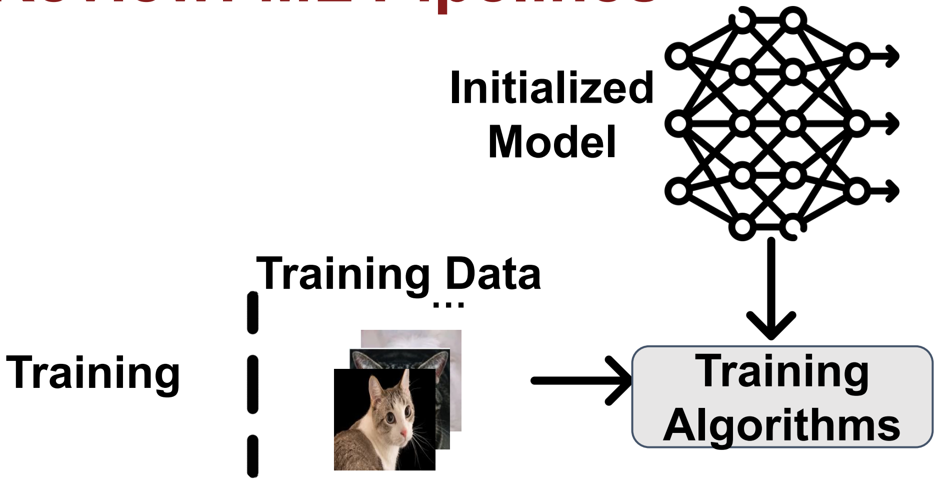


Training Data

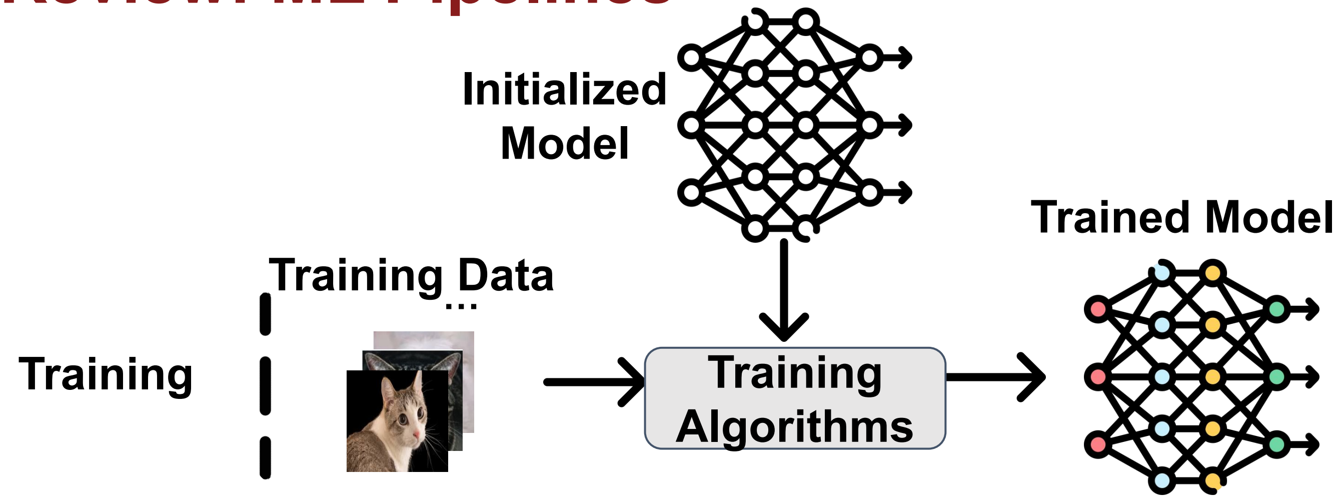
...

