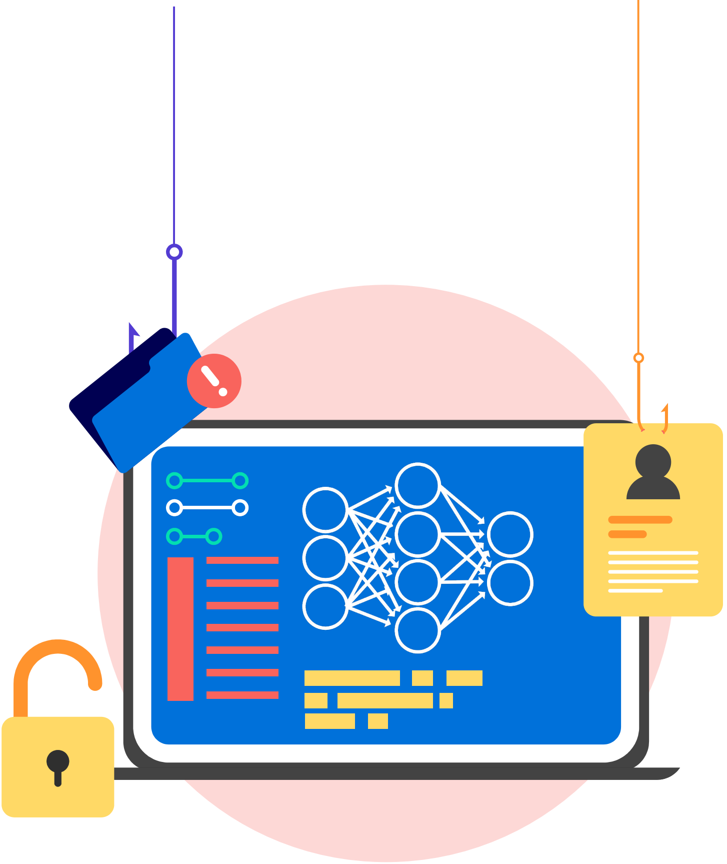
Privacy Preserving Machine Learning:

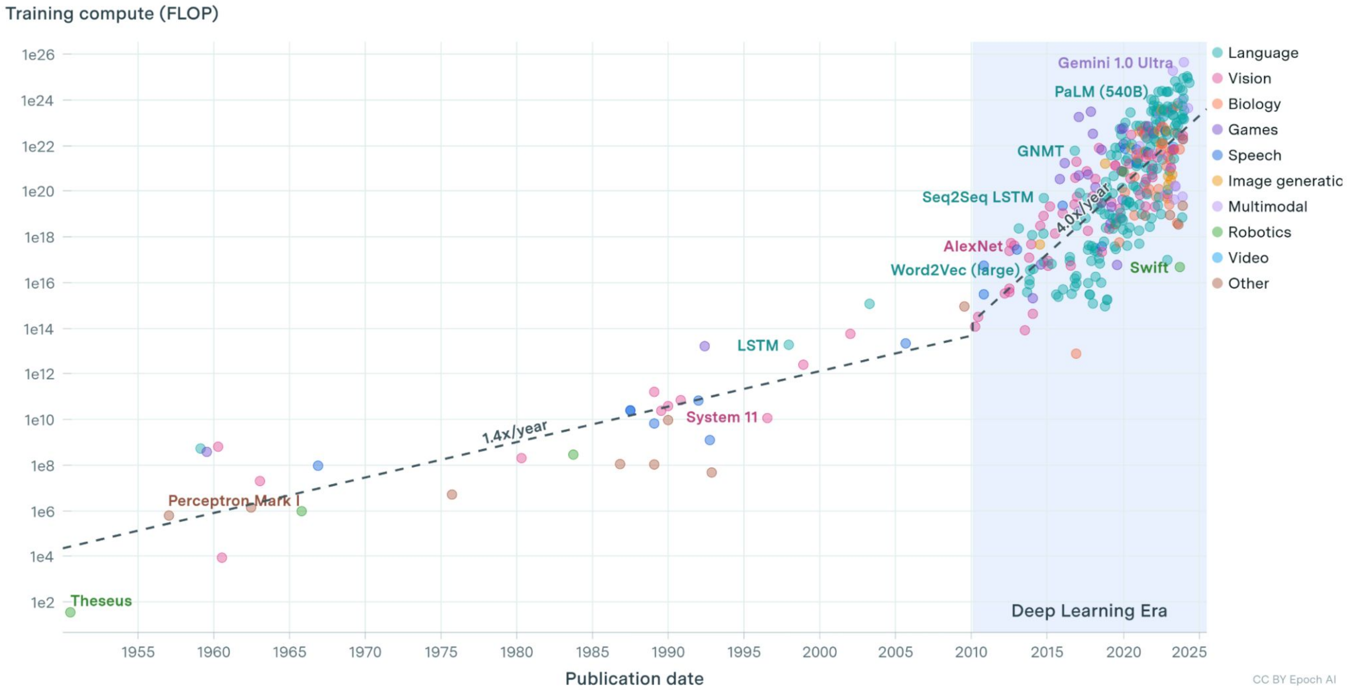
Utilizing Image Datasets With and Without Consent



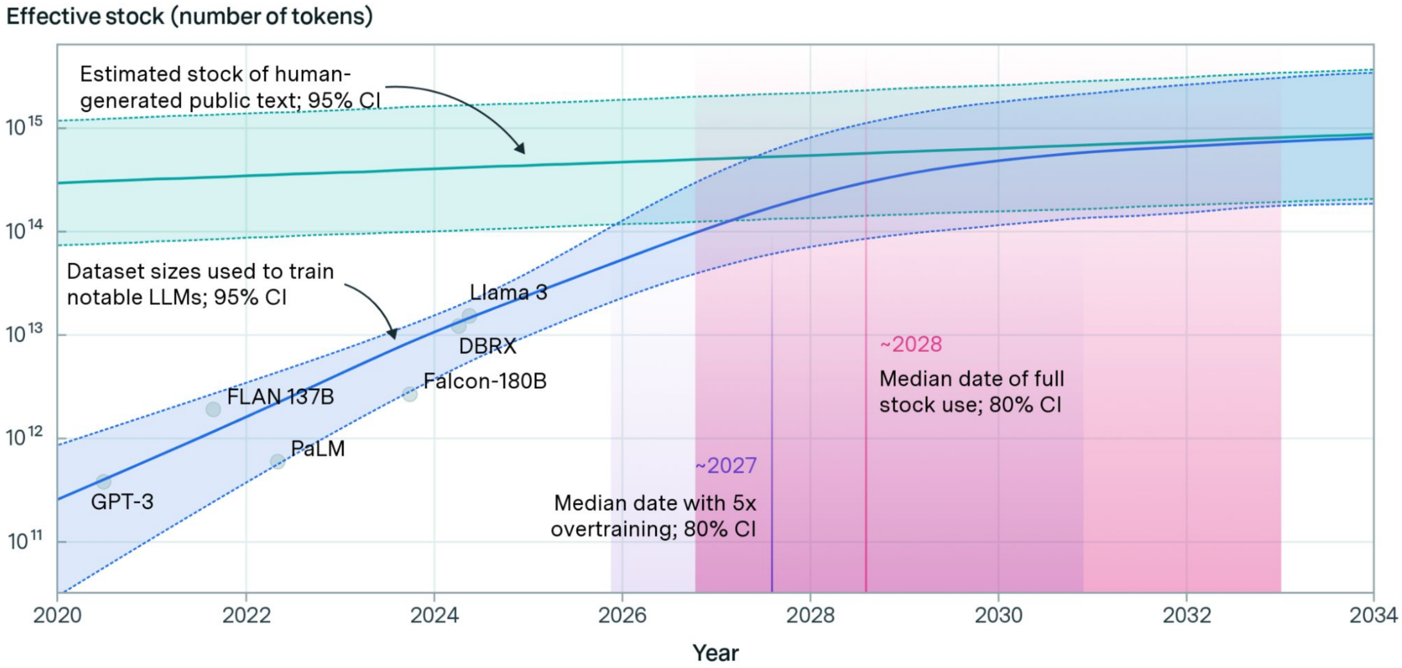
Kartik Patwari

March 10, 2025

ML Performance and Data Scaling Trends



ML Performance and Data Scaling Trends



Pretraining Datasets!

1.4 M Images

400 M
Images-Text Pairs

Data Privacy in AI

Google hit with lawsuit alleging it stole data from millions of users to train its AI tools

By Catherine Thorbecke, CNN
Updated 8:48 AM EDT, Wed July 12, 2023

Source: CNN Business

TECH

Google Exposed User Data, Feared Repercussions of Disclosing to Public

Google opted not to disclose to users its discovery of a bug that gave outside developers access to private data. It found no evidence of misuse.

Source: The Wall Street Journal

Artificial intelligence (AI)

'I didn't give permission': Do AI's backers care about data law breaches?

