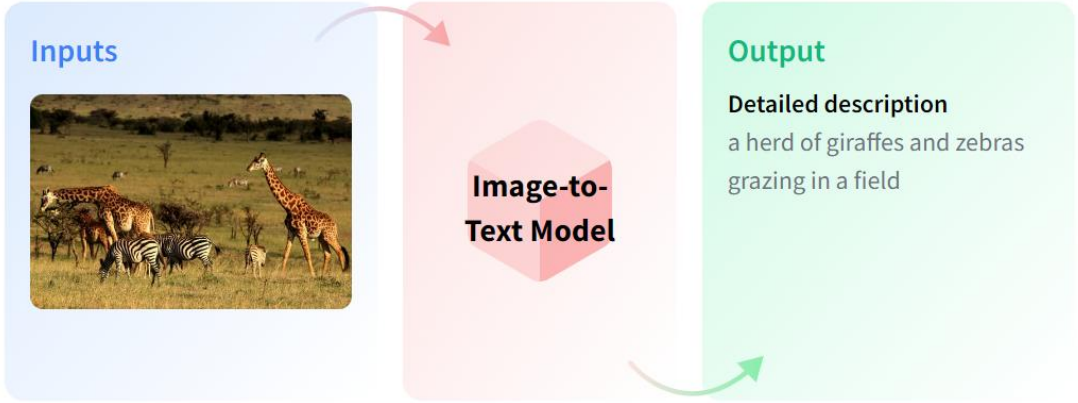
Variance-Based Membership Inference Attacks Against Large-Scale Image Captioning Models

Daniel Samira Edan Habler Yuval Elovici Asaf Shabtai
Ben-Gurion University of the Negev, Israel

Motivation

- **Membership Inference Attack (MIA)** - The main goal of MIA is to determine whether a specific sample was as a part of the model's training set
- **Image Captioning Model** - Image captioning is a challenging AI problem that involves generating a descriptive and appropriate sentence for a given image



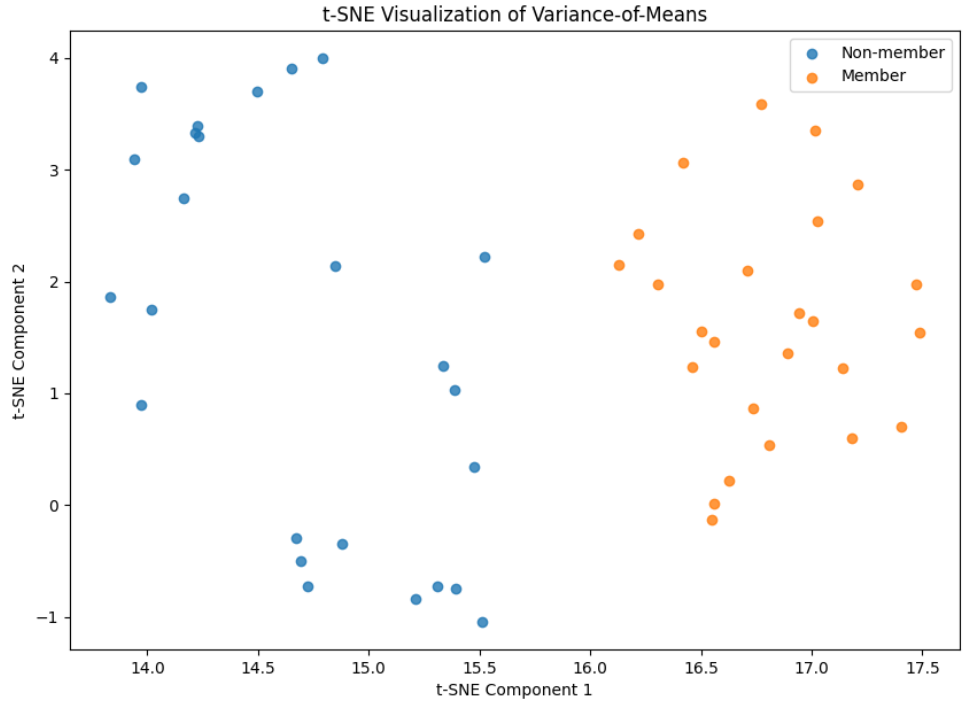
Research Purpose

- **Prove a given image-to-text generation model M_{target} was trained on my data**