Training

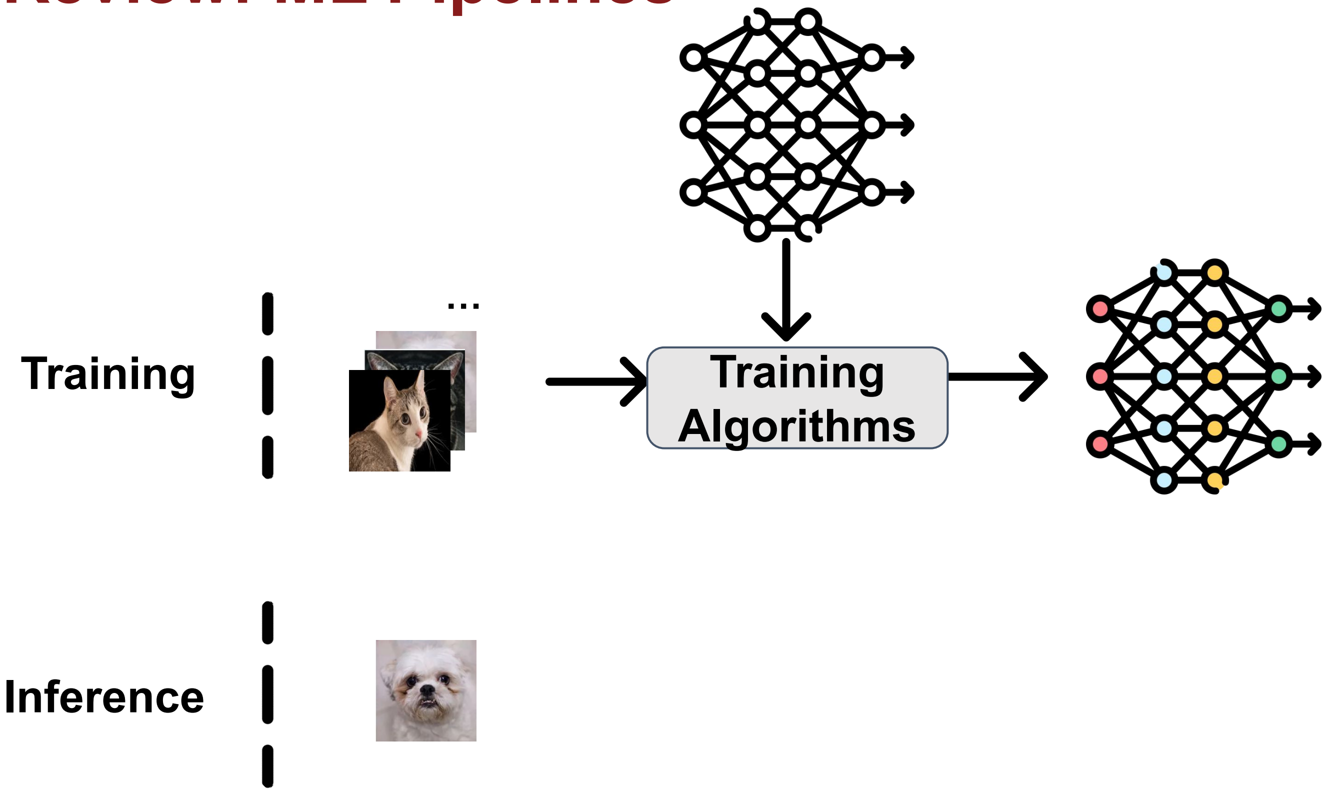
Review: ML Pipelines



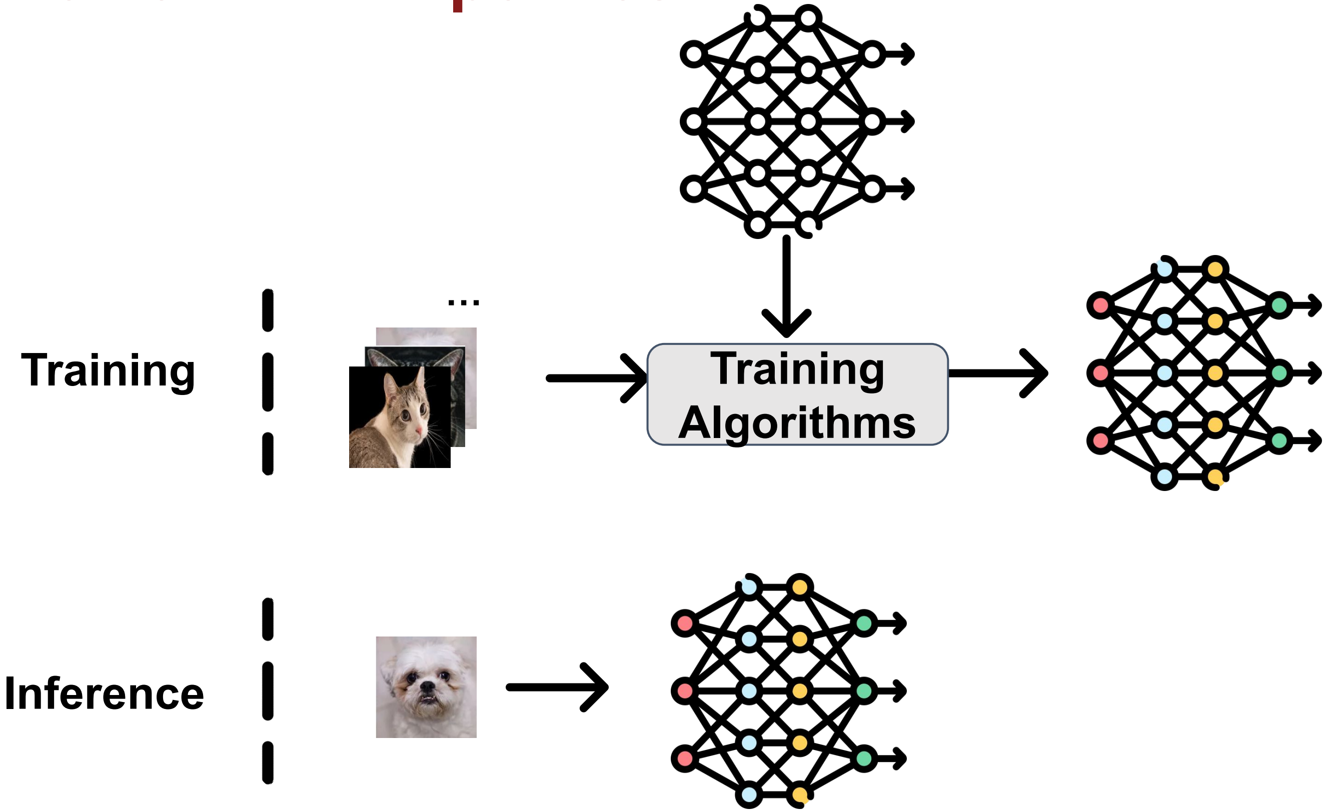
Review: ML Pipelines



Review: ML Pipelines

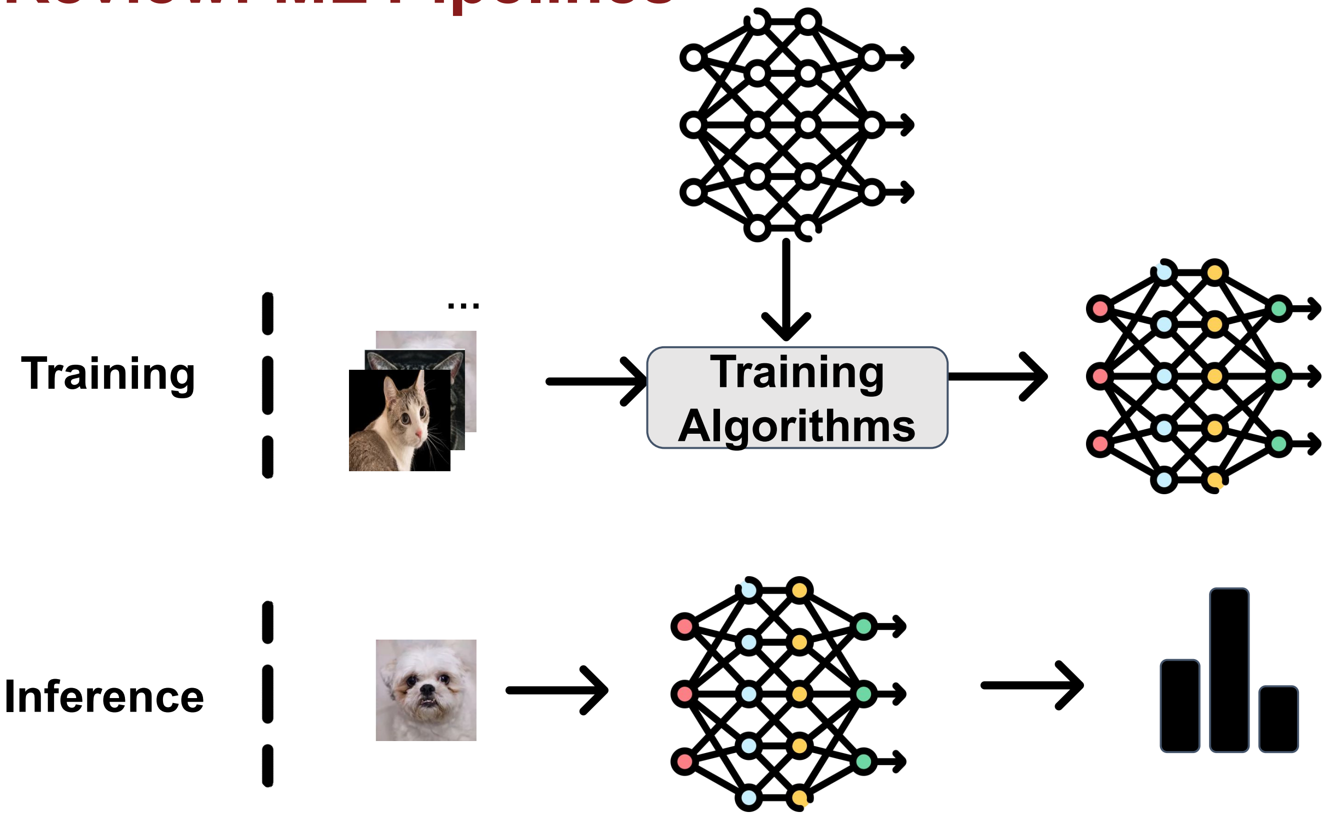


Review: ML Pipelines

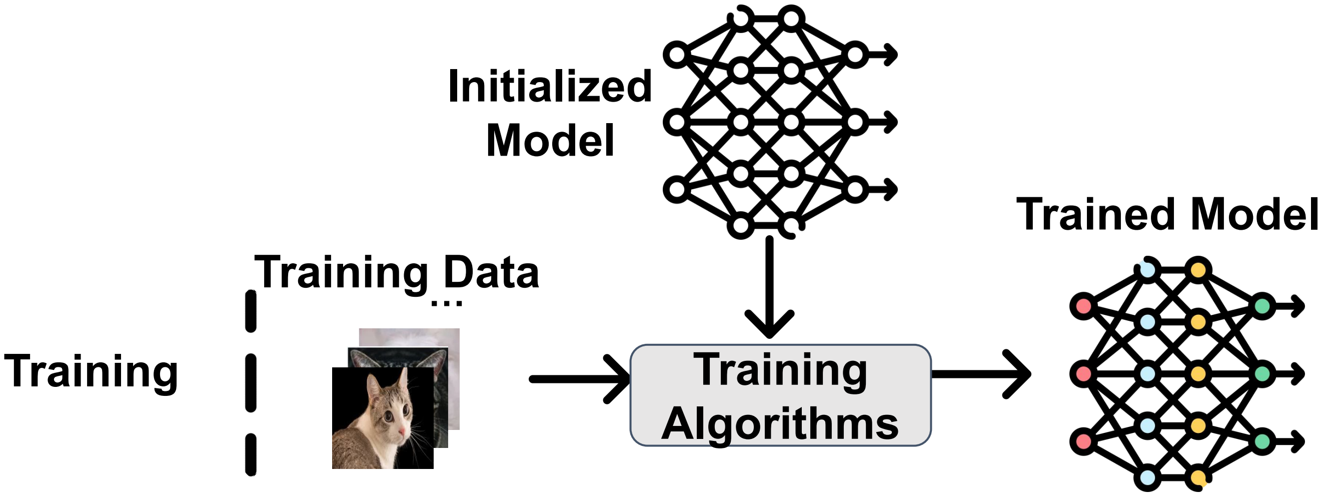


University of Massachusetts Amherst

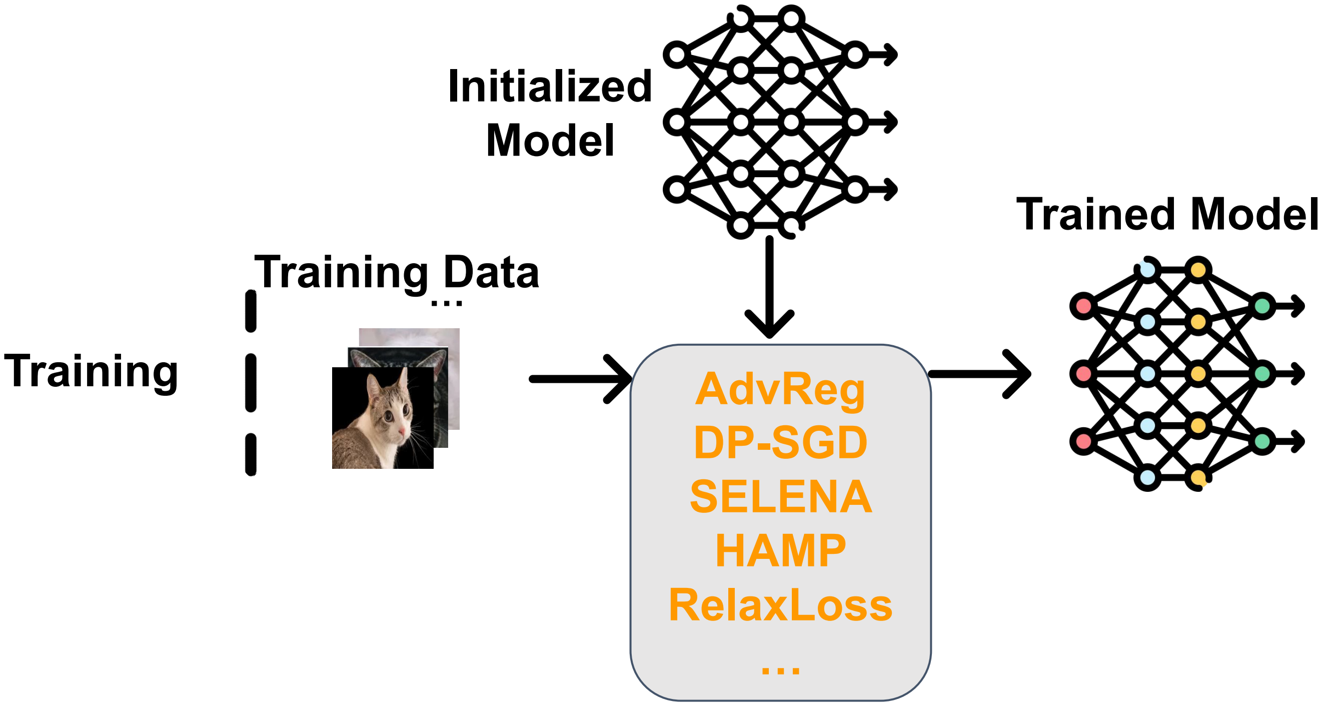
Review: ML Pipelines



Existing MIA Defenses: Training Time



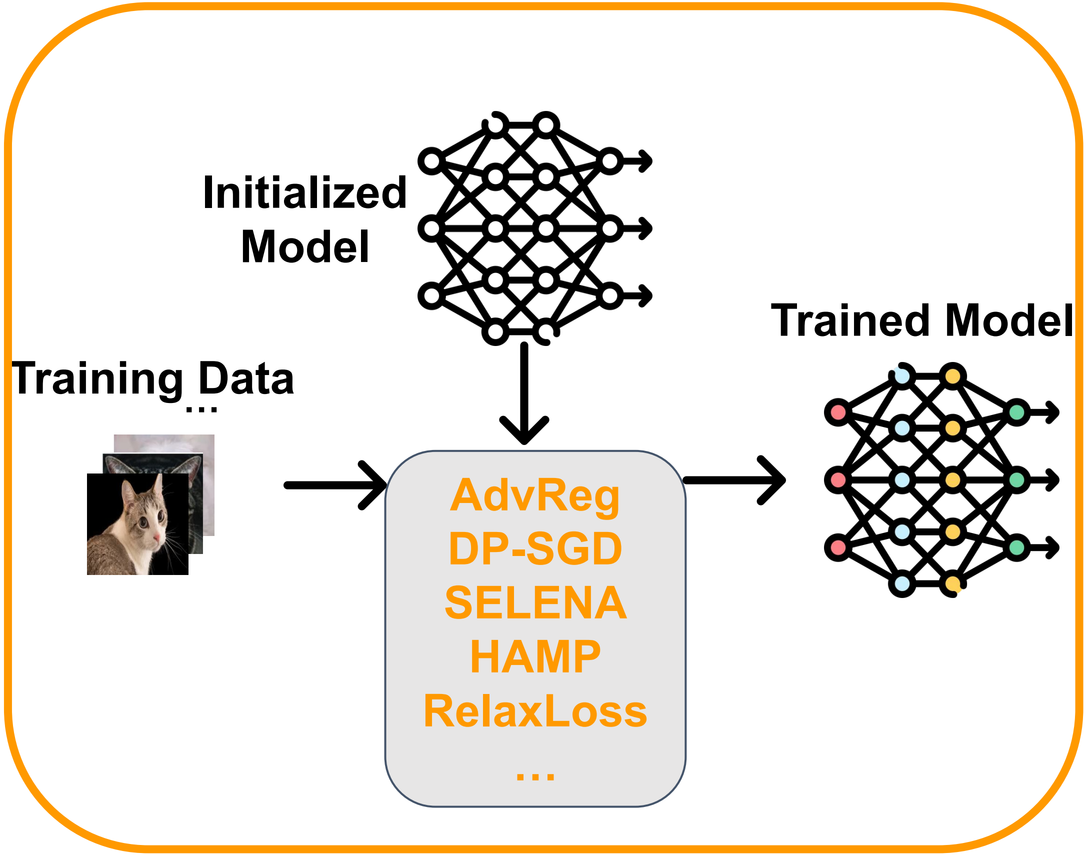
Existing MIA Defenses: Training Time



University of Massachusetts Amherst

Existing MIA Defenses: Training Time

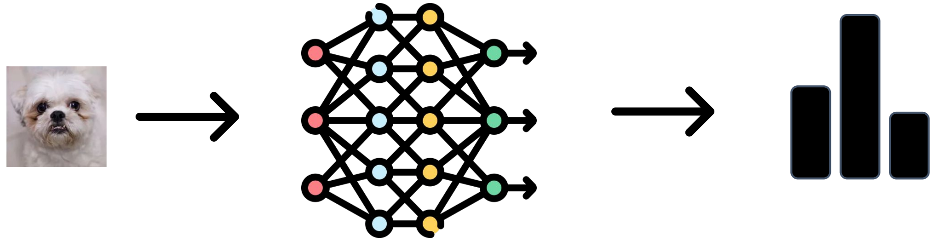
Training Phase Defense



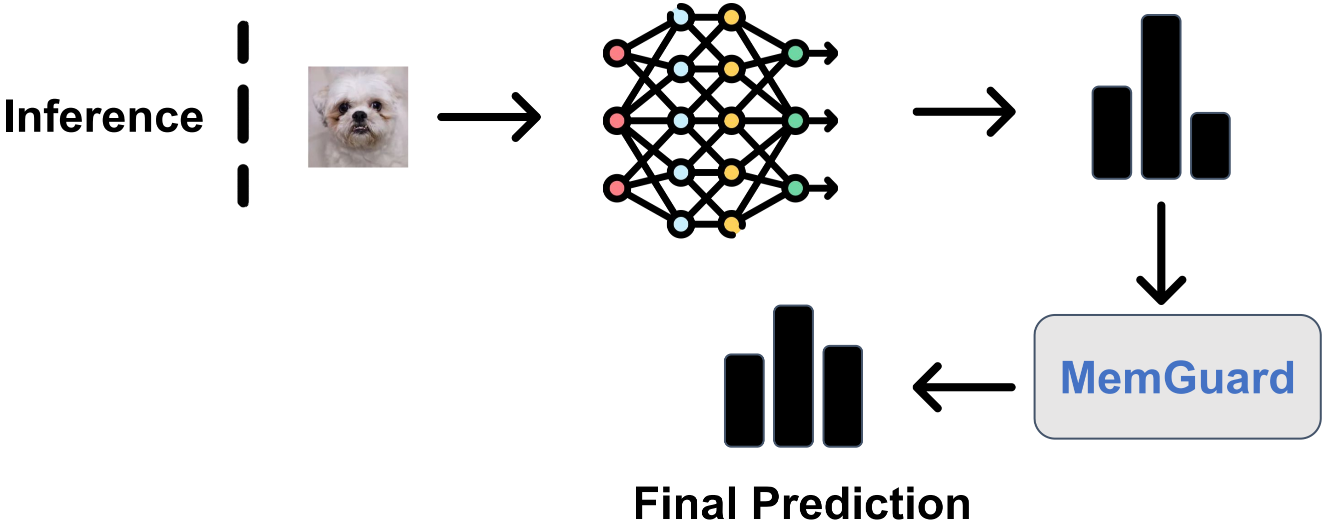
University of Massachusetts Amherst

Existing MIA Defenses: Post-Inference Time

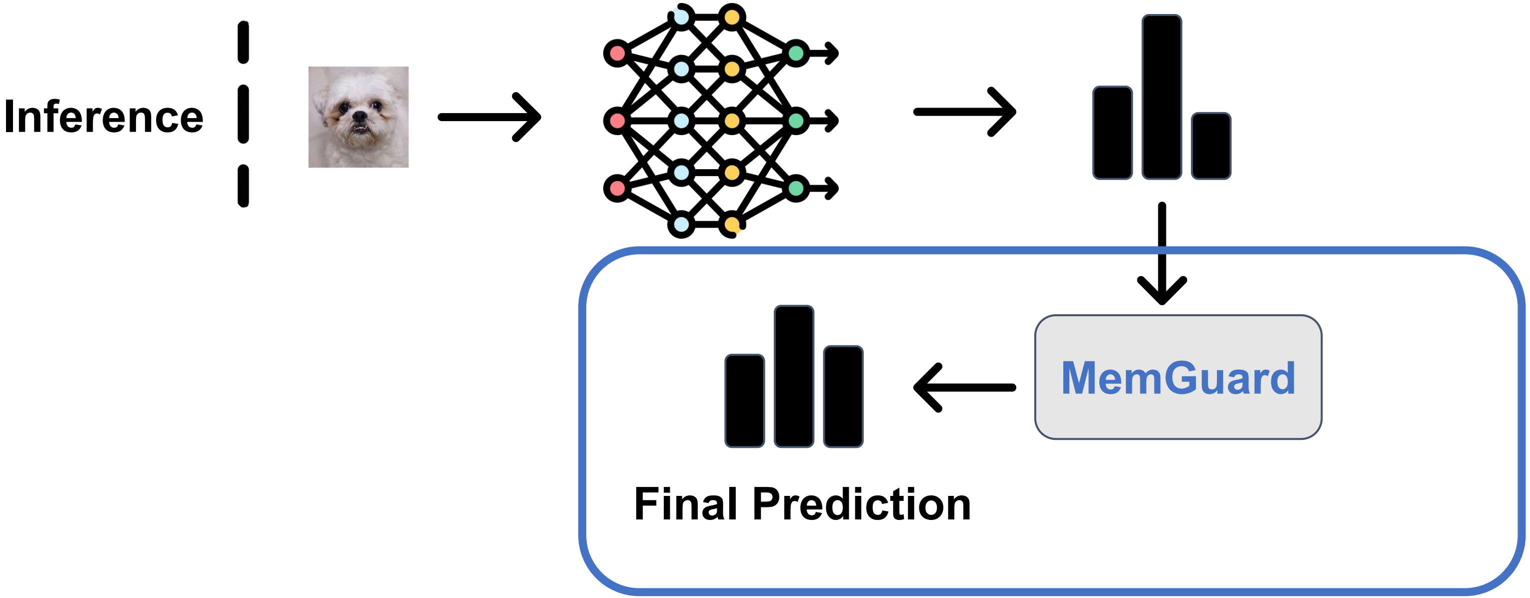
Inference



Existing MIA Defenses: Post-Inference Time



Existing MIA Defenses: Post-Inference Time



Post-Inference
Phase Defense

Shortcomings of Existing Defenses

- **Requires re-training**

TABLE I: **A summary of existing defenses.** ✓means the information is required by the adversary, - otherwise.