Regulators around world are cracking down on content being hoovered up by ChatGPT, Stable Diffusion and others

Alex Hern *and* Dan Milmo
Mon 10 Apr 2023 10.10 BST

Source: The Guardian

AI and Privacy: The privacy concerns surrounding AI, its potential impact on personal data

ET Online • Last Updated: Apr 25, 2023, 08:31 PM IST

SHARE FONT SIZE SAVE PRINT COMMENT

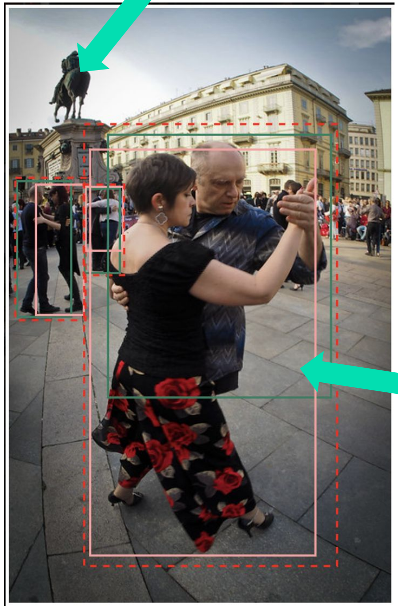
Source: The Economic Times

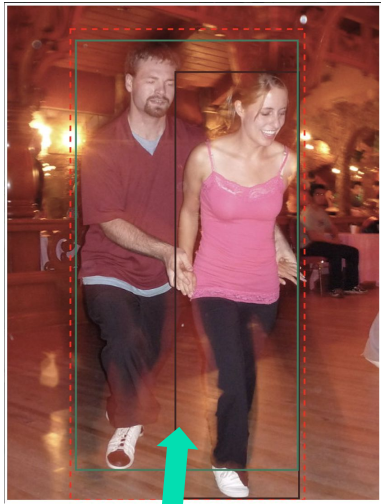
General Data Protection Regulation

CONSUMER PRIVACY ACT CALIFORNIA CCPA

Images Contain Rich Private/Personal Data!

Location





Identifiers

Occupation

Identity &
Relations

Images from Google's OpenImages Dataset!
Scraped without consent

Data Rights, Access, and Privacy Regulation Compliance

Restricted Data Access

How can we train models without access to data?

Data-Free Learning!

Data Access without Consent

How can we make data without consent usable?

Image Anonymization!

PPML:
Utilizing Image Datasets With and Without Consent

Table of Contents

RCL: Empowering Source-Free Domain Adaptation via MLLM-Guided Reliability-Based Curriculum Learning

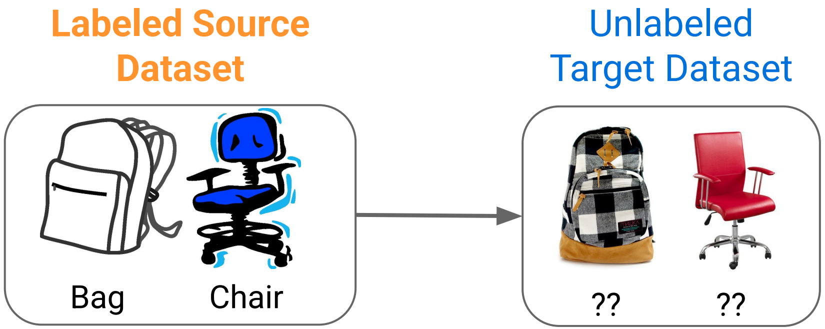
Dongjie Chen*, **Kartik Patwari***, Xiaoguang Zhu, Zhengfeng Lai, Samson Cheung, Chen-Nee Chuah

Under Submission

Background: UDA and SFDA

Unsupervised Domain Adaptation (UDA)

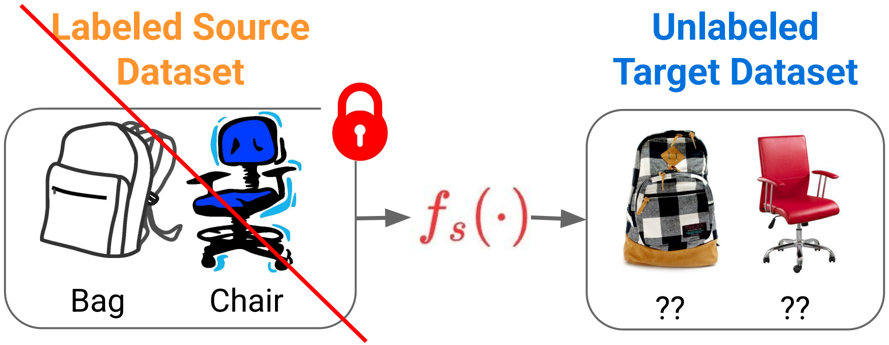
Transfer of knowledge from a **labeled source domain** to an **unlabeled target domain** under domain-shift



Source-Free Domain Adaptation (SFDA)

Transfer of knowledge from a **pre-trained source model** to an **unlabeled target domain** under domain-shift **without access to source data**.

- Source data unavailable
- Privacy or copyright reasons



Background: UDA and SFDA

Unsupervised Domain Adaptation (UDA)

$$\mathcal{D}_S = \{(\mathbf{x}_i^S, y_i^S)\}_{i=1}^N + \mathcal{D}_T = \{(\mathbf{x}_i^T)\}_{i=1}^M$$

$$\downarrow$$

$$f_{\mathrm{T}}(x)$$

Source-Free Domain Adaptation (SFDA)

$$\mathcal{D}_S = \{(\mathbf{x}_i^S, y_i^S)\}_{i=1}^N$$

$$\downarrow \text{ Source Training}$$

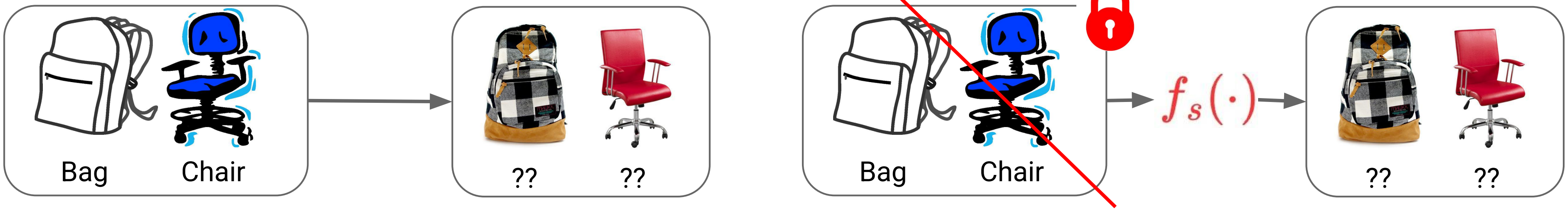
$$f_S(\mathbf{x}) + \mathcal{D}_T = \{(\mathbf{x}_i^T)\}_{i=1}^M$$

$$\downarrow \text{ Target Adaptation}$$

$$f_{\mathrm{T}}(x)$$

$$p_S(X,Y) \neq p_T(X,Y)$$

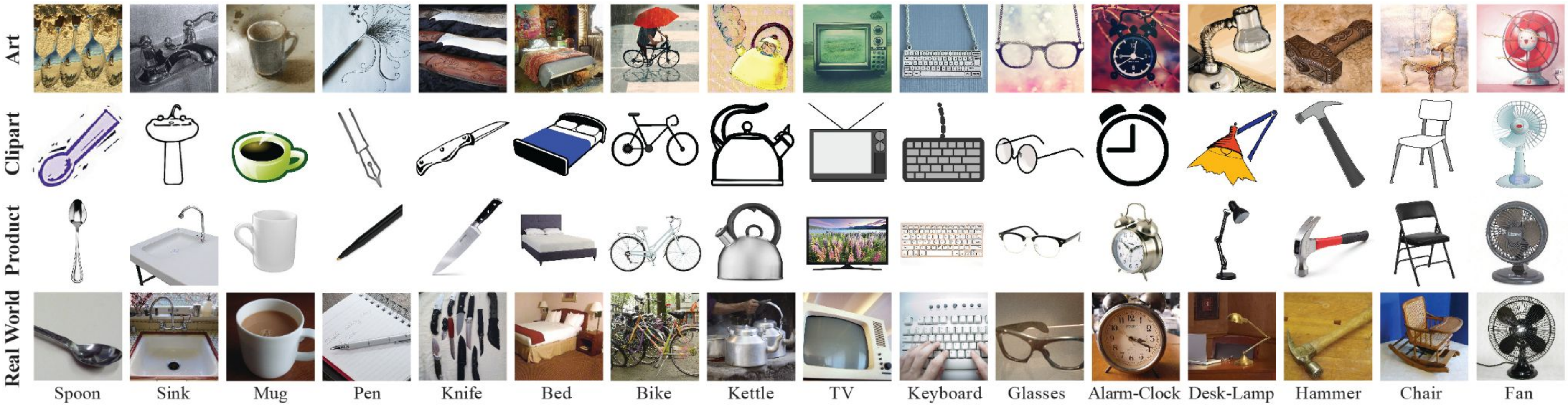
Labeled Source Dataset | Unlabeled Target Dataset | ~~Labeled Source Dataset~~ | Unlabeled Target Dataset



SFDA Datasets: Office-Home

Domains: 4 (Art, Clipart, Product, Real World)