- Settings:
  - Adversary's knowledge - Black-Box, Query-Only Access
  - Data – Image only
  - Task – Classification (Member or not)
  - Assumption - Access to verified non-member data D_{no}

Intuition Behind the Proposed Methods

- The **variance** in the *generated texts* for images included in the training set should be smaller than the variance for images not included in the training set



Text Generation and Multi-modal Feature Encoding

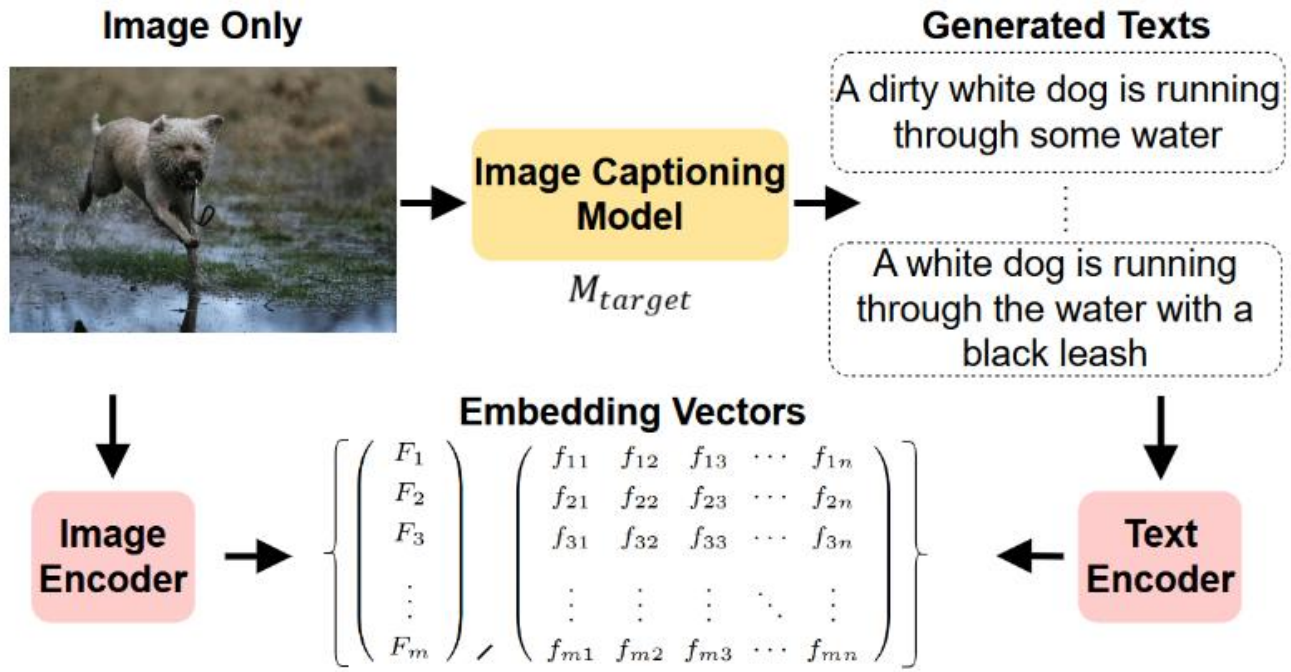


Figure 1. Overview of our method's text generation and multi-modal feature encoding stage. This stage consists of two steps: text generation, in which captions are generated using the target model, and feature encoding, in which features from multiple modalities (text and image) are encoded using pre-trained models.