Technique	**Requires Re-training**	**Requires Additional Data**	**Impact on Model Accuracy**	**Deployment Stage**
AdvReg [1]	✓	✓	High	Training
MemGuard [2]	-	✓	None	Post-Inference
DPSGD [3]	✓	-	High	Training
SELENA [4]	✓	-	Low	Training
RelaxLoss [5]	✓	-	None	Training
HAMP [6]	✓	-	Low	Training

University of Massachusetts Amherst

Shortcomings of Existing Defenses

- Requires re-training
- **Require additional data**

TABLE I: **A summary of existing defenses.** ✓means the information is required by the adversary, - otherwise.

Technique	**Requires Re-training**	**Requires Additional Data**	**Impact on Model Accuracy**	**Deployment Stage**
AdvReg [1]	✓	✓	High	Training
MemGuard [2]	-	✓	None	Post-Inference
DPSGD [3]	✓	-	High	Training
SELENA [4]	✓	-	Low	Training
RelaxLoss [5]	✓	-	None	Training
HAMP [6]	✓	-	Low	Training

University of Massachusetts Amherst

Shortcomings of Existing Defenses

- Requires re-training
- Require additional Data
- **Poor Privacy-utility trade-off**

TABLE I: **A summary of existing defenses.** ✓means the information is required by the adversary, - otherwise.

Technique	Requires Re-training	Requires Additional Data	Impact on Model Accuracy	Deployment Stage
AdvReg [1]	✓	✓	High	Training
MemGuard [2]	-	✓	None	Post-Inference
DPSGD [3]	✓	-	High	Training
SELENA [4]	✓	-	Low	Training
RelaxLoss [5]	✓	-	None	Training
HAMP [6]	✓	-	Low	Training

University of Massachusetts Amherst

A New Type of Defense: DIFFENCE

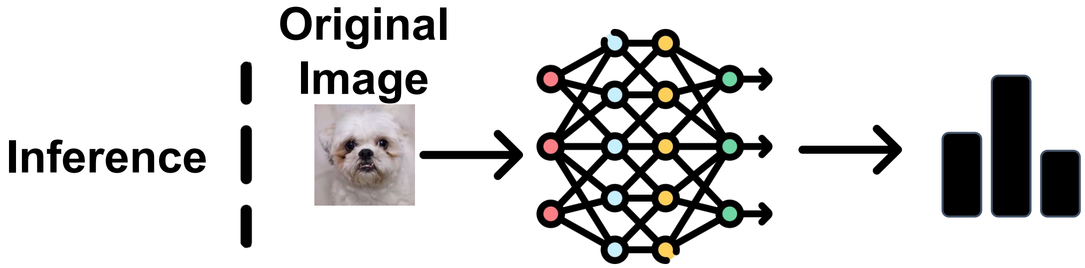
- No retraining
- Additional Data is optional
- Enhanced privacy-utility trade-off

TABLE I: **A comparison to prior works.** ✓means the information is required by the adversary, - otherwise.

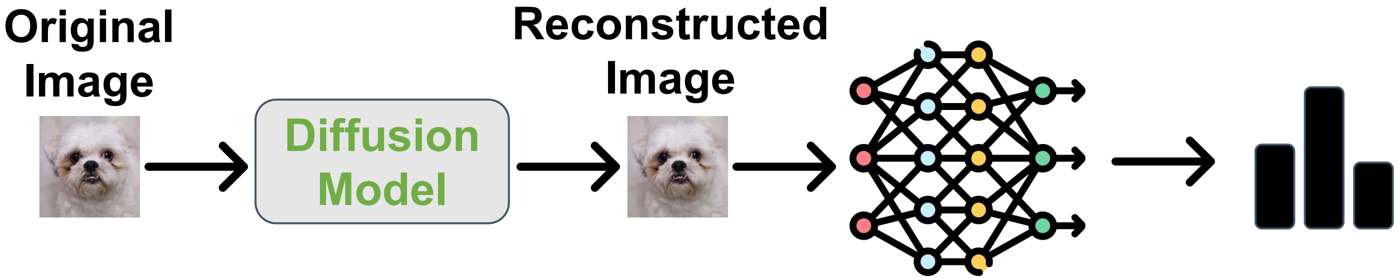
Technique	Requires Re-training	Requires Additional Data	Impact on Model Accuracy	Deployment Stage
AdvReg [1]	✓	✓	High	Training
MemGuard [2]	-	✓	None	Post-Inference
DPSGD [3]	✓	-	High	Training
SELENA [4]	✓	-	Low	Training
RelaxLoss [5]	✓	-	None	Training
HAMP [6]	✓	-	Low	Training
DIFFENCE (Scenario 1)	-	✓	None	Pre-inference
DIFFENCE (Scenario 2)	-	-	None	Pre-inference
DIFFENCE (Scenario 3)	-	-	None	Pre-inference

University of Massachusetts Amherst

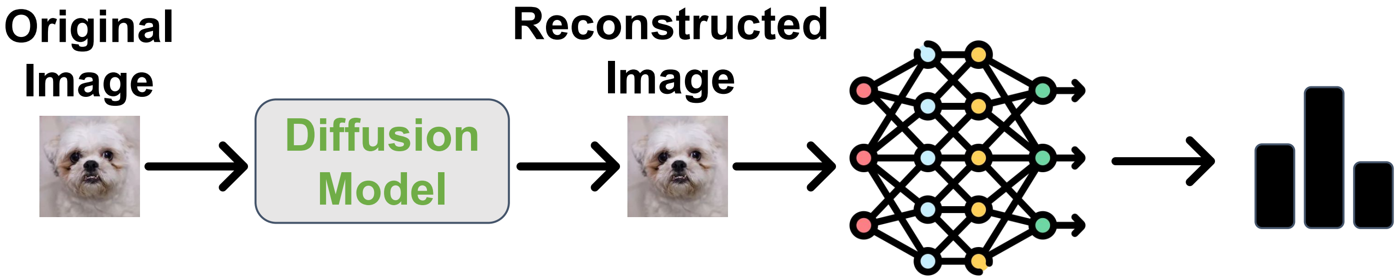
DIFFENCE: High-Level Overview



DIFFENCE: High-Level Overview



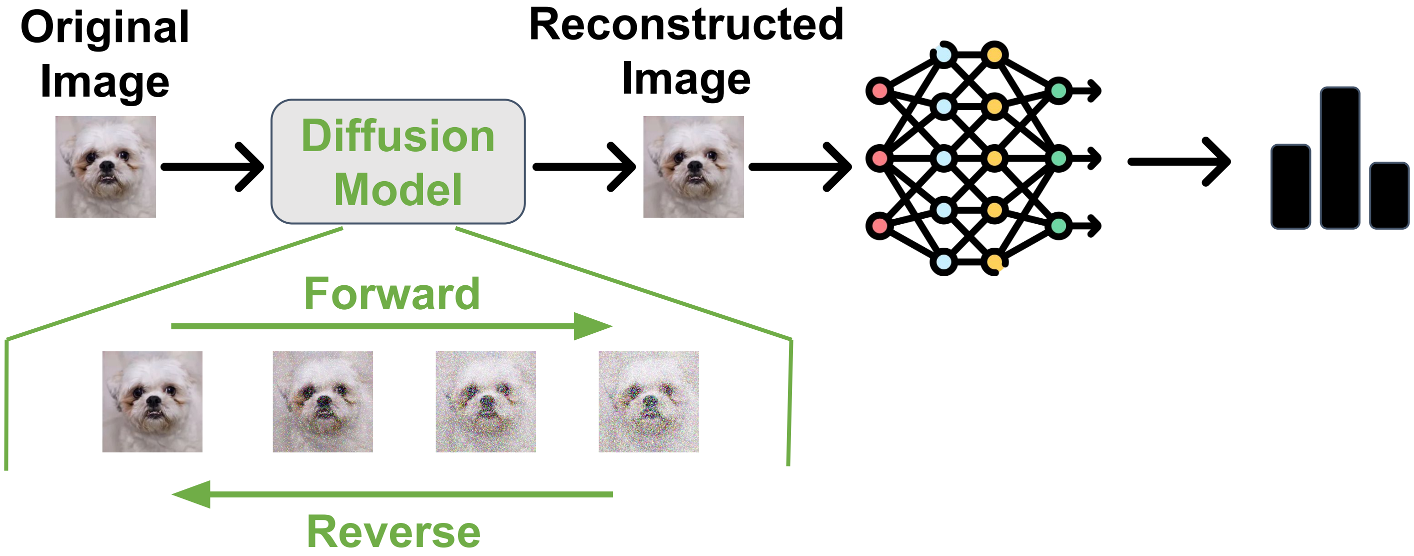
DIFFENCE: High-Level Overview



The model encounters samples that are not exact replicas of those observed during training