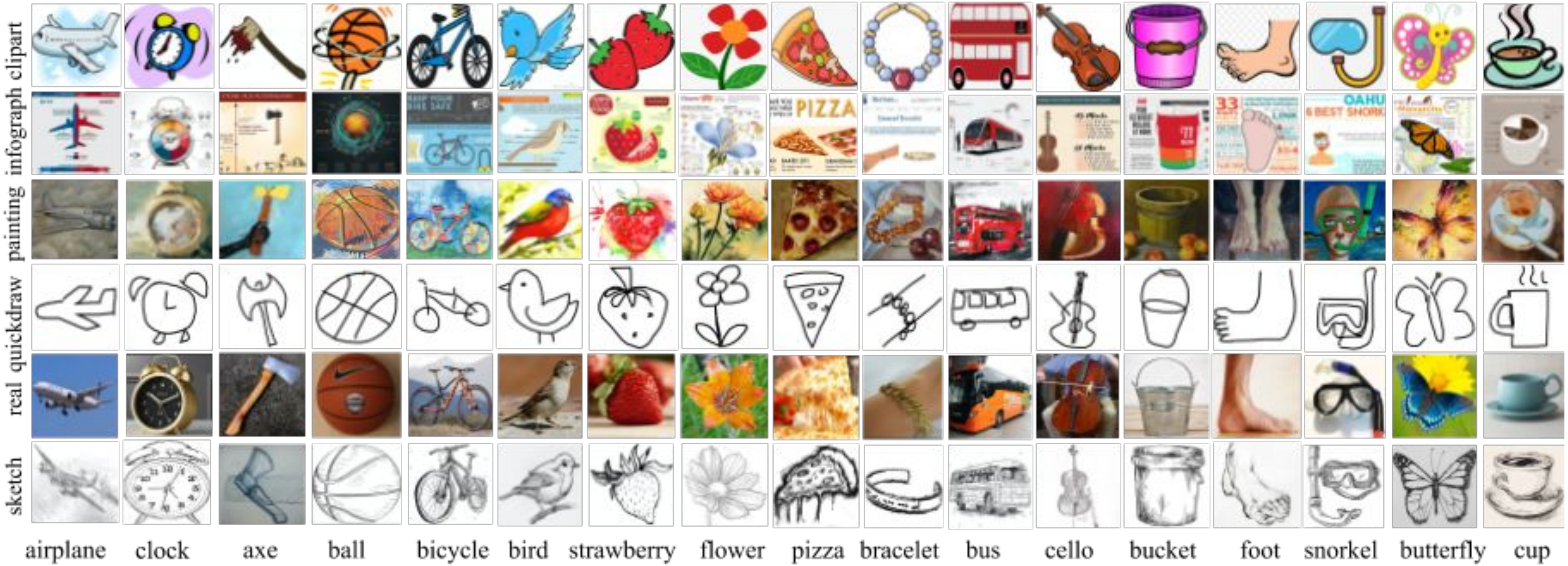
65 Classes, 15,000 Images



SFDA Datasets: DomainNet

Domains: 6 (Real, Art, Painting, Clipart, Quickdraw, Infograph)

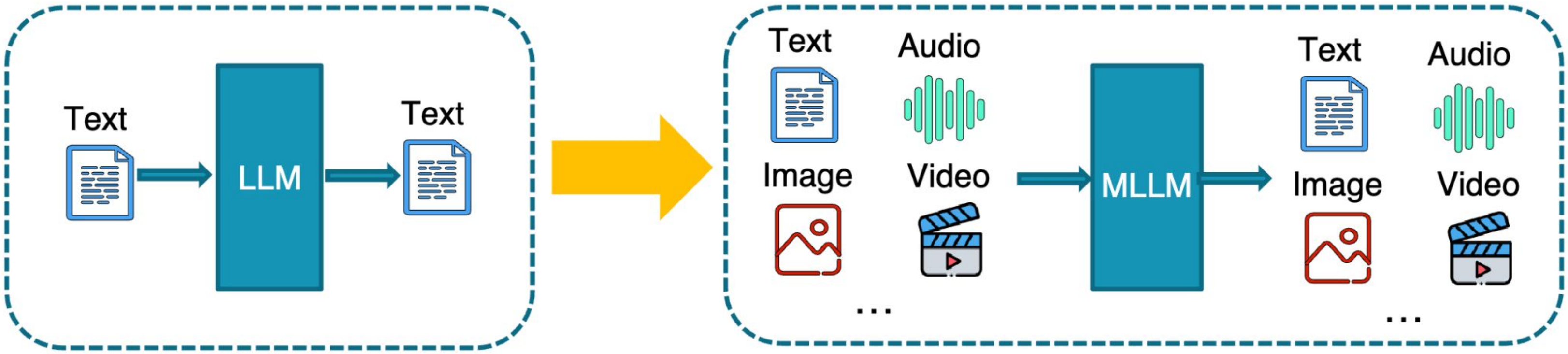
126 Classes, 596k Images



MLLMs: Background

What are MLLMs?

- Multimodal LLMs
- Enable LLMs to comprehend multimodal information



MLLMs: Visual Question Answering (VQA)

VQA: Computer vision task that involves answering questions about an image

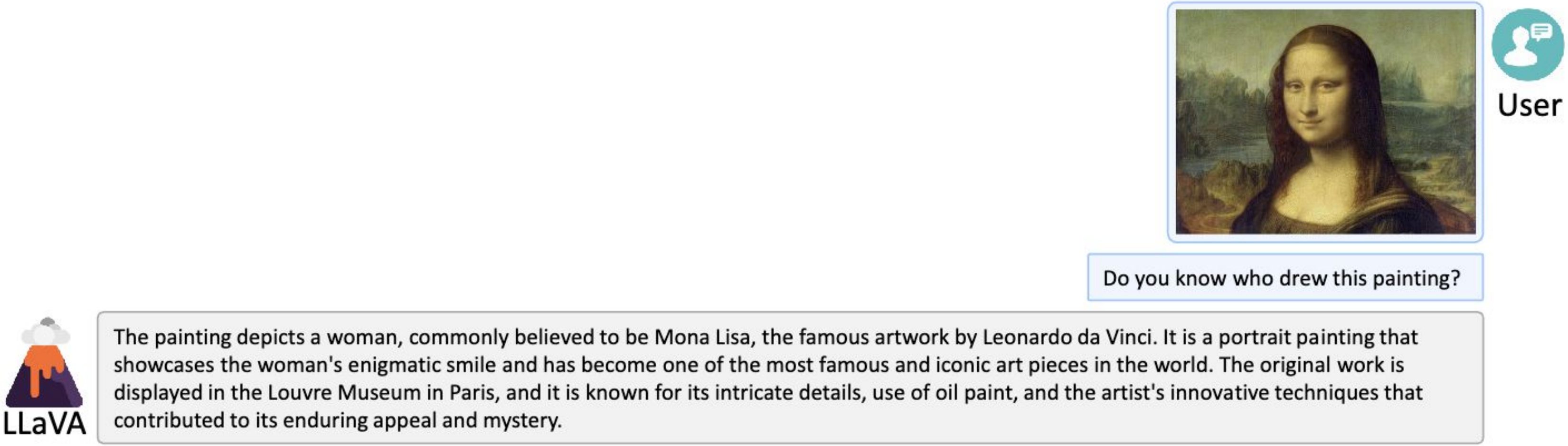


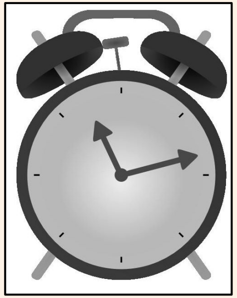
Image Classification as VQA with MLLMs

Input Text Prompt:

```
"What is the closest name from this
list to describe the object in the
image? return the name only.
{str(class_names)}"

class_names = ["Alarm_clock",
"Car", "Plane", … ]
```

Input Image:



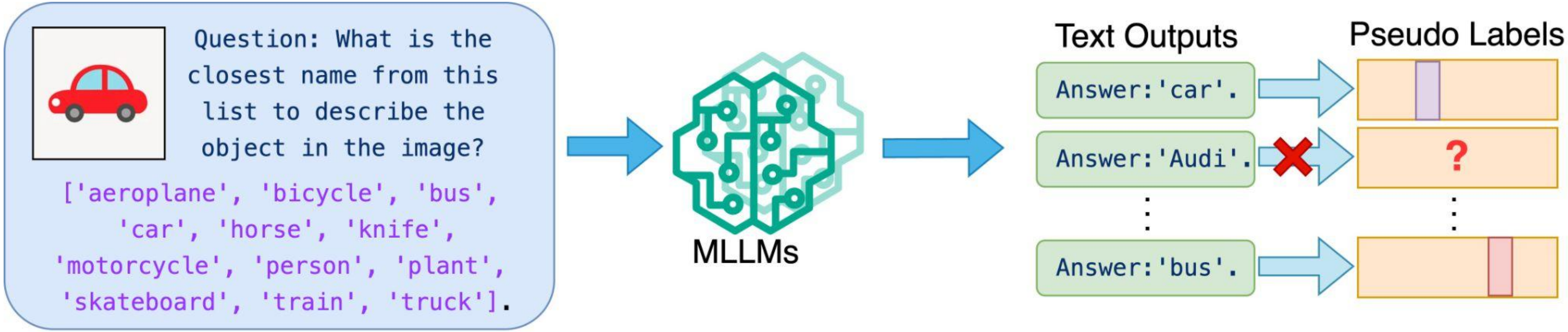
MLLM

Output:

`Alarm_clock`

Image Classification as VQA with MLLMs

Issue: MLLM do not always follow prompts!!