Means-of-Variance (MV) Metric

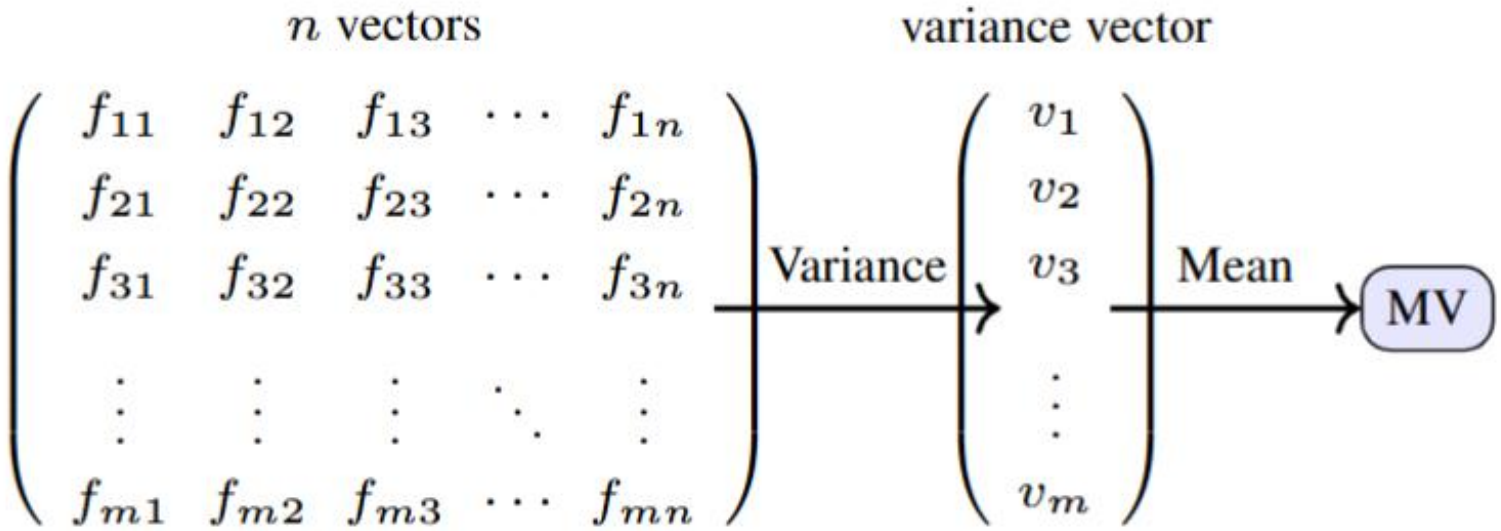
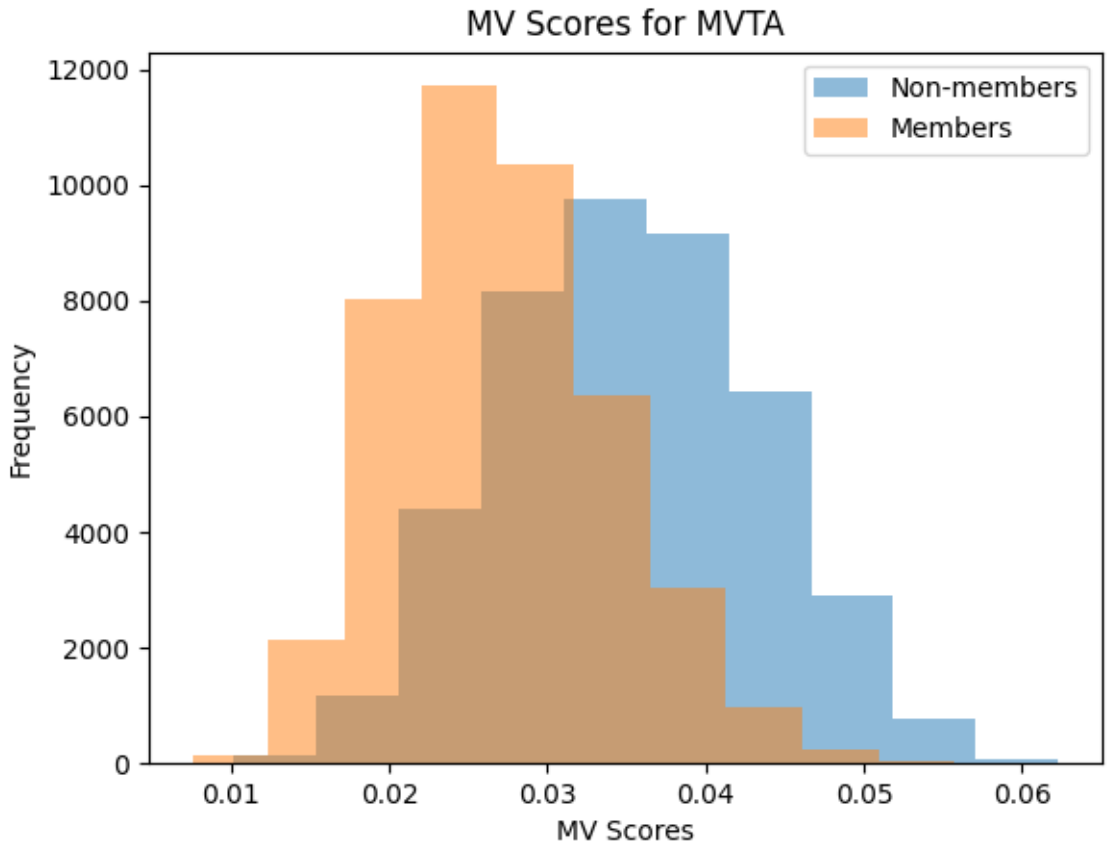