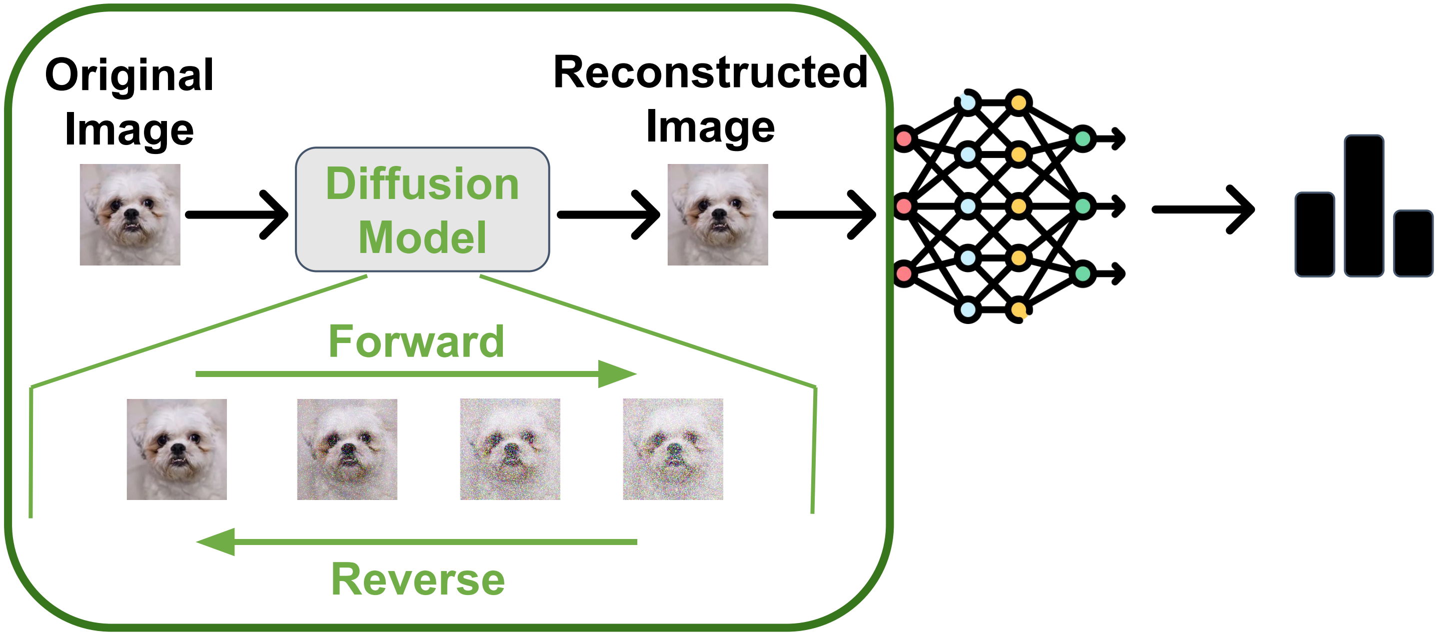
University *of* Massachusetts Amherst

DIFFENCE: High-Level Overview



University of Massachusetts Amherst

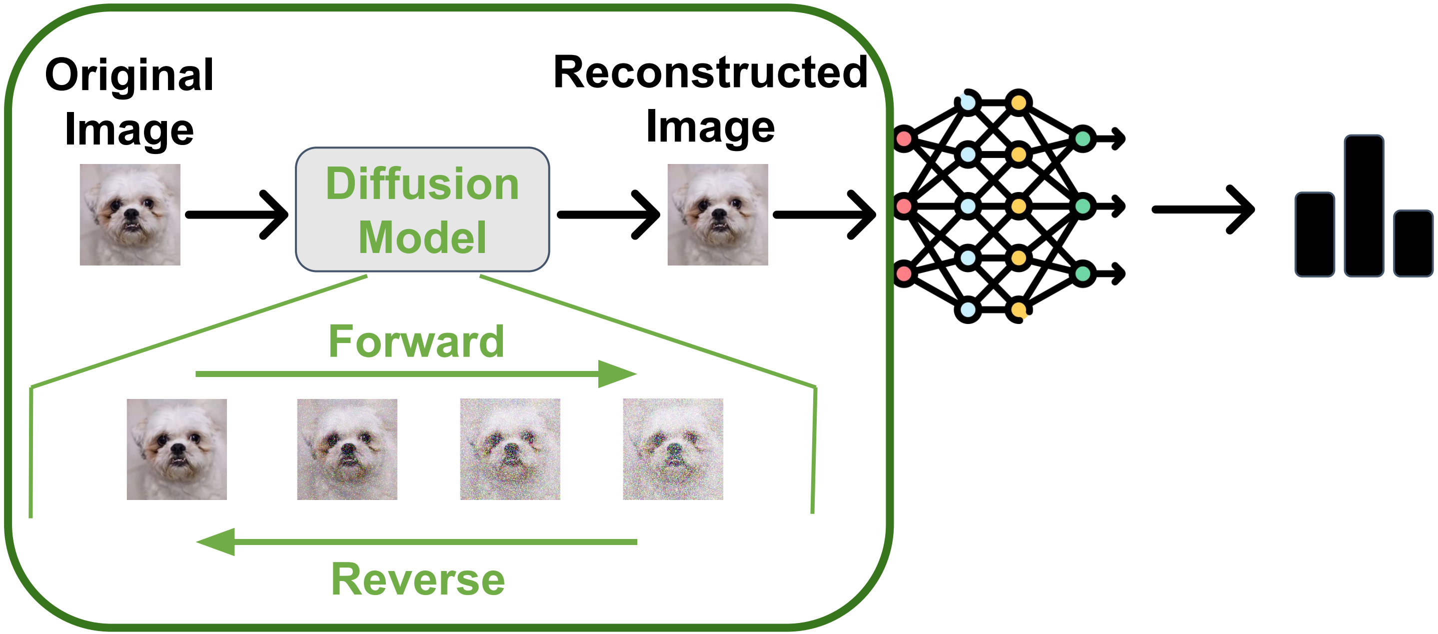
DIFFENCE: High-Level Overview



Pre-Inference Phase Defense

DIFFENCE: High-Level Overview

- No re-training

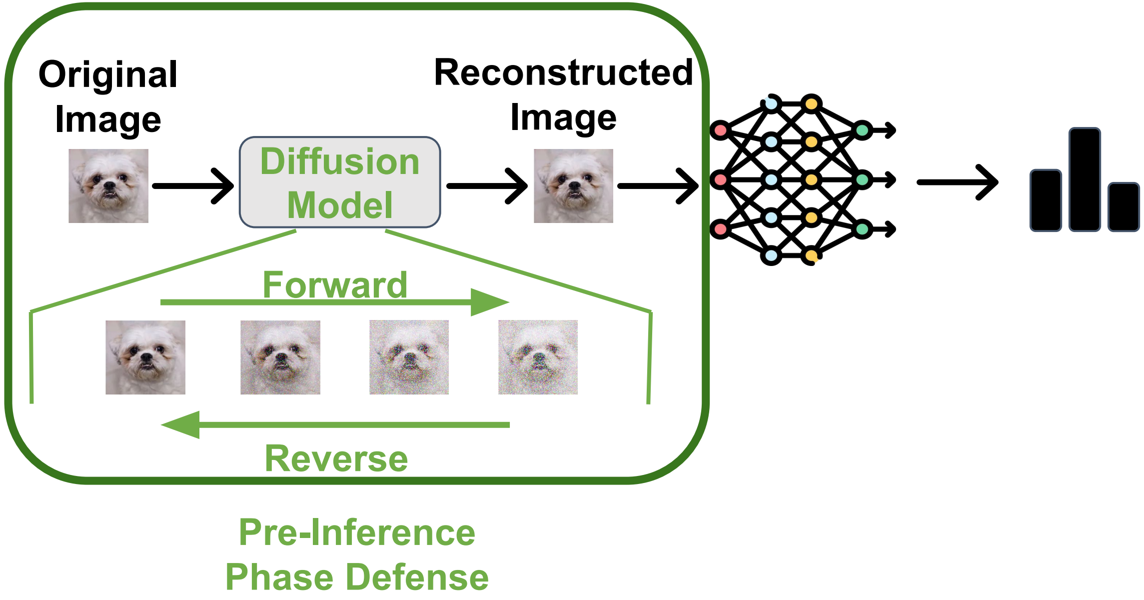


Pre-Inference Phase Defense

University of Massachusetts Amherst

DIFFENCE: High-Level Overview

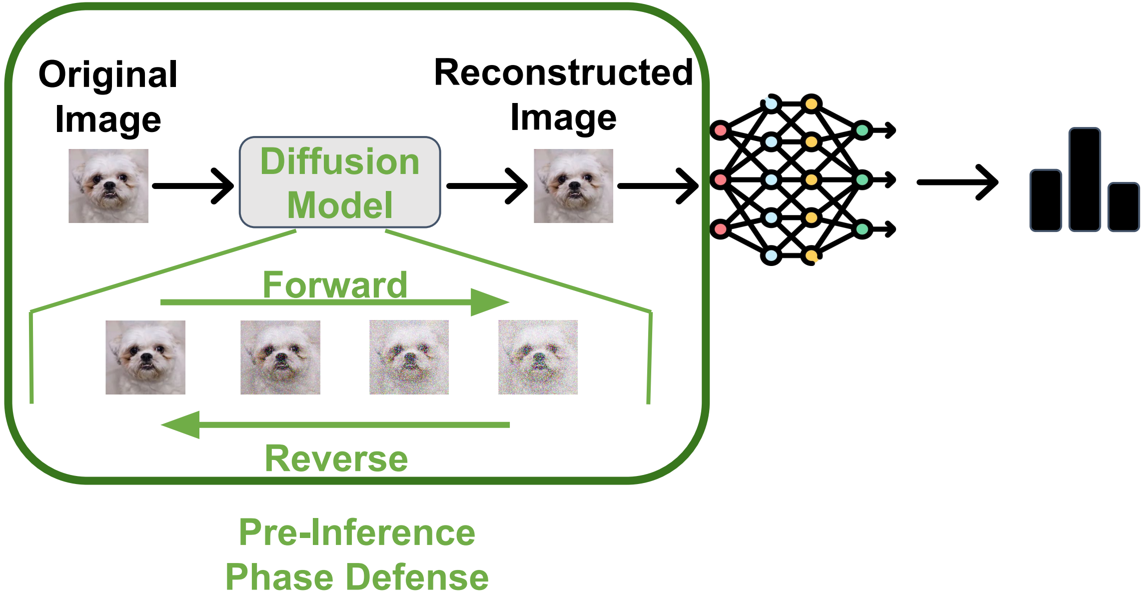
- **No re-training**



University of Massachusetts Amherst

DIFFENCE: High-Level Overview

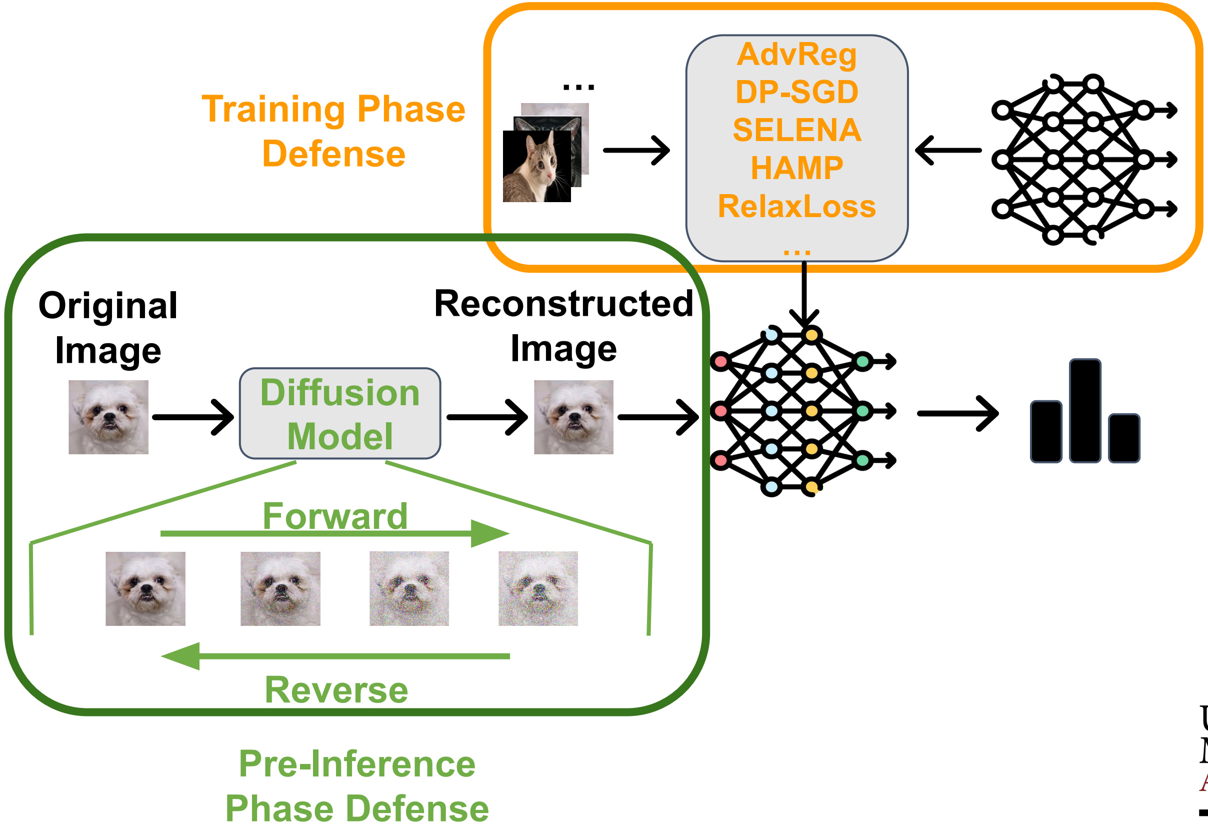
- No re-training
- **Plug-n-play**



University of Massachusetts Amherst

DIFFENCE: High-Level Overview

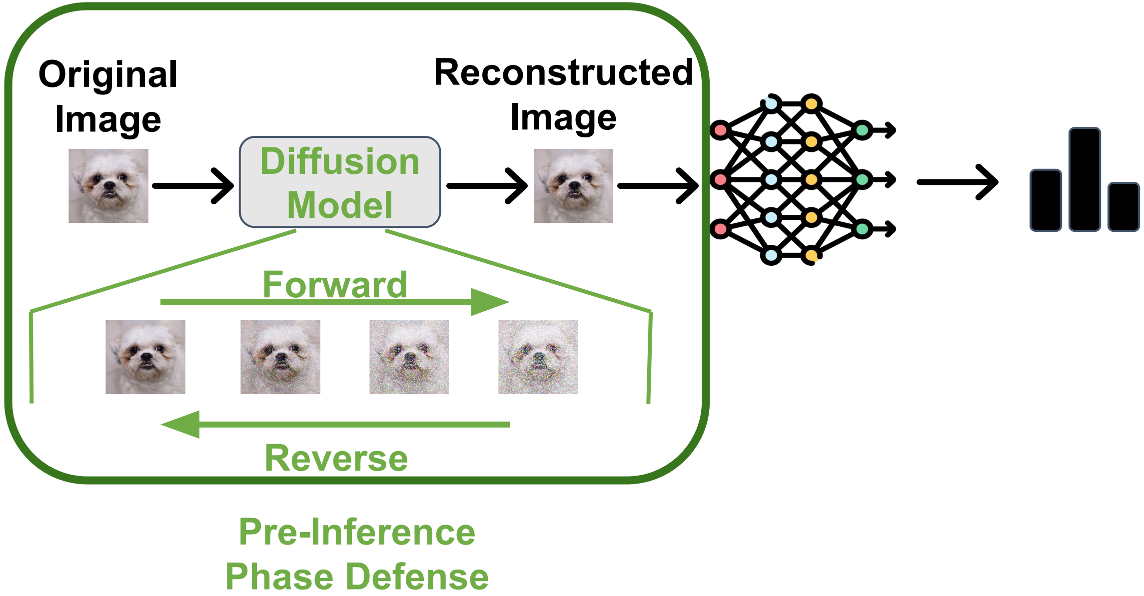
- No re-training
- **Plug-n-play**



University of Massachusetts Amherst