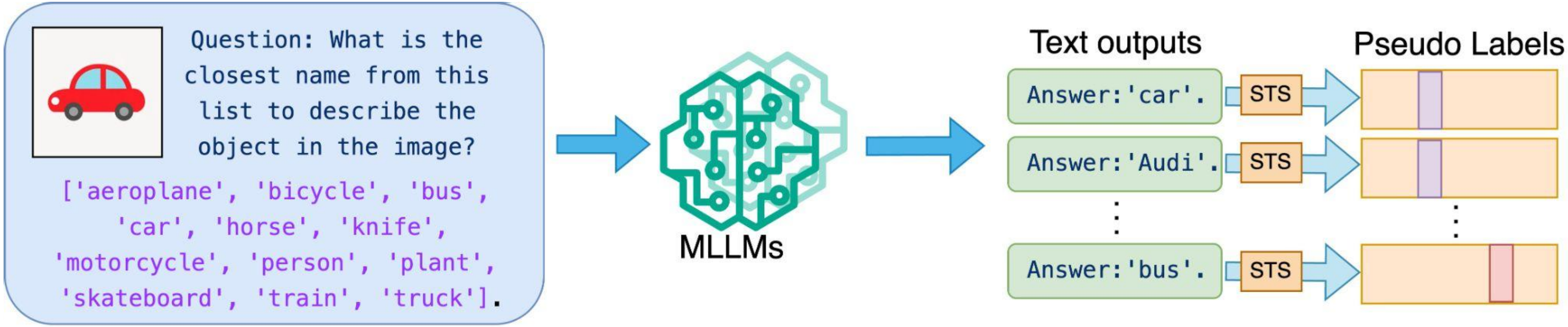


Semantic Text Similarity (STS)

Solution: Proposed STS!

$$\hat{y}^{mi} = \operatorname*{argmax}_{c} \mathbf{STS}(T_m^i, T_t^c),$$

$$\mathrm{STS}(T_1, T_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|_2 \|\mathbf{v}_2\|_2} - 1,$$

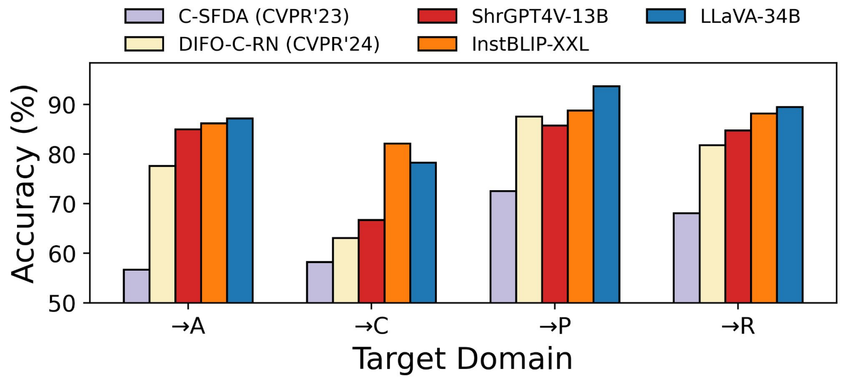


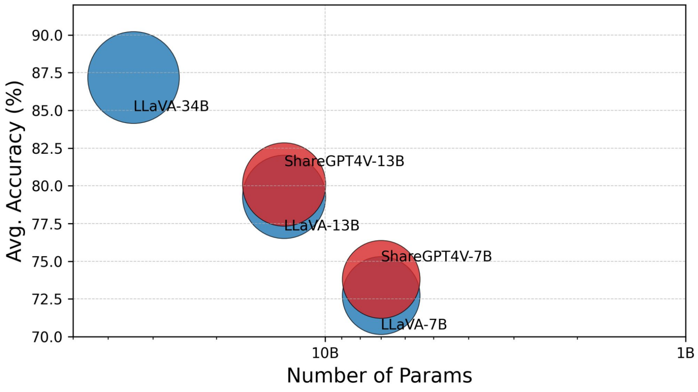
MLLM with STS

MLLMs: ShareGPT4V-13B, InstructBLIP-XXL, LLaVA-34B

↳ Zero-Shot with STS already beats SOTA SFDA!

Issues: (1) MLLMs are large! (2) Inconsistency between MLLMs

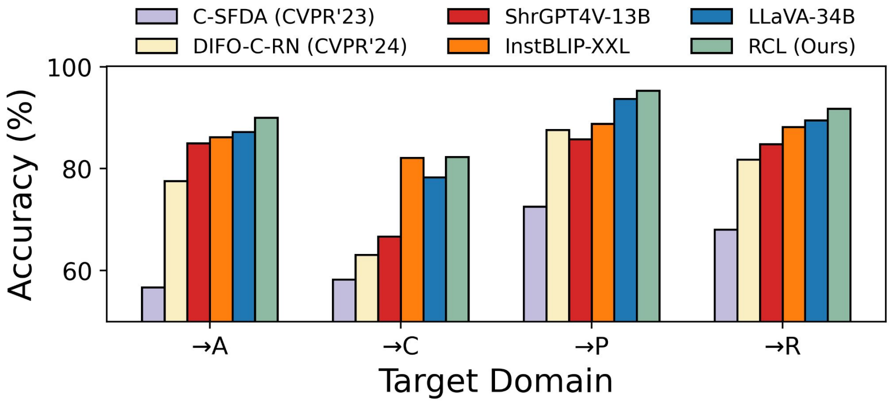


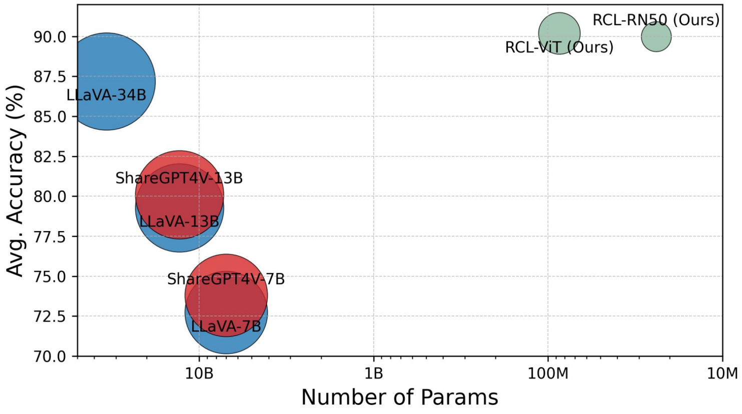


Reliability-based Curriculum Learning (RCL)

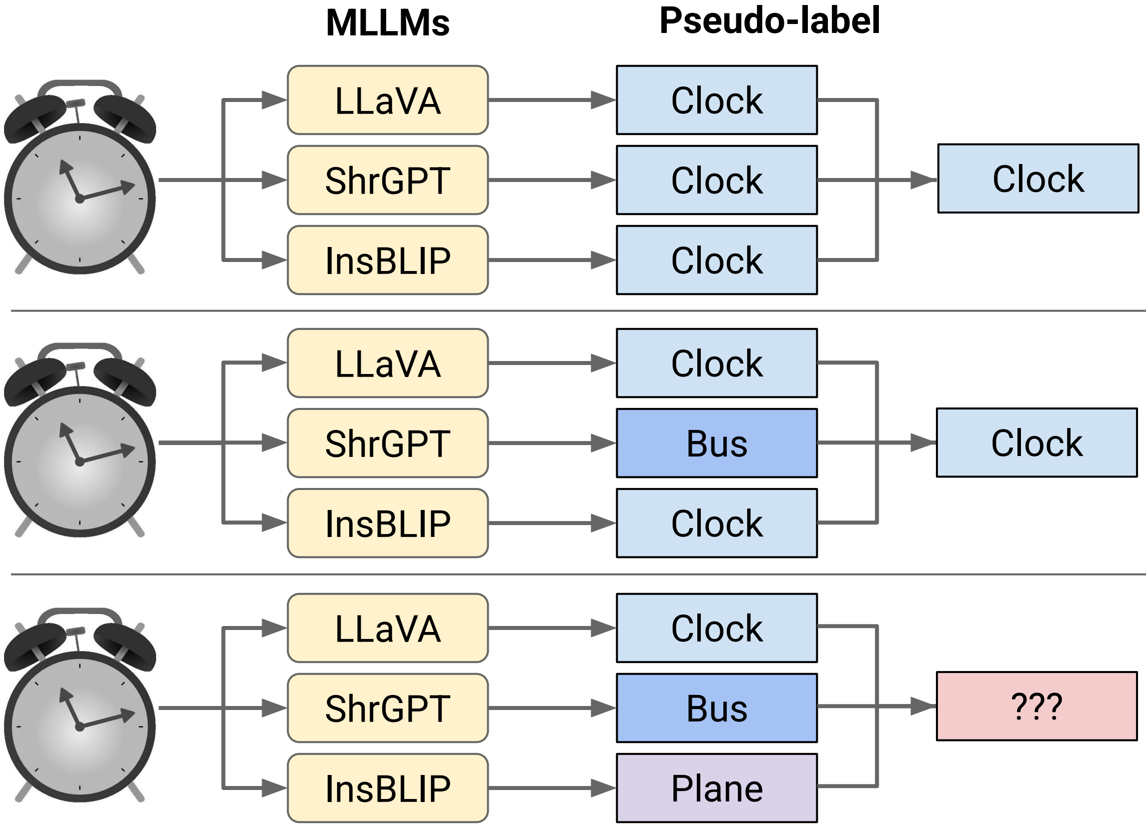
Issue 1: MLLMs are large!	**Solution:** Knowledge Distillation (KD)
Issue 2: Inconsistency between MLLMs	**Solution:** Multi-Teacher KD (MTKD)