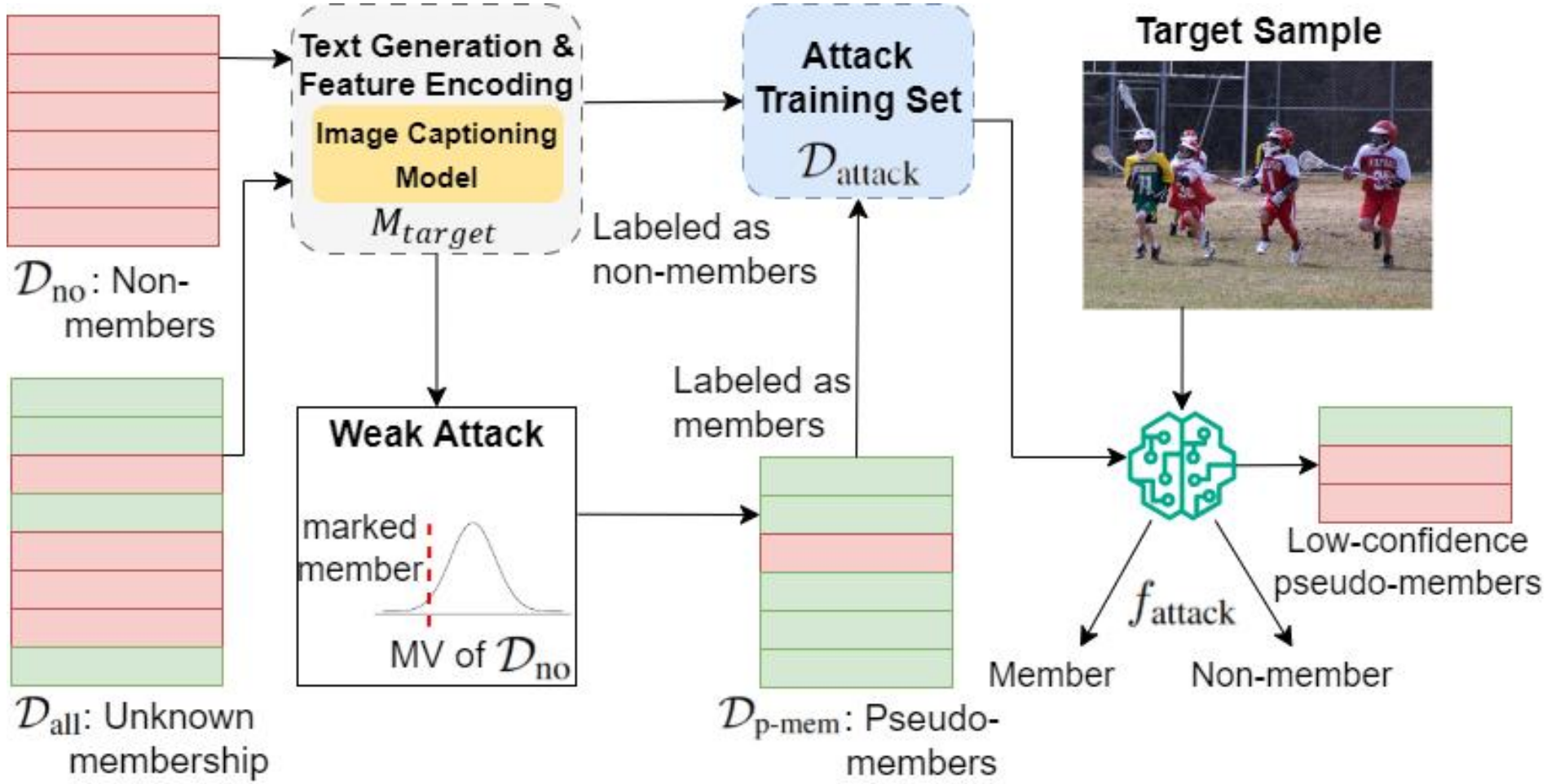
Figure 2. Calculation of the MV metric. The feature matrix (f) consists of n vectors (columns), each containing m features (rows). First, the variance for each feature across the n vectors is calculated to form the variance vector (v); then, the mean is calculated to produce the final MV score.