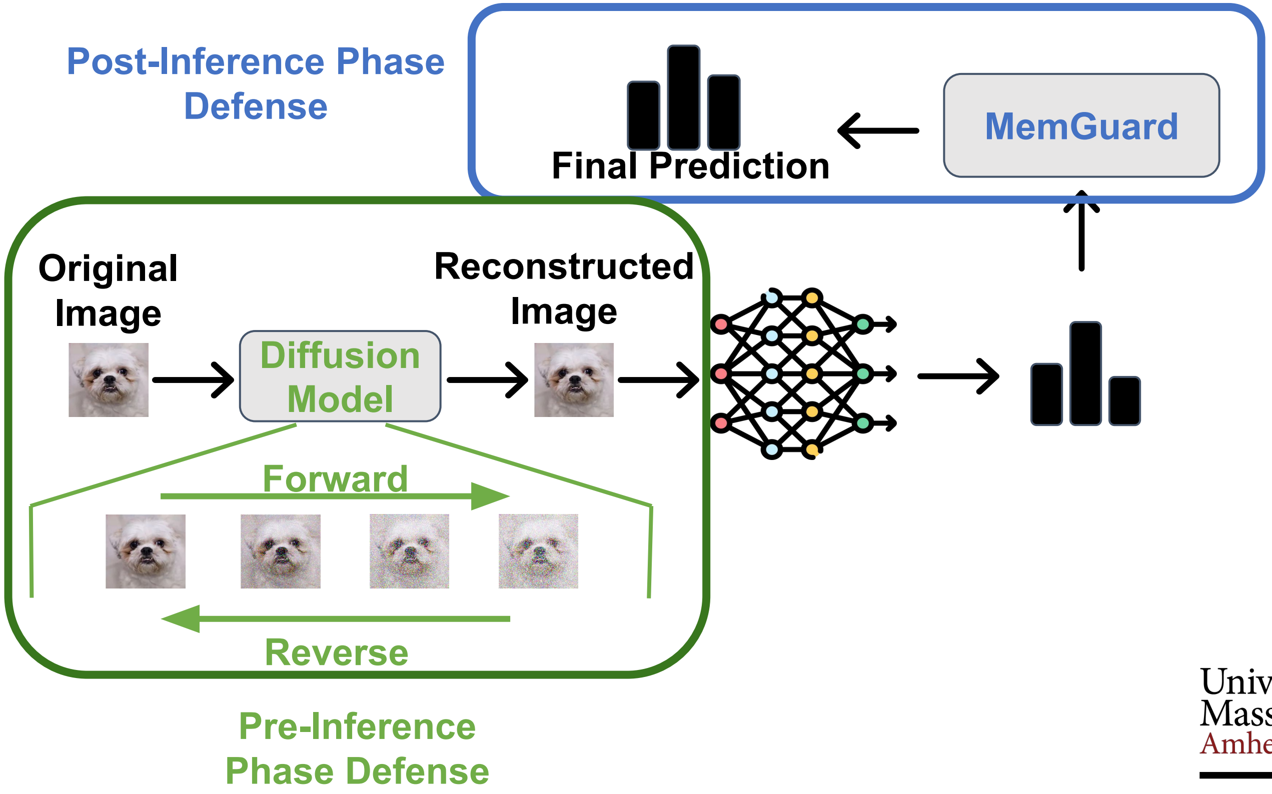
DIFFENCE: High-Level Overview

- No re-training
- **Plug-n-play**

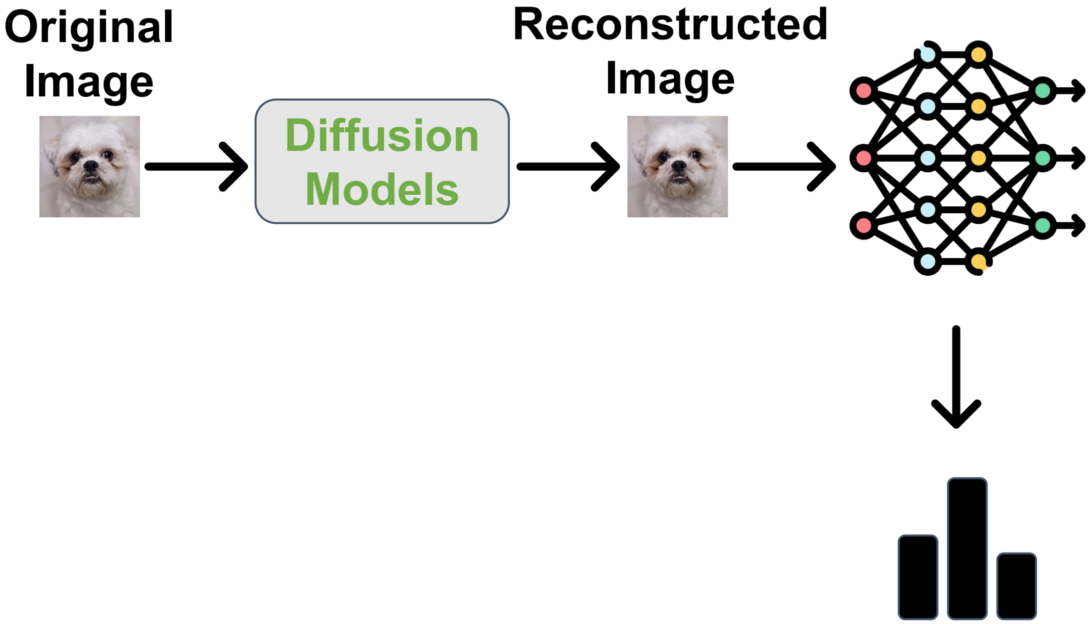


DIFFENCE: High-Level Overview

- No re-training
- **Plug-n-play**

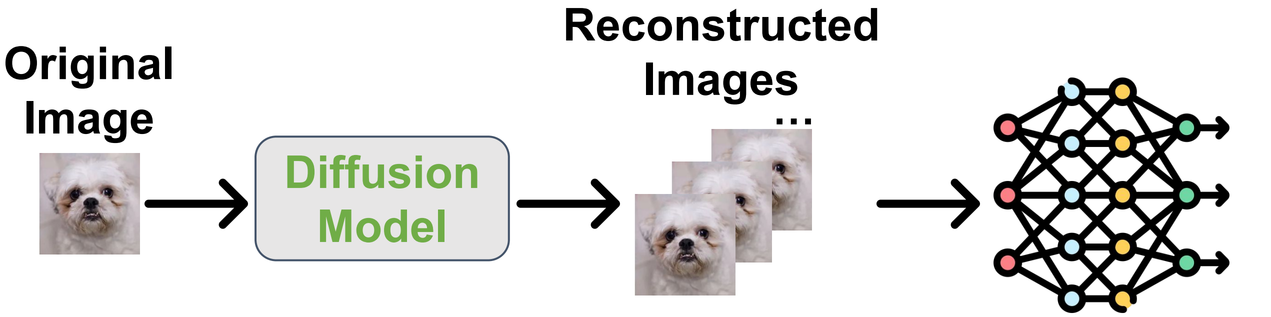


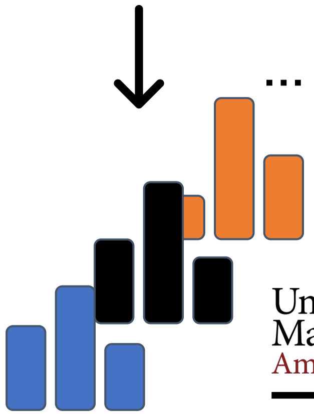
How DIFFENCE Works



- A closer look

How DIFFENCE Works

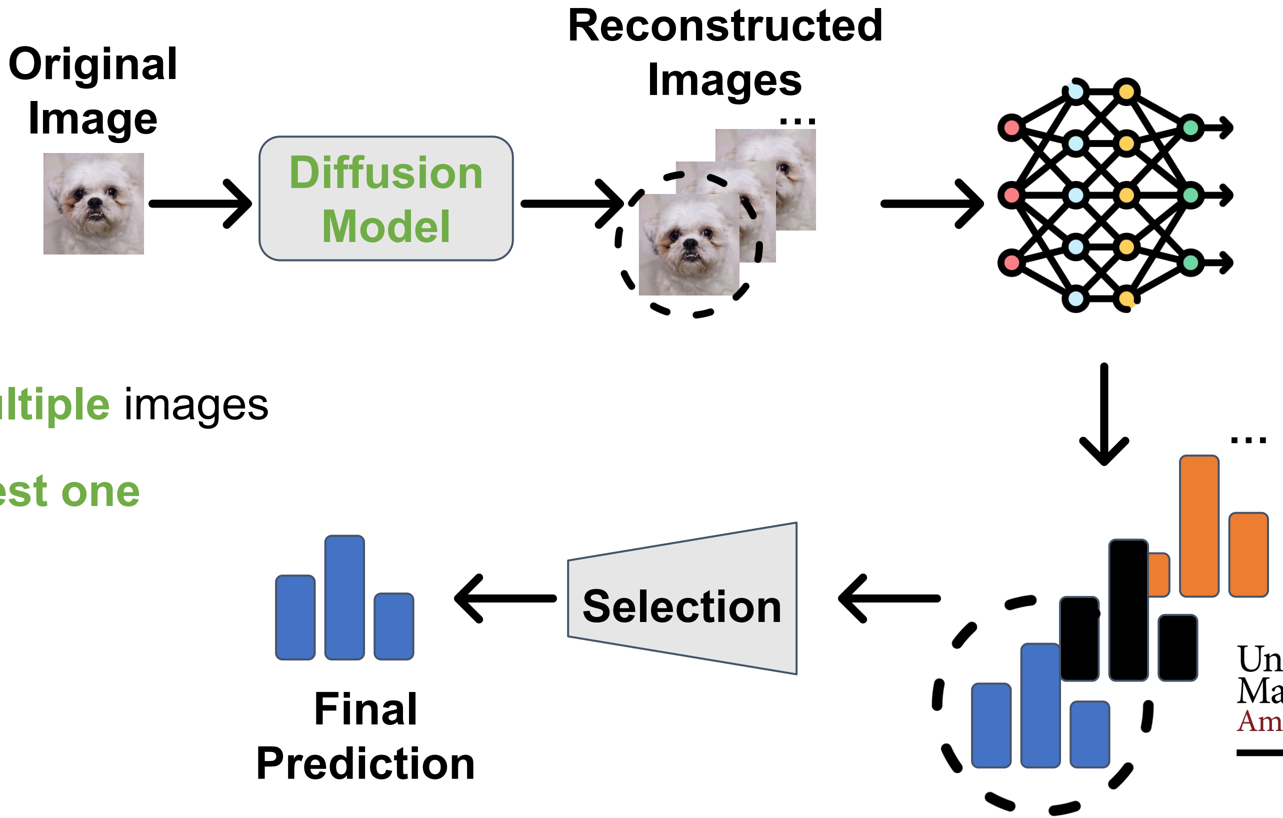




- **A closer look**
  - Generate **multiple** images

University of Massachusetts Amherst

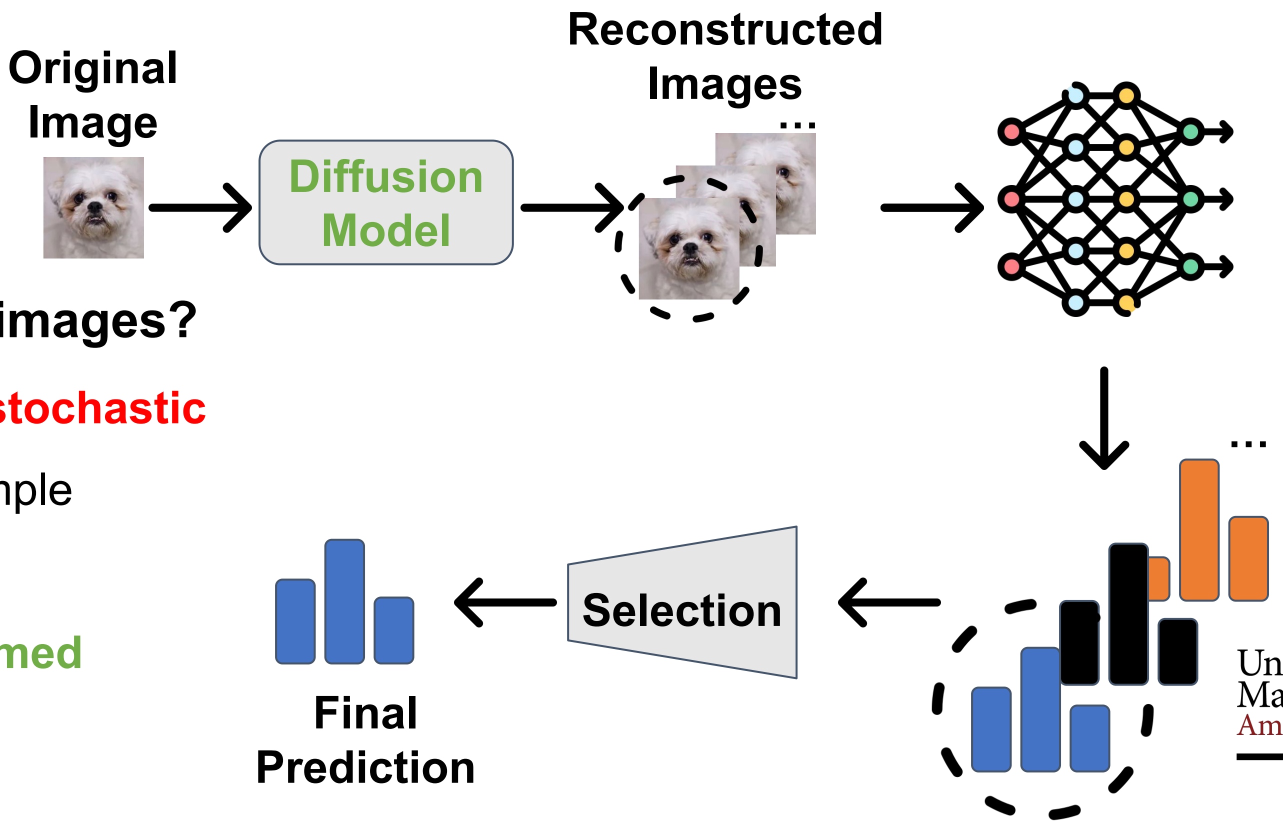
How DIFFENCE Works



- **A closer look**
  - Generate **multiple** images
  - Select **the best one**

University of Massachusetts Amherst

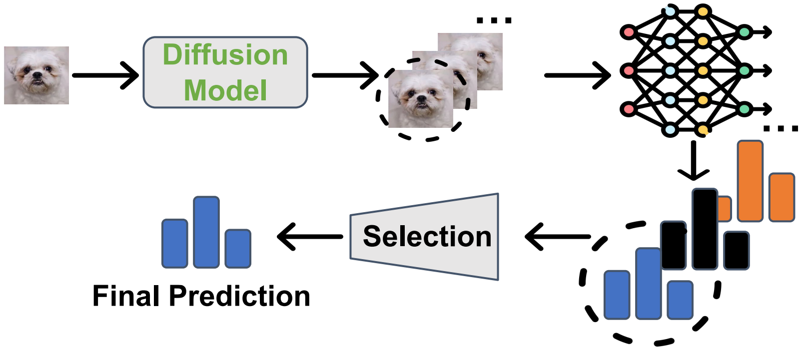
How DIFFENCE Works



- **Why multiple images?**
  - Mitigate the **stochastic** nature of sample generation
  - Enable **informed selection**