→ **RCL** uses MLLMs for MTKD with Consensus-based Reliability and Curriculum Learning





Consensus-based Reliability Measurement



Full Agreement → Reliable

$$\mathcal{D}_{\mathrm{R}} : \mathcal{R}(x_t^i) = 1$$

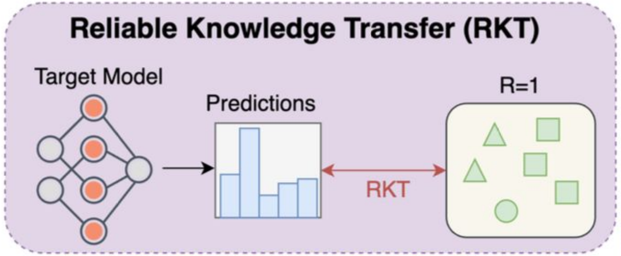
Partial Agreement → Less Reliable

$$\mathcal{D}_{\mathrm{LR}} : 0 < \mathcal{R}(x_t^i) < 1$$

No Agreement → Un-Reliable

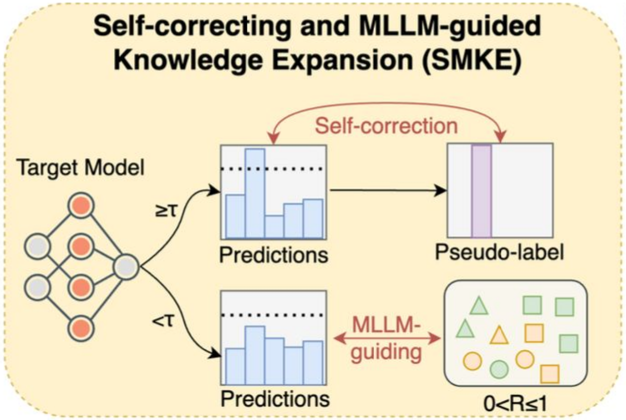
$$\mathcal{D}_{\mathrm{UR}} : \mathcal{R}(x_t^i) = 0$$

Stage 1: Reliable Knowledge Transfer



$$\mathcal{L}_{RKT} = -\frac{1}{|\mathcal{D}_R|} \sum_{(x_r^i, y_r^i) \in \mathcal{D}_R} y_r^i \cdot \log f_{\theta_t}(x_r^i),$$

Stage 2: Self-Correcting and MLLM-guided Knowledge Expansion



$$\mathcal{L}_{\text{SMKE}} = -\frac{1}{|\mathcal{D}_R \cup \mathcal{D}_{LR}|} \sum_{x_t^i \in \{\mathcal{D}_R \cup \mathcal{D}_{LR}\}} \tilde{y}^i \cdot \log f_{\theta_t}(x_t^i),$$

Stage 3: Multi-Hot Masking Refinement

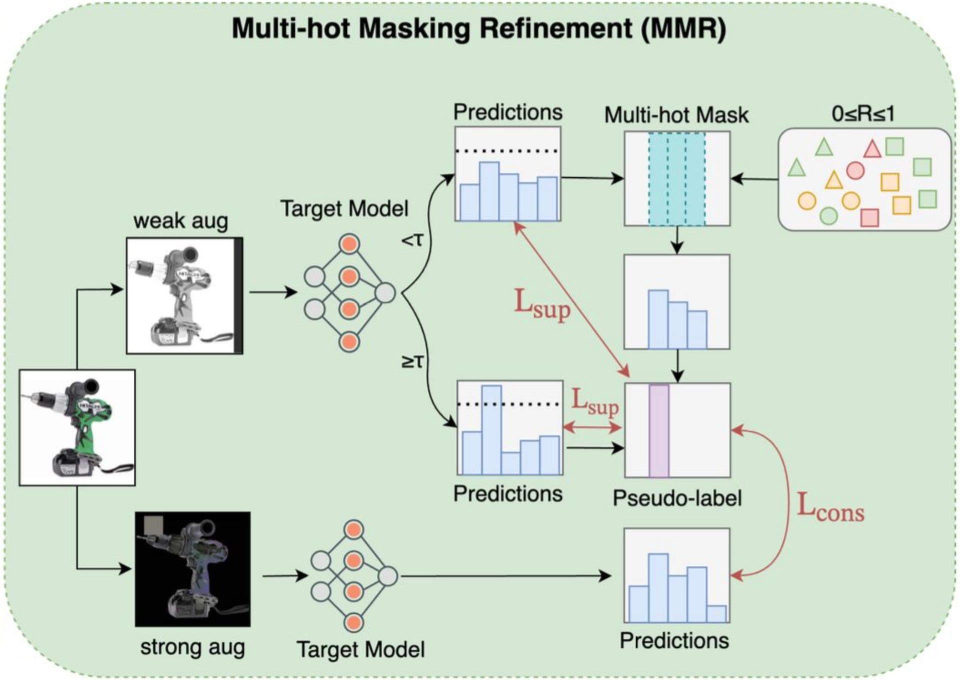
$$\mathcal{D} = \mathcal{D}_R \cup \mathcal{D}_{LR} \cup \mathcal{D}_{UR}$$

$$\tilde{y}^i = \begin{cases} \arg\max_C(\mathbf{z}_t^i), & \text{if } p_t^i \geq \tau, \\ \arg\max_C(\mathbf{z}_t^i \odot \mathbf{m}^i), & \text{if } p_t^i < \tau, \end{cases}$$

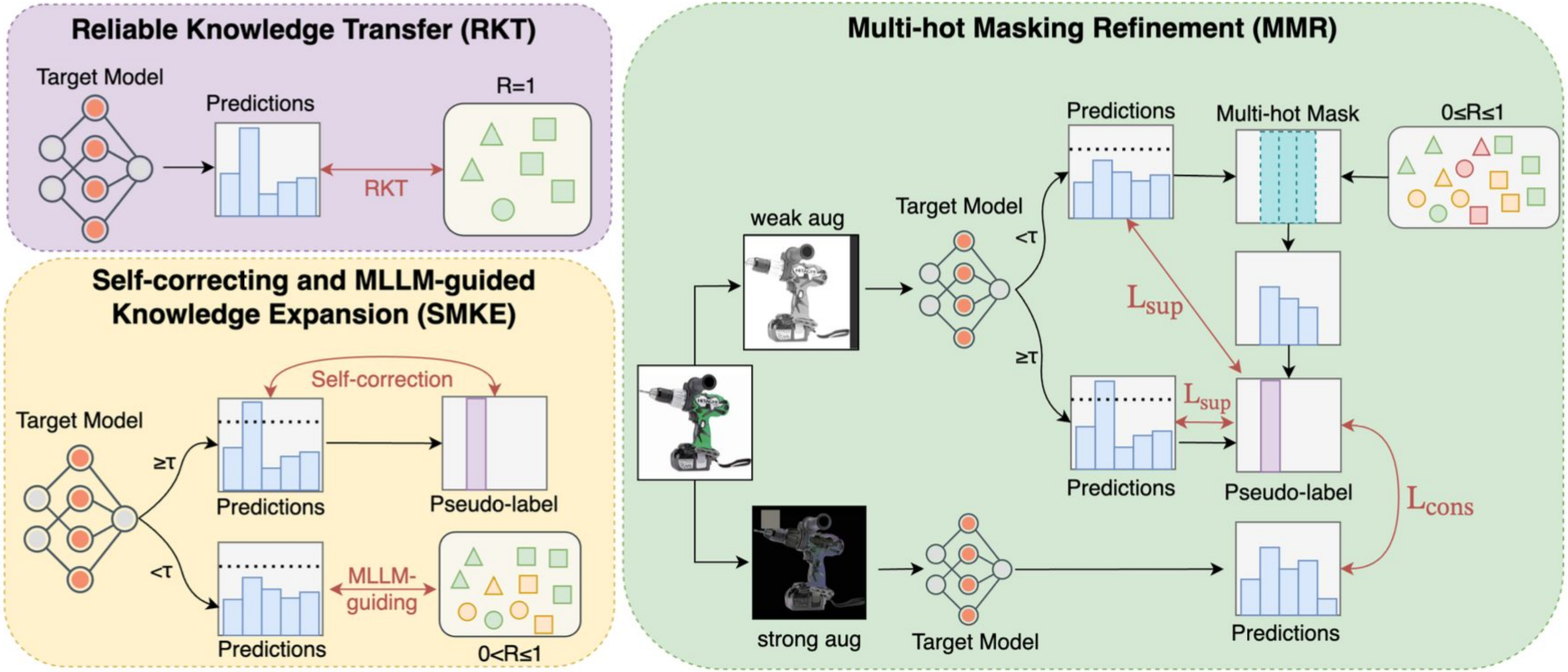
$$\mathcal{L}_{\text{sup}} = -\frac{1}{\mathcal{D}} \sum_{x_t^i \in \mathcal{D}} \tilde{y}^i \cdot \log f_{\theta_t}(x_t^i)$$

$$\mathcal{L}_{\text{cons}} = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{N_t} \mathcal{L}_{\text{CE}}(\tilde{y}^i, \mathbf{z}_{st}^i),$$

$$\mathcal{L}_{\text{MMR}} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{cons}}$$