Means-of-Variance Threshold Attack (MVTA)

$$\text{Predicted Membership} = \begin{cases} \text{Member} & \text{if } MV < \tau \\ \text{Non-Member} & \text{otherwise} \end{cases}$$



Confidence-Based Weakly Supervised Attack (C-WSA)



Evaluation Settings

- **Member Datasets** – MSCOCO | Textcaps
- **Non-member Datasets** - Flickr30k | NoCaps | IAPR TC-12
- **Metrics** – Accuracy(ACC)| AUC | TPR at low FPR
- **Target Model** – BLIP Large | ViT-GPT2 | GIT Base
- **Membership Classifier** – MLP
- **Text Encoder** – clip_base
- **Image Encoder** – clip_base

Results - Public Models

Table 2. Evaluation of attack performance on different public target models (BLIP, ViT-GPT2, and GIT) across the MSCOCO and TextCaps datasets. We bold the best results for each knowledge setting.

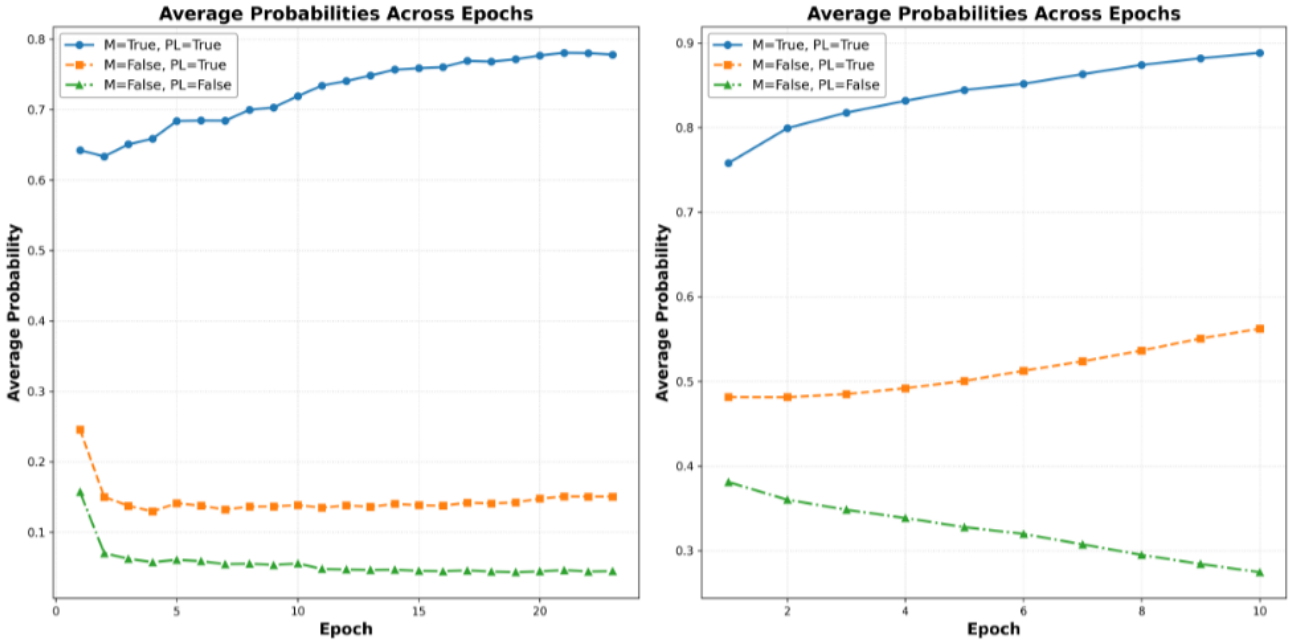
Dataset [Model]	MSCOCO [BLIP]			MSCOCO [ViT-GPT2]			MSCOCO [GIT]			Textcaps [GIT]		
Method	AUC	ACC	TPR@FPR=1%	AUC	ACC	TPR@FPR=1%	AUC	ACC	TPR@FPR=1%	AUC	ACC	TPR@FPR=1%
CSA	0.5444	0.5358	0.0087	0.7207	0.6630	0.0466	0.5942	0.6265	0.0197	0.4484	0.5010	0.0110
MVTA	**0.6601**	**0.6183**	**0.0224**	**0.7727**	**0.7026**	**0.0690**	**0.7630**	**0.7011**	**0.0441**	**0.5026**	**0.5124**	**0.0083**
WSA	0.8398	0.7914	0.3144	0.8931	0.8415	0.3758	0.8626	0.8034	0.3328	0.8165	0.7945	0.3773