University of Massachusetts Amherst

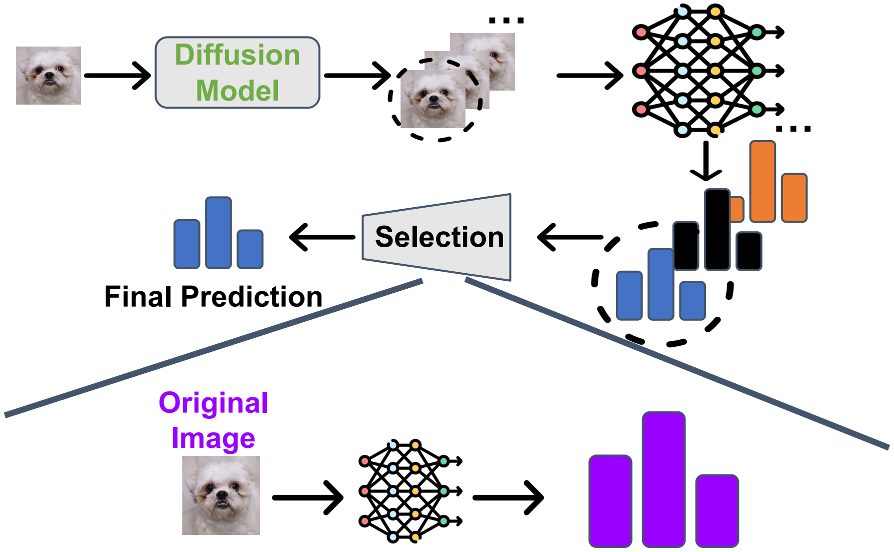
DIFFENCE: Selection Methodology



- **Sample selection**

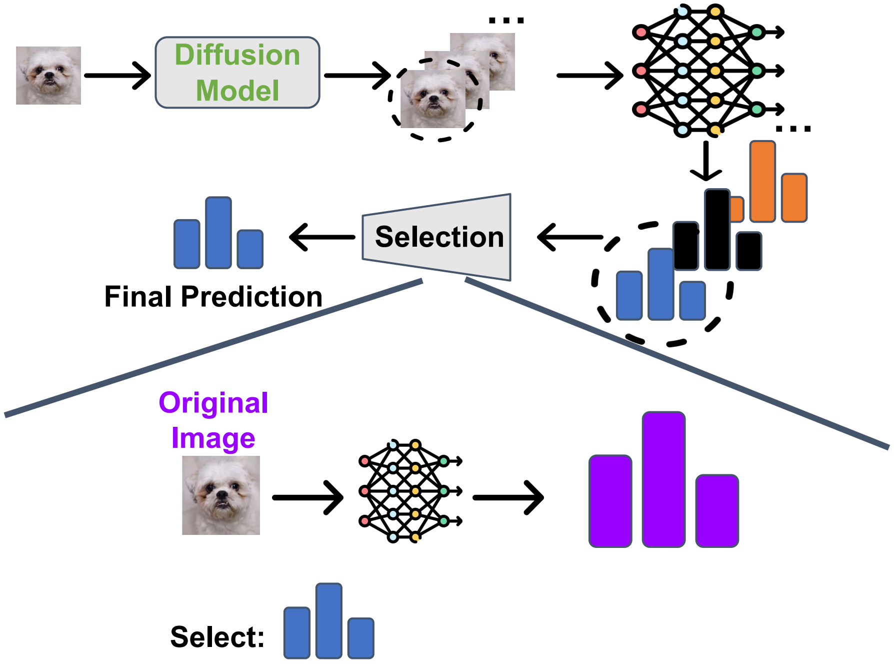
DIFFENCE: Selection Methodology

- **Sample selection**



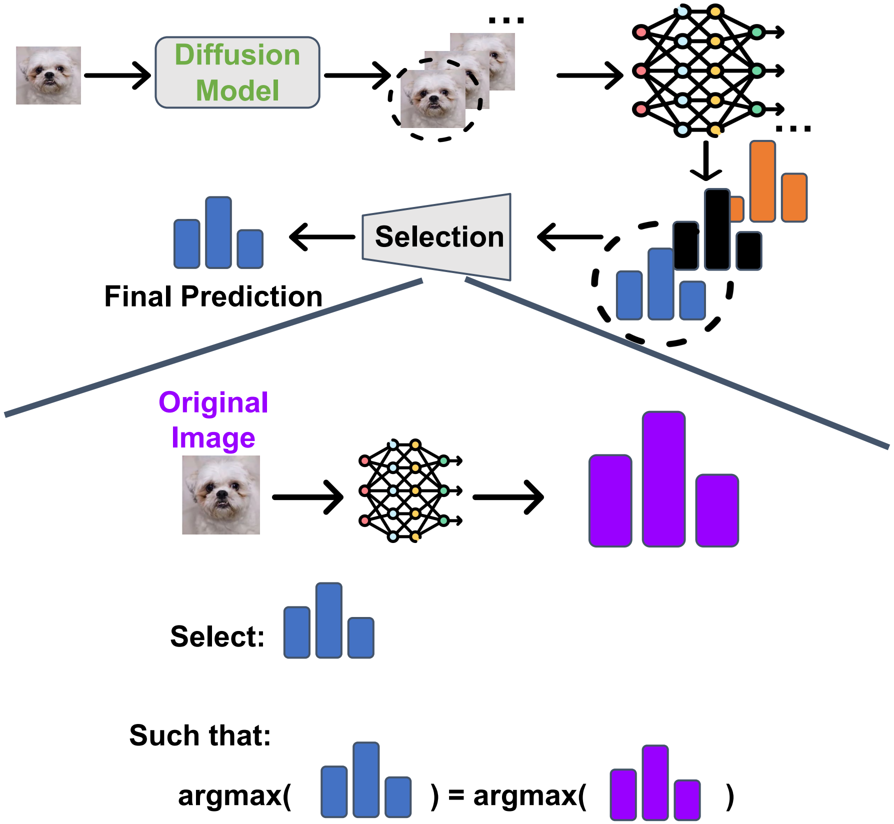
DIFFENCE: Selection Methodology

- **Sample selection**



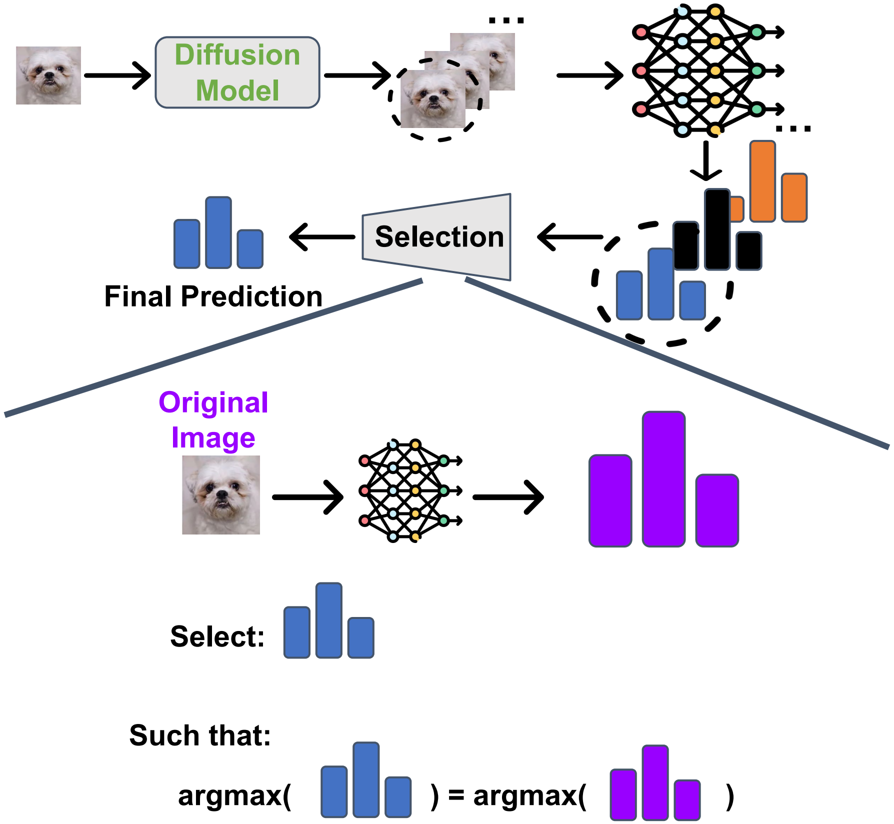
DIFFENCE: Selection Methodology

- **Sample selection**



DIFFENCE: Selection Methodology

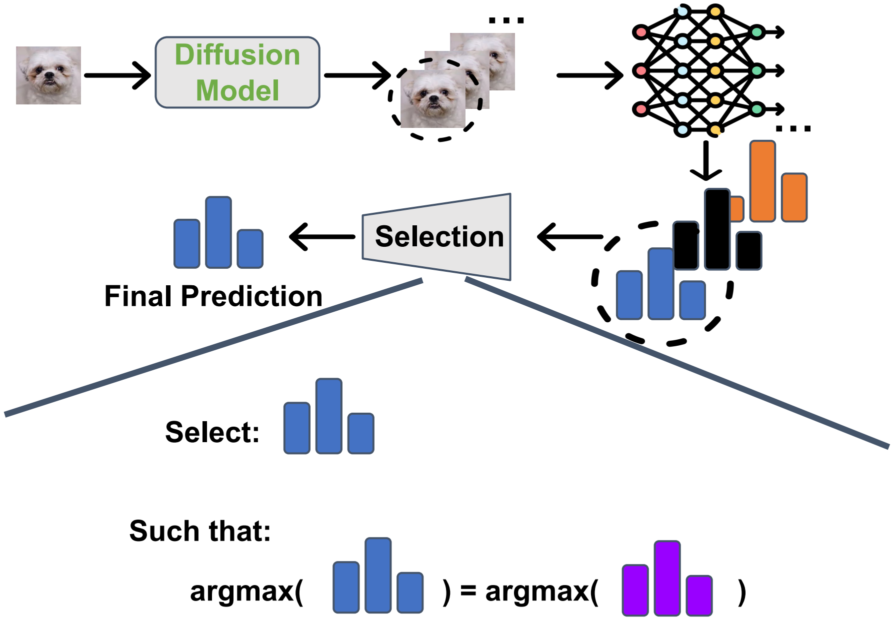
- **Sample selection**
  - Predicted label **matches the original sample**



University of Massachusetts Amherst

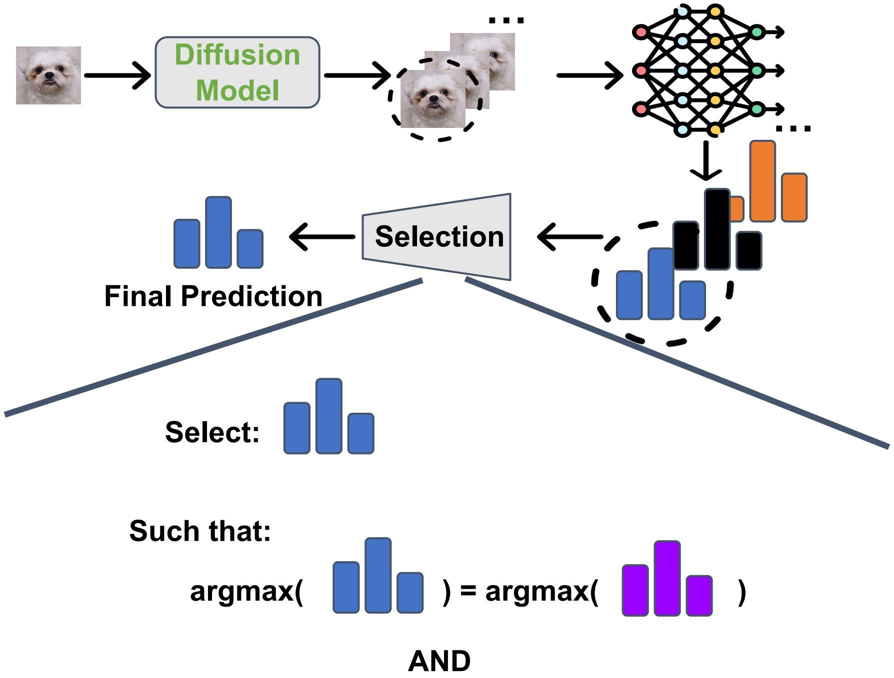
DIFFENCE: Selection Methodology

- **Sample selection**
    - Predicted label **matches the original sample**



DIFFENCE: Selection Methodology

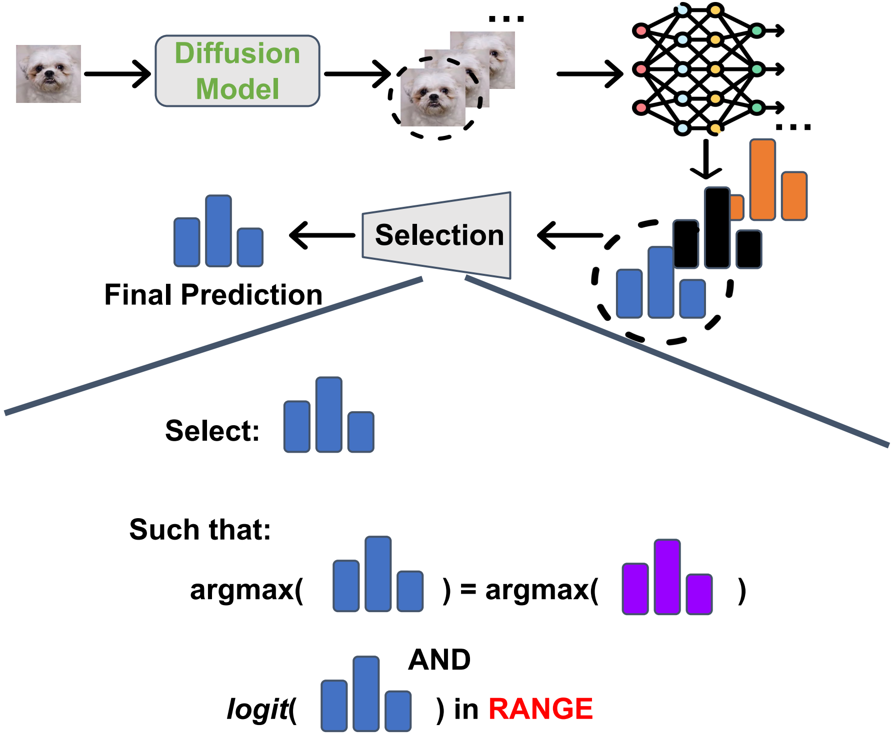
- **Sample selection**
  - Predicted label **matches the original sample**