Reliability-based Curriculum Learning (RCL)



Main Results: Office-Home

Method	SF	CP	ViT	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg.
Source	-	✗	✗	44.7	64.2	69.4	48.3	57.9	60.3	49.5	40.3	67.2	59.7	45.6	73.0	56.7
PADCLIP-RN [15]	✗	✓	✗	57.5	84.0	83.8	77.8	85.5	84.7	76.3	59.2	85.4	78.1	60.2	86.7	76.6
ADCLIP-RN [33]	✗	✓	✗	55.4	85.2	85.6	76.1	85.8	86.2	76.7	56.1	85.4	76.8	56.1	85.5	75.9
ELR [48]	✓	✗	✗	58.4	78.7	81.5	69.2	79.5	79.3	66.3	58.0	82.6	73.4	59.8	85.1	72.6
PLUE [23]	✓	✗	✗	49.1	73.5	78.2	62.9	73.5	74.5	62.2	48.3	78.6	68.6	51.8	81.5	66.9
C-SFDA [13]	✓	✗	✗	60.3	80.2	82.9	69.3	80.1	78.8	67.3	58.1	83.4	73.6	61.3	86.3	73.5
PSAT-GDA [39]	✓	✗	✓	73.1	88.1	89.2	82.1	88.8	88.9	83.0	72.0	89.6	83.3	73.7	91.3	83.6
DIFO-C-RN [41]	✓	✓	✗	62.6	87.5	87.1	79.5	87.9	87.4	78.3	63.4	88.1	80.0	63.3	87.7	79.4
DIFO-C-B32 [41]	✓	✓	✓	70.6	90.6	88.8	82.5	90.6	88.8	80.9	70.1	88.9	83.4	70.5	91.2	83.1
CLIP-RN [30]*	-	✓	✗	51.7	85.0	83.7	69.3	85.0	83.7	69.3	51.7	83.7	69.3	51.7	85.0	72.4
LLaVA-34B (w/ STS) [25]*	-	✓	✓	78.3	93.7	89.5	87.0	93.7	89.5	87.0	78.3	89.5	87.0	78.3	93.7	87.2
InstBLIP-XXL (w/ STS) [4]*	-	✓	✓	82.0	91.6	88.8	82.2	91.6	88.8	82.2	82.0	88.8	82.2	82.0	91.6	86.2
ShrGPT4V-13B (w/ STS) [2]*	-	✓	✓	66.7	85.8	84.8	83.2	85.8	84.8	83.2	66.7	84.8	83.2	66.7	85.8	80.1
RCL (Ours)	✓	✗	✗	82.5	95.3	**93.3**	89.1	**95.3**	**92.7**	**89.3**	**82.4**	92.8	89.4	82.1	95.4	90.0
RCL-ViT (Ours)	✓	✗	✓	**83.1**	**95.7**	93.1	**89.2**	**95.3**	92.6	89.2	82.3	**92.9**	**90.0**	**83.2**	**95.5**	**90.2**

Table 1: Accuracy (%) on Office-Home dataset.

Main Results

Method	SF	CP	ViT	DomainNet													VisDA
				C→P	C→R	C→S	P→C	P→R	P→S	R→C	R→P	R→S	S→C	S→P	S→R	Avg.	S→R
Source	-	✗	✗	42.6	53.7	51.9	52.9	66.7	51.6	49.1	56.8	43.9	60.9	48.6	53.2	52.7	45.3
DAPL-RN [7]	✗	✓	✗	72.4	87.6	65.9	72.7	87.6	65.6	73.2	72.4	66.2	73.8	72.9	87.8	74.8	86.9
ADCLIP-RN [15]	✗	✓	✗	71.7	88.1	66.0	73.2	86.9	65.2	73.6	73.0	68.4	72.3	74.2	89.3	75.2	88.5
PLUE [23]	✓	✗	✗	59.8	74.0	56.0	61.6	78.5	57.9	61.6	65.9	53.8	67.5	64.3	76.0	64.7	88.3
TPDS [38]	✓	✗	✗	62.9	77.1	59.8	65.6	79.0	61.5	66.4	67.0	58.2	68.6	64.3	75.3	67.1	87.6
DIFO-C-RN [41]	✓	✓	✗	73.8	89.0	69.4	74.0	88.7	70.1	74.8	74.6	69.6	74.7	74.3	88.0	76.7	88.8
DIFO-C-B32 [41]	✓	✓	✓	76.6	87.2	74.9	80.0	87.4	75.6	80.8	77.3	75.5	80.5	76.7	87.3	80.0	90.3
LLaVA-34B (w/ STS) [25]*	-	✓	✓	84.4	91.0	83.7	85.5	91.0	83.7	85.5	84.4	83.7	85.5	84.4	91.0	86.1	92.1
InstBLIP-XXL (w/ STS) [4]*	-	✓	✓	82.5	89.0	83.0	86.7	89.0	83.0	86.7	82.5	83.0	86.7	82.5	89.0	85.3	86.7
ShrGPT4V-13B (w/ STS) [2]*	-	✓	✓	79.7	87.9	79.2	79.9	87.9	79.2	79.9	79.7	79.2	79.9	79.7	87.9	81.7	90.4
RCL (Ours)	✓	✗	✗	87.6	92.8	87.9	89.2	92.7	87.8	89.6	87.7	87.6	89.4	87.5	92.7	89.4	93.2
RCL-ViT (Ours)	✓	✗	✓	**88.1**	**93.3**	**88.0**	**89.7**	**93.3**	**88.0**	**89.7**	**88.0**	**87.8**	**89.7**	**88.1**	**93.3**	**89.7**	**93.3**

Table 2: Accuracy (%) on DomainNet and VisDA datasets.

More Results & Ablation Studies

Fig 1: Direct distillation from single MLLMs

Method	BB	Office-Home →A	→C	→P	→R	Avg.
DIFO-C-RN	RN50	79.3	63.1	87.7	87.5	79.4
DIFO-C-B32	RN50	82.3	70.4	90.8	88.3	83.1
RCL (Ours)	RN18	89.1	81.5	95.1	92.6	89.6
RCL (Ours)	RN50	**89.3**	**82.3**	**95.3**	**92.9**	**90.0**

Table 4: RCL backbone choice

RCL RKT	SMKE	MMR	Office-Home →A	→C	→P	→R	Avg.
✓	✗	✗	82.8	73.3	89.3	88.1	83.3
✓	✗	✓	87.7	80.2	93.3	92.0	88.3
✓	✓	✗	88.5	80.9	95.1	92.5	89.3
✓	✓	✓	**89.3**	**82.3**	**95.3**	**92.9**	**90.0**

Table 3: Ablation on RCL components

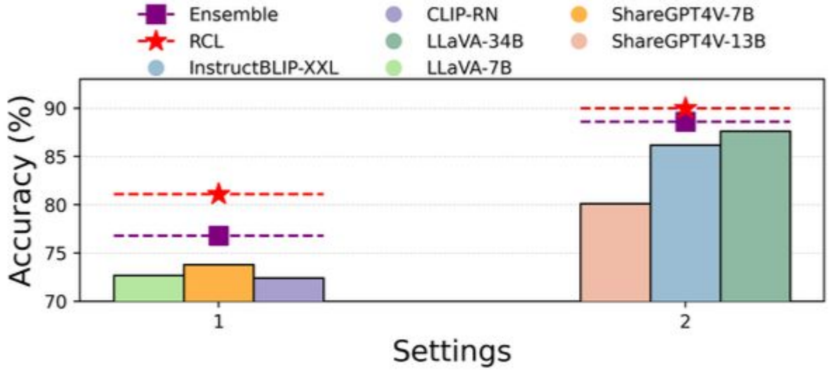


Fig 2: Choice of MLLM ensemble

PPML:

Utilizing Image Datasets With and Without Consent

Table of Contents

RefSD: Rendering-Refined Stable Diffusion for Privacy Compliant Synthetic Data

Kartik Patwari*, David Schneider*, Xiaoxiao Sun, Chen-Nee Chuah, Lingjuan Lyu, Vivek Sharma*

Under Submission

Image Anonymization

Images contain PIIs (Personal identifiable Information)

Including: People, Faces, License Plates, Text, etc.