C-WSA	**0.9131**	**0.8345**	**0.3826**	**0.9335**	**0.8619**	**0.4251**	**0.9255**	**0.8520**	**0.4075**	**0.8967**	**0.8479**	**0.4152**

Results - Fine-tuning on Public Pre-trained Models

Table 4. Evaluation of attack performance on the publicly available BLIP model fine-tuned by us on the Textcaps and Flickr30K datasets after 5 and 10 epochs.

Dataset [Model]	Textcaps[BLIP]						Flickr30K[BLIP]					
Epochs	5 Epochs			10 Epochs			5 Epochs			10 Epochs		
Method	AUC	ACC	TPR@FPR=1%	AUC	ACC	TPR@FPR=1%	AUC	ACC	TPR@FPR=1%	AUC	ACC	TPR@FPR=1%
CSA	0.5474	0.5373	**0.0401**	0.5565	0.5424	**0.0417**	0.6200	0.5872	0.0169	0.6403	0.5996	0.0190
MVTA	**0.6120**	**0.5826**	0.0250	**0.6482**	**0.6080**	0.0346	**0.6245**	**0.5899**	**0.0196**	**0.6612**	**0.6185**	**0.0242**
WSA	0.8483	0.8255	0.4721	0.8441	0.8182	0.4478	0.9074	0.8513	**0.4893**	0.9173	0.8615	0.4333
C-WSA	**0.9363**	**0.8820**	**0.5269**	**0.9419**	**0.8953**	**0.5560**	**0.9456**	**0.8869**	0.4450	**0.9610**	**0.9050**	**0.5436**

Confidence-Threshold



(a) C-WSA with threshold. (b) WSA without threshold.

Figure 5. Comparison of C-WSA with a confidence threshold and WSA without one, depicting the average probability of D_{p-mem} across the training epochs. M denotes membership status, and PL represents the pseudo-label assignment.

End of Presentation

Thank you