University of Massachusetts Amherst

DIFFENCE: Selection Methodology

- **Sample selection**
  - Predicted label **matches the original sample**
  - Predicted logit falls within a pre-computed RANGE

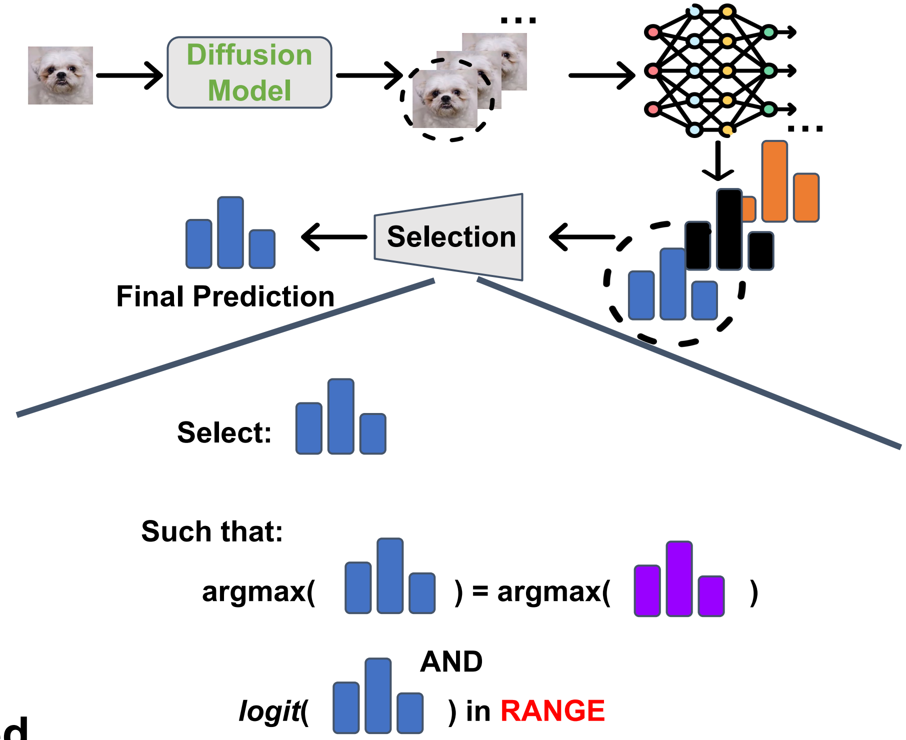


University of Massachusetts Amherst

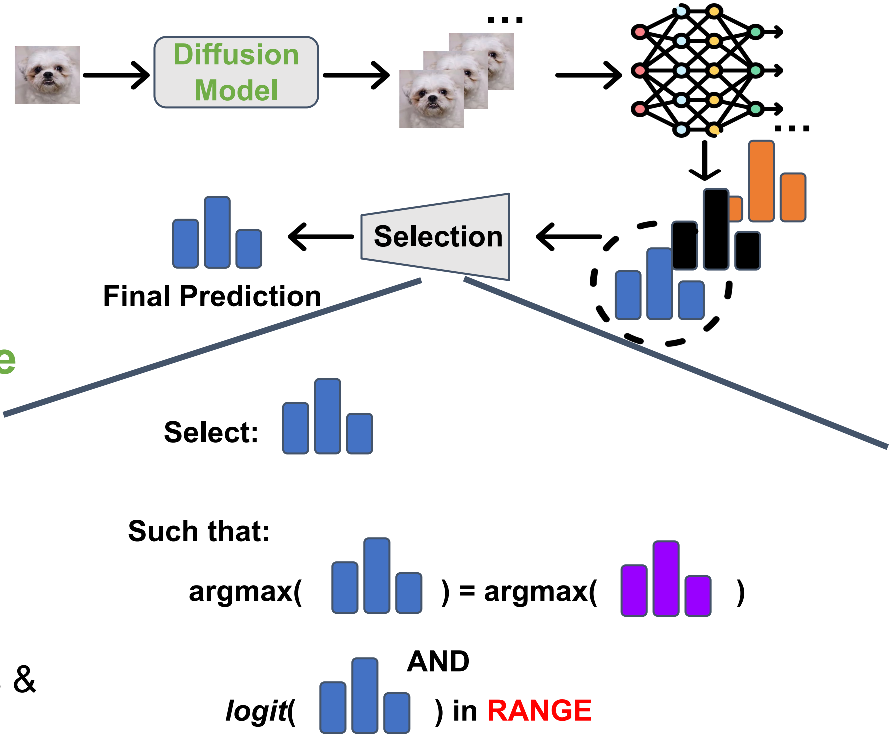
DIFFENCE: Selection Methodology

- **Sample selection**
  - Predicted label **matches the original sample**
  - Predicted logit falls within a pre-computed RANGE

RANGE is calculated based on defender's knowledge



DIFFENCE: Selection Methodology



- **Sample selection**
  - Predicted label **matches the original sample**
  - Predicted logit falls within a pre-computed RANGE
    - **Scenario 1:** Knows members & non-members
    - **Scenario 2:** Knows members
    - **Scenario 3:** No knowledge

University of Massachusetts Amherst

Evaluation

📌 **Datasets (🗂️ 5)**

- 🗂️ **CIFAR-10**
- 🗂️ **CIFAR-100**
- 🗂️ **SVHN**
- 🗂️ **CelebA**
- 🗂️ **UTKFace**

📌 **Models (🤖 4)**

- 🤖 **ResNet-18**
- 🤖 **DenseNet-121**
- 🤖 **VGG-16**
- 🤖 **Vision Transformers**

📌 **Attacks (⚔️ 6)**

- ⚔️ **NN-based**
- ⚔️ **Loss-based**
- ⚔️ **Confidence-based**
- ⚔️ **Entropy-based**
- ⚔️ **Modified-entropy-based**
- ⚔️ **Likelihood-ratio (LiRA)**

📌 **Defenses (🛡️ 6)**

- 🛡️ **AdvReg (CCS'18)**
- 🛡️ **MemGuard (CCS'19)**
- 🛡️ **SELENA (USENIX'22)**
- 🛡️ **DP-SGD (CCS'16)**
- 🛡️ **HAMP (NDSS'24)**
- 🛡️ **RelaxLoss (ICLR'22)**

Key Results

TABLE: **Average attack AUC** (lower is better). The best (lowest) AUC under each defense is in **bold**. Columns "Δ" show how much the AUC decreases compared to "w/o DIFFENCE".

Defenses	**Prediction Accuracy Delta (%)**	**w/o DIFFENCE** (AUC %)	**w/ DIFFENCE (Scenario 1)**		**w/ DIFFENCE (Scenario 2)**		**w/ DIFFENCE (Scenario 3)**	
			AUC (%)	**Δ (%)**	**AUC (%)**	**Δ (%)**	**AUC (%)**	**Δ (%)**
Undefended	0	79.14	**68.08**	−11.06	70.79	−8.35	69.12	−10.02
SELENA	−2.13	62.22	**56.00**	−6.22	60.30	−1.92	57.81	−4.41
AdvReg	−5.53	61.32	**59.17**	−2.15	61.33	0.01	60.87	−0.45
HAMP	−0.23	78.96	**67.60**	−11.36	71.23	−7.73	69.18	−9.78
RelaxLoss	0.97	75.81	**67.13**	−8.68	69.56	−6.25	68.60	−7.21
DP-SGD	−9.13	56.61	**55.47**	−1.14	58.40	−1.79	56.60	−0.01
Memguard	0	69.53	**66.76**	−2.77	67.23	−2.30	67.48	−2.05

University of Massachusetts Amherst

Key Results

TABLE: **Average attack AUC** (lower is better). The best (lowest) AUC under each defense is in **bold**. Columns "Δ" show how much the AUC decreases compared to "w/o DIFFENCE".

Defenses	Prediction Accuracy Delta (%)	w/o DIFFENCE (AUC %)	w/ DIFFENCE (Scenario 1)		w/ DIFFENCE (Scenario 2)		w/ DIFFENCE (Scenario 3)	
			AUC (%)	Δ (%)	AUC (%)	Δ (%)	AUC (%)	Δ (%)
Undefended	0	79.14	**68.08**	−11.06	70.79	−8.35	69.12	−10.02
SELENA	−2.13	62.22	**56.00**	−6.22	60.30	−1.92	57.81	−4.41
AdvReg	−5.53	61.32	**59.17**	−2.15	61.33	0.01	60.87	−0.45
HAMP	−0.23	78.96	**67.60**	−11.36	71.23	−7.73	69.18	−9.78
RelaxLoss	0.97	75.81	**67.13**	−8.68	69.56	−6.25	68.60	−7.21
DP-SGD	−9.13	56.61	**55.47**	−1.14	58.40	−1.79	56.60	−0.01
Memguard	0	69.53	**66.76**	−2.77	67.23	−2.30	67.48	−2.05

- **DIFFENCE** enhances membership privacy for both **undefended** models and models **defended with other methods**

Key Results

TABLE: **Average attack AUC** (lower is better). The best (lowest) AUC under each defense is in **bold**. Columns "Δ" show how much the AUC decreases compared to "w/o DIFFENCE".

Defenses	Prediction Accuracy Delta (%)	w/o DIFFENCE (AUC %)	w/ DIFFENCE (Scenario 1)		w/ DIFFENCE (Scenario 2)		w/ DIFFENCE (Scenario 3)	
			AUC (%)	Δ (%)	AUC (%)	Δ (%)	AUC (%)	Δ (%)
Undefended	0	$\boxed{79.14}$	**68.08**	−11.06	70.79	−8.35	69.12	−10.02
SELENA	−2.13	62.22	**56.00**	−6.22	60.30	−1.92	57.81	−4.41
AdvReg	−5.53	61.32	**59.17**	−2.15	61.33	0.01	60.87	−0.45
HAMP	−0.23	78.96	**67.60**	−11.36	71.23	−7.73	69.18	−9.78
RelaxLoss	0.97	75.81	**67.13**	−8.68	69.56	−6.25	68.60	−7.21
DP-SGD	−9.13	56.61	**55.47**	−1.14	58.40	−1.79	56.60	−0.01
Memguard	0	69.53	**66.76**	−2.77	67.23	−2.30	67.48	−2.05

- Example:
  - Undefended (79.14%)

University of Massachusetts Amherst

Key Results

TABLE: **Average attack AUC** (lower is better). The best (lowest) AUC under each defense is in **bold**. Columns "Δ" show how much the AUC decreases compared to "w/o DIFFENCE".

Defenses	Prediction Accuracy Delta (%)	w/o DIFFENCE (AUC %)	w/ DIFFENCE (Scenario 1)		w/ DIFFENCE (Scenario 2)		w/ DIFFENCE (Scenario 3)	
			AUC (%)	Δ (%)	AUC (%)	Δ (%)	AUC (%)	Δ (%)
Undefended	0	79.14	**68.08**	−11.06	70.79	−8.35	69.12	−10.02
SELENA	−2.13	$\boxed{62.22}$	**56.00**	−6.22	60.30	−1.92	57.81	−4.41
AdvReg	−5.53	61.32	**59.17**	−2.15	61.33	0.01	60.87	−0.45
HAMP	−0.23	78.96	**67.60**	−11.36	71.23	−7.73	69.18	−9.78
RelaxLoss	0.97	75.81	**67.13**	−8.68	69.56	−6.25	68.60	−7.21
DP-SGD	−9.13	56.61	**55.47**	−1.14	58.40	−1.79	56.60	−0.01
Memguard	0	69.53	**66.76**	−2.77	67.23	−2.30	67.48	−2.05

- Example:
    - Undefended (79.14%) → SELENA (62.22%)

University of Massachusetts Amherst

Key Results

TABLE: **Average attack AUC** (lower is better). The best (lowest) AUC under each defense is in **bold**. Columns "Δ" show how much the AUC decreases compared to "w/o DIFFENCE".

Defenses	Prediction Accuracy Delta (%)	w/o DIFFENCE (AUC %)	w/ DIFFENCE (Scenario 1)		w/ DIFFENCE (Scenario 2)		w/ DIFFENCE (Scenario 3)	
			AUC (%)	Δ (%)	AUC (%)	Δ (%)	AUC (%)	Δ (%)
Undefended	0	79.14	**68.08**	−11.06	70.79	−8.35	69.12	−10.02
SELENA	−2.13	62.22	**56.00**	−6.22	60.30	−1.92	57.81	−4.41
AdvReg	−5.53	61.32	**59.17**	−2.15	61.33	0.01	60.87	−0.45
HAMP	−0.23	78.96	**67.60**	−11.36	71.23	−7.73	69.18	−9.78
RelaxLoss	0.97	75.81	**67.13**	−8.68	69.56	−6.25	68.60	−7.21
DP-SGD	−9.13	56.61	**55.47**	−1.14	58.40	−1.79	56.60	−0.01
Memguard	0	69.53	**66.76**	−2.77	67.23	−2.30	67.48	−2.05

- Example:
  - Undefended (79.14%) → SELENA (62.22%) → SELENA w/ DIFFENCE (**56.0%**)

University of Massachusetts Amherst

Key Results

TABLE: **Average attack AUC** (lower is better). The best (lowest) AUC under each defense is in **bold**. Columns "Δ" show how much the AUC decreases compared to "w/o DIFFENCE".