Detect PIIs

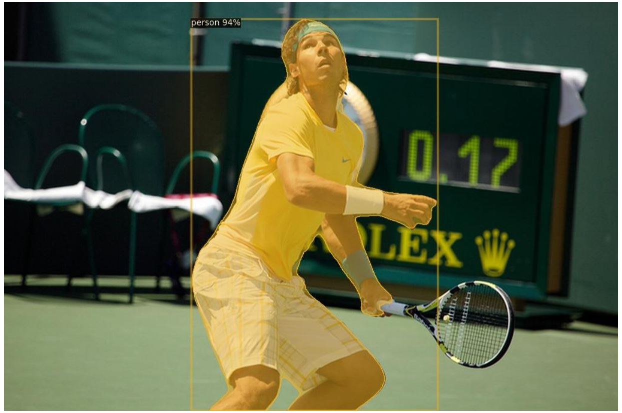
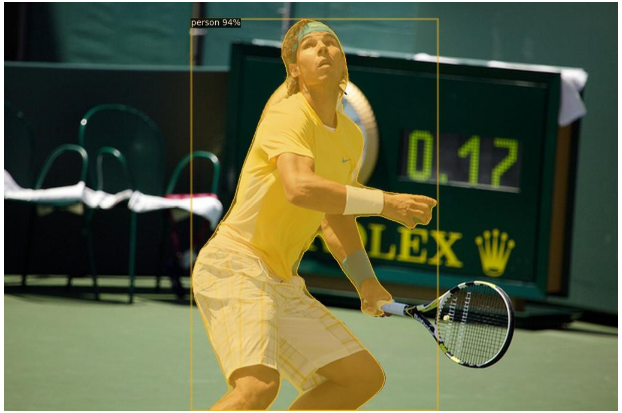


Image Anonymization

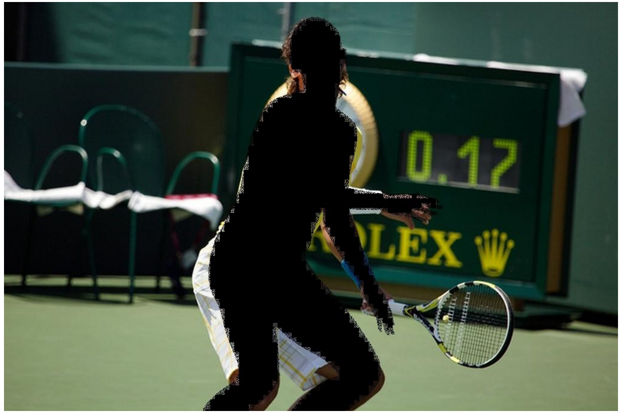
Images contain PIIs (Personal identifiable Information)

Including: People, Faces, License Plates, Text, etc.

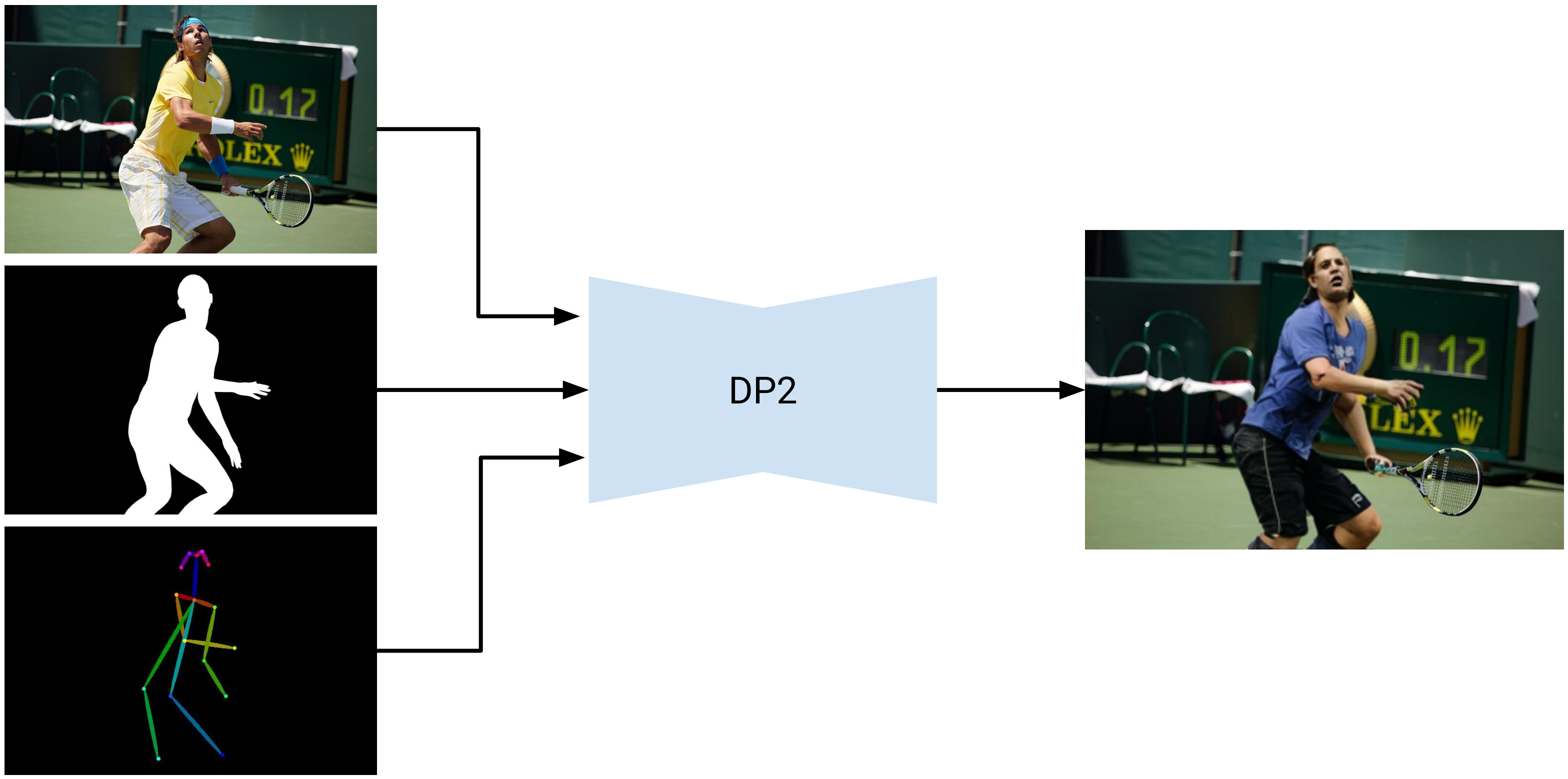
Anonymization is process of removing PIIs