Defenses	Prediction Accuracy Delta (%)	w/o DIFFENCE (AUC %)	w/ DIFFENCE (Scenario 1)		w/ DIFFENCE (Scenario 2)		w/ DIFFENCE (Scenario 3)	
			AUC (%)	Δ (%)	AUC (%)	Δ (%)	AUC (%)	Δ (%)
Undefended	0	79.14	**68.08**	−11.06	70.79	−8.35	69.12	−10.02
SELENA	−2.13	62.22	**56.00**	−6.22	60.30	−1.92	57.81	−4.41
AdvReg	−5.53	61.32	**59.17**	−2.15	61.33	0.01	60.87	−0.45
HAMP	−0.23	78.96	**67.60**	−11.36	71.23	−7.73	69.18	−9.78
RelaxLoss	0.97	75.81	**67.13**	−8.68	69.56	−6.25	68.60	−7.21
DP-SGD	−9.13	56.61	**55.47**	−1.14	58.40	−1.79	56.60	−0.01
Memguard	0	69.53	**66.76**	−2.77	67.23	−2.30	67.48	−2.05

- **DIFFENCE** enhances membership privacy for both **undefended** models and models **defended with other methods**
- **DIFFENCE** is most effective when the attacker know some member and non-member samples (Scenario 1).

University of Massachusetts Amherst

Key Results

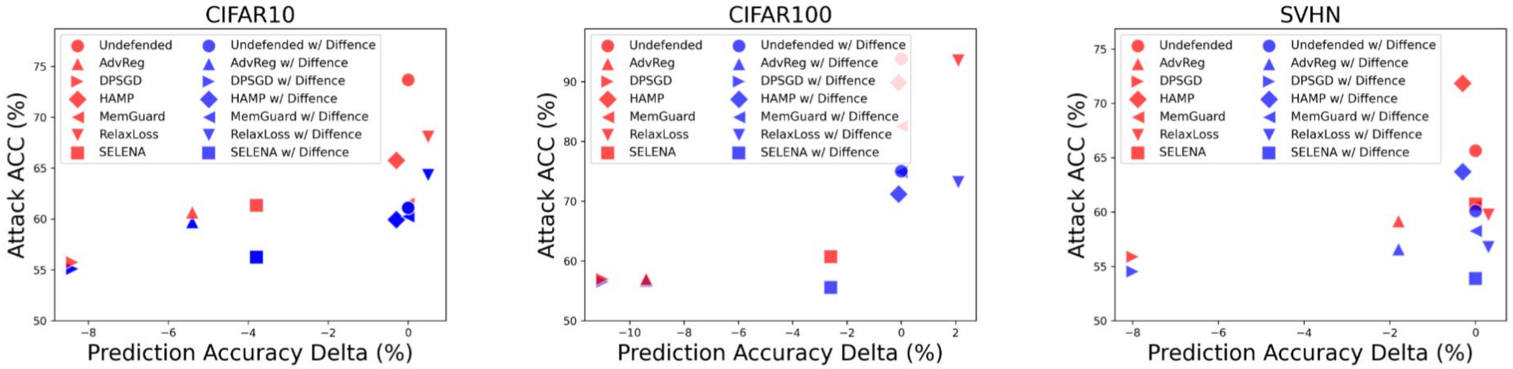
TABLE: **Average attack AUC** (lower is better). The best (lowest) AUC under each defense is in **bold**. Columns "Δ" show how much the AUC decreases compared to "w/o DIFFENCE".

Defenses	Prediction Accuracy Delta (%)	w/o DIFFENCE (AUC %)	w/ DIFFENCE (Scenario 1)		w/ DIFFENCE (Scenario 2)		w/ DIFFENCE (Scenario 3)	
			AUC (%)	Δ (%)	AUC (%)	Δ (%)	AUC (%)	Δ (%)
Undefended	0	79.14	**68.08**	−11.06	70.79	−8.35	69.12	−10.02
SELENA	−2.13	62.22	**56.00**	−6.22	60.30	−1.92	57.81	−4.41
AdvReg	−5.53	61.32	**59.17**	−2.15	61.33	0.01	60.87	−0.45
HAMP	−0.23	78.96	**67.60**	−11.36	71.23	−7.73	69.18	−9.78
RelaxLoss	0.97	75.81	**67.13**	−8.68	69.56	−6.25	68.60	−7.21
DP-SGD	−9.13	56.61	**55.47**	−1.14	58.40	−1.79	56.60	−0.01
Memguard	0	69.53	**66.76**	−2.77	67.23	−2.30	67.48	−2.05

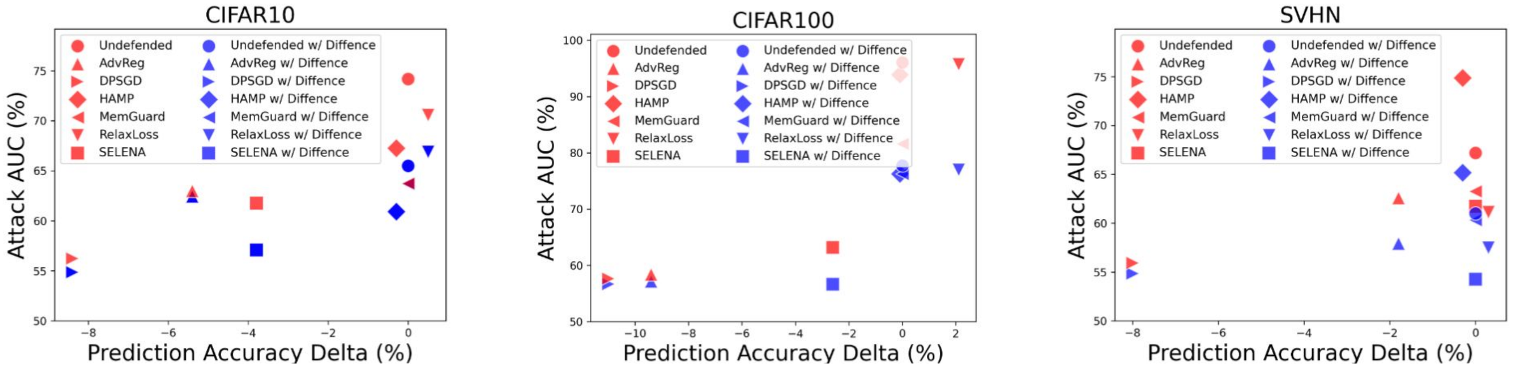
- **DIFFENCE** enhances membership privacy for both **undefended** models and models **defended with other methods**
- **DIFFENCE** is still effective when the attacker have no knowledge about the membership of any samples (Scenario 3).

University of Massachusetts Amherst

Key Results

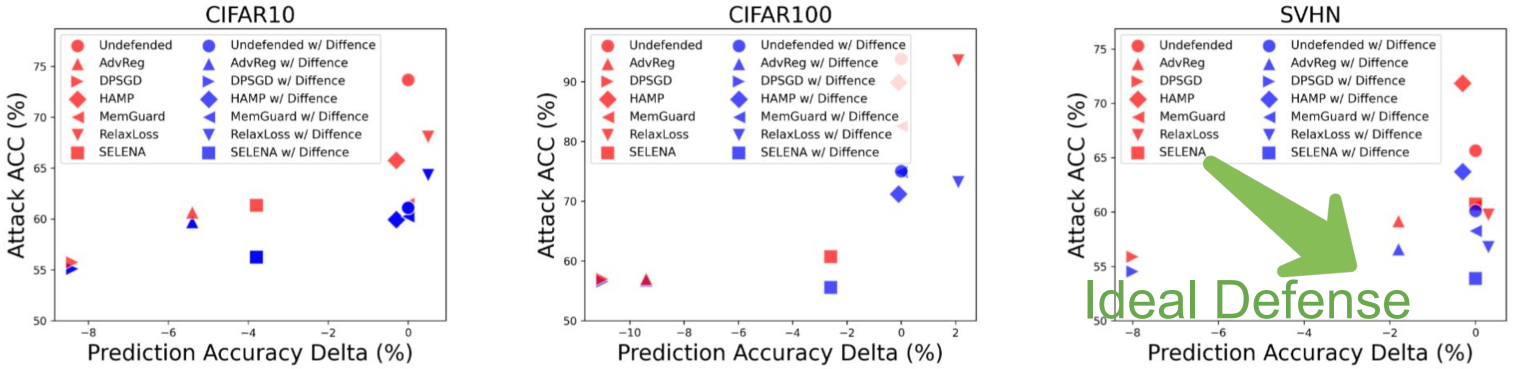


(a) Attack Accuracy

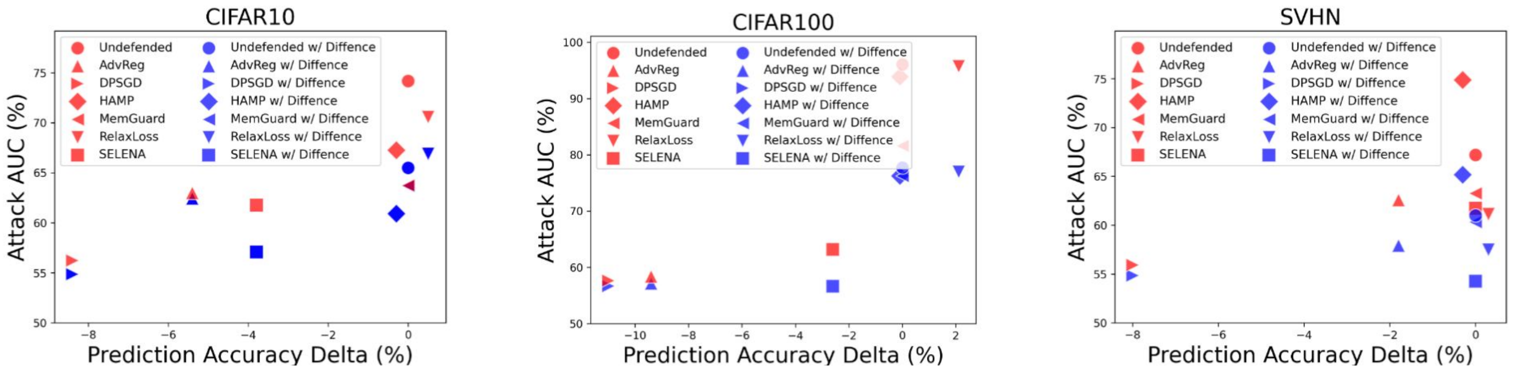


University of Massachusetts Amherst

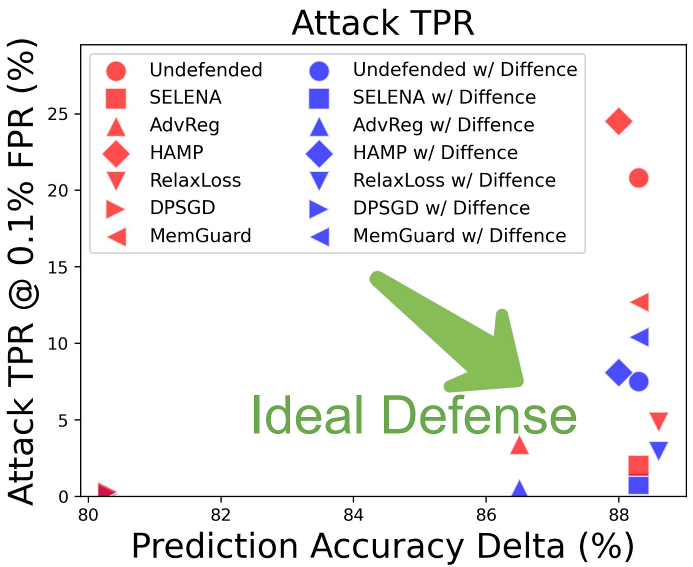
Key Results

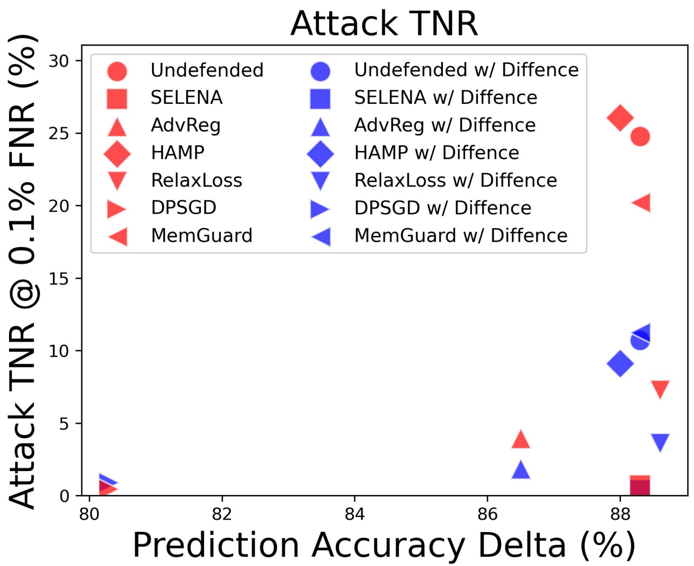


(a) Attack Accuracy



Key Results





University of Massachusetts Amherst

Takeaways

Artifact Evaluated
NDSS
Available
Functional
Reproduced

Existing MIA defenses focus on **training** or **post-inference** stages. We introduce **DIFFENCE** as a **new defense paradigm**, operating at the **pre-inference** stage

DIFFENCE is designed to **work with other defenses**. It is **plug-and-play**, requiring **no retraining**, and seamlessly integrates with all existing methods.

DIFFENCE is **Effective & Lossless**: Enhances MIA privacy **without utility loss**.

Paper

Code

University of Massachusetts Amherst

Thank You

University of Massachusetts Amherst