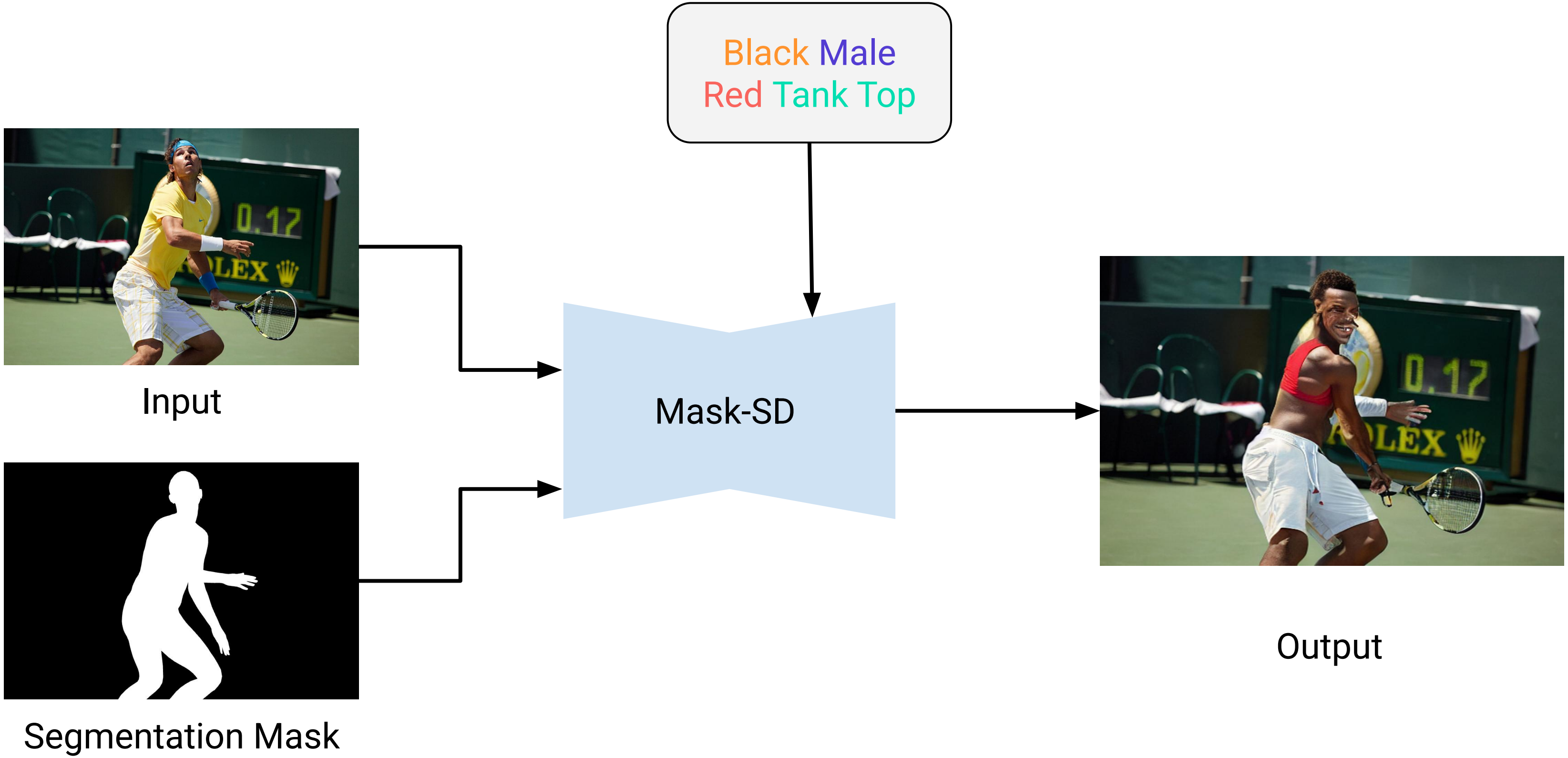
Remove PIIs



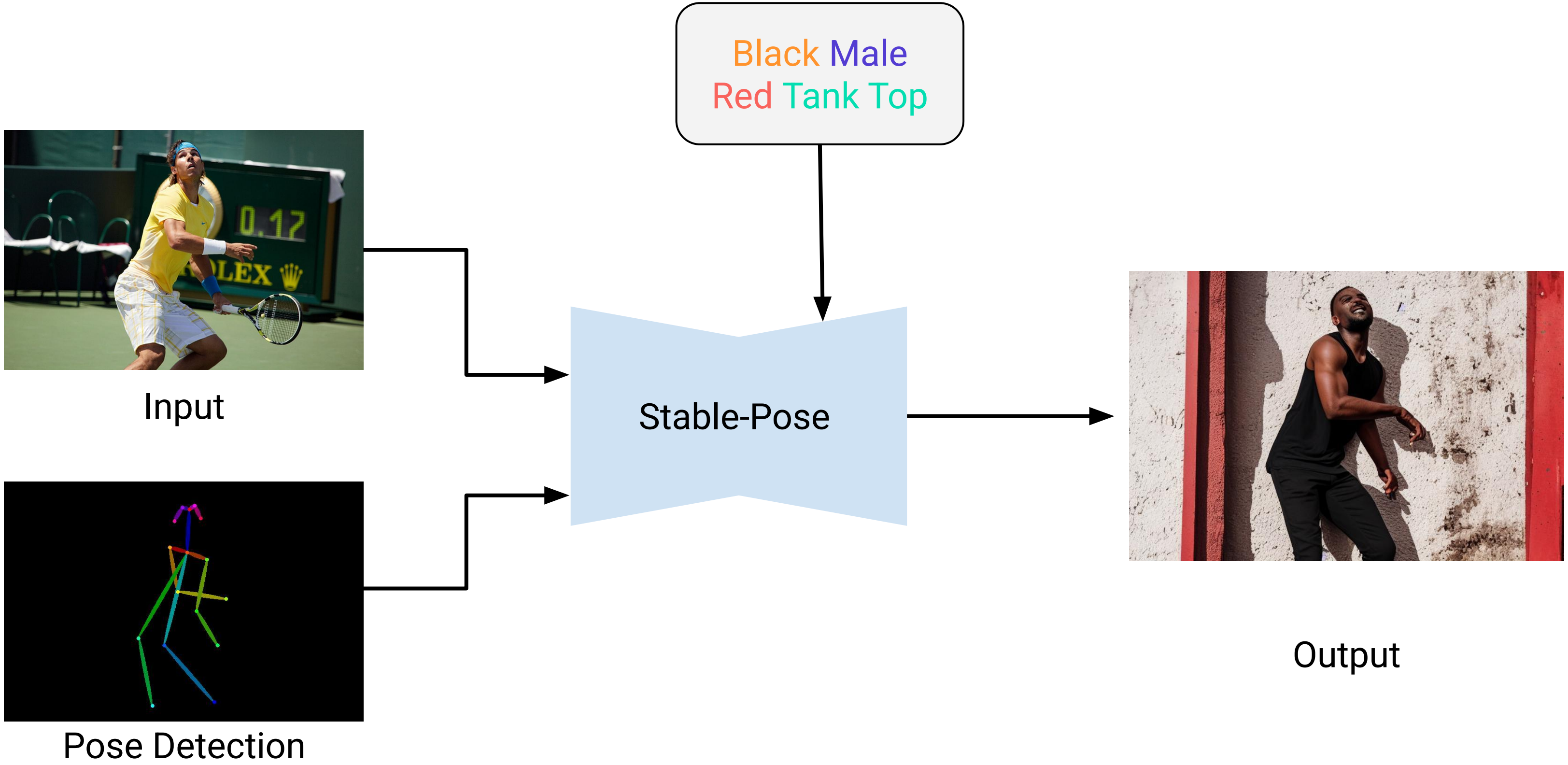
Human Synthesis: GANs (DeepPrivacy2, WACV'24)



Human Synthesis: Stable Diffusion Inpaint (ICCV'23)

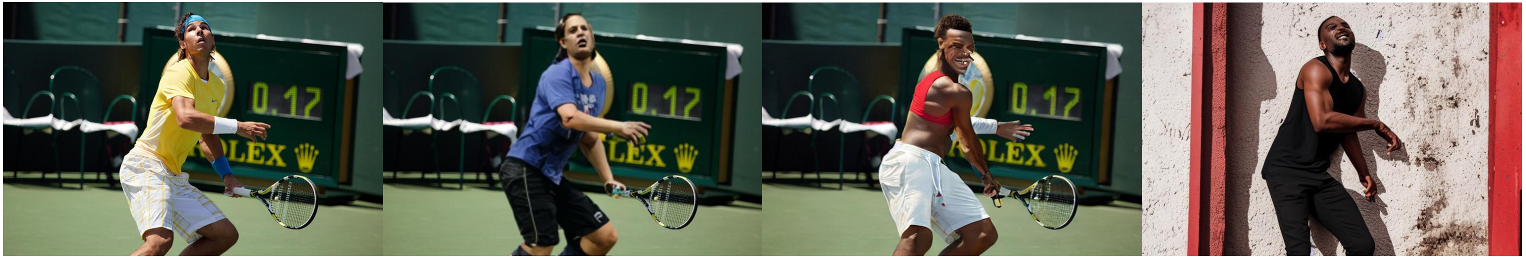


Human Synthesis: Stable Diffusion with Stable Pose (NeurIPS'24)



Human Synthesis

Original	DP2 (WACV'24)	Mask-SD (ICCV'23)	Pose-SD (NeurIPS'24)



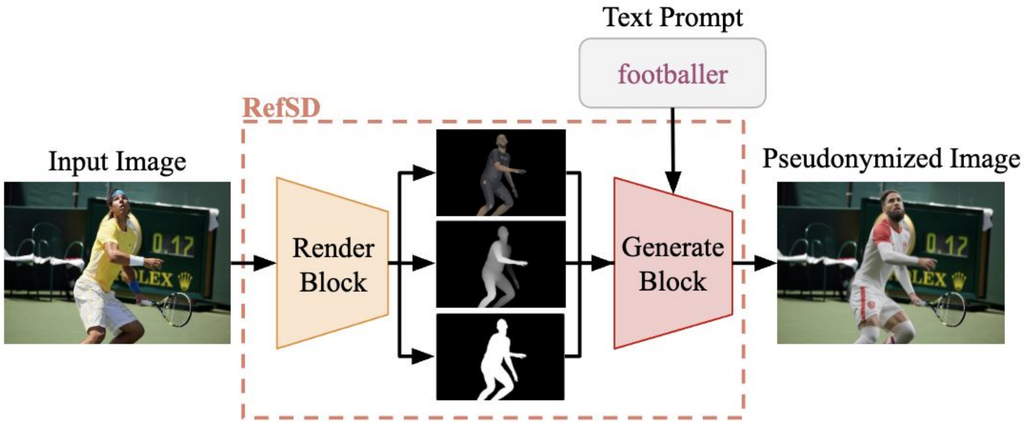
	DP2 (WACV'24)	Mask-SD (ICCV'23)	Pose-SD (NeurIPS'24)
Privacy	✓	✓	✗
Pose	✓	✗	✓
Scene Integrity	✓	✓	✗
Prompt Control	✗	✗	✗
Image Quality	✗	✗	✓

Human Synthesis

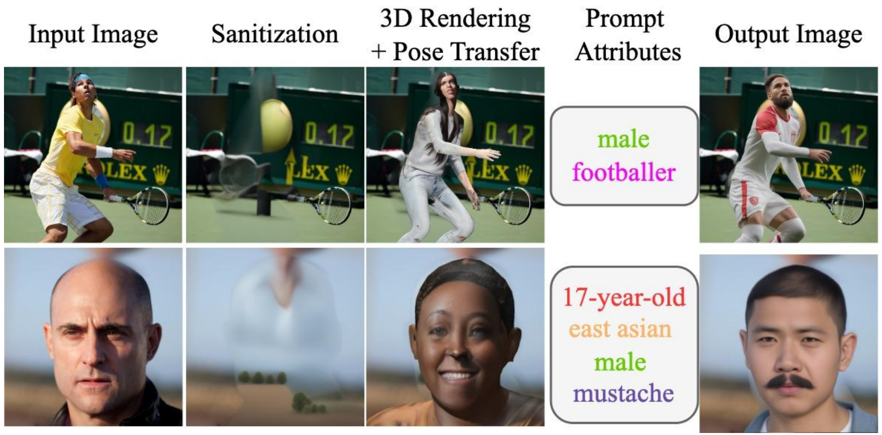
	Original	DP2 (WACV’24)	Mask-SD (ICCV’23)	Pose-SD (NeurIPS’24)	**RefSD (Ours)**
Privacy		✓	✓	✗	✓
Pose		✓	✗	✓	✓
Scene Integrity		✓	✓	✗	✓
Prompt Control		✗	✗	✗	✓
Image Quality		✗	✗	✓	✓

Rendering Refined Stable Diffusion (RefSD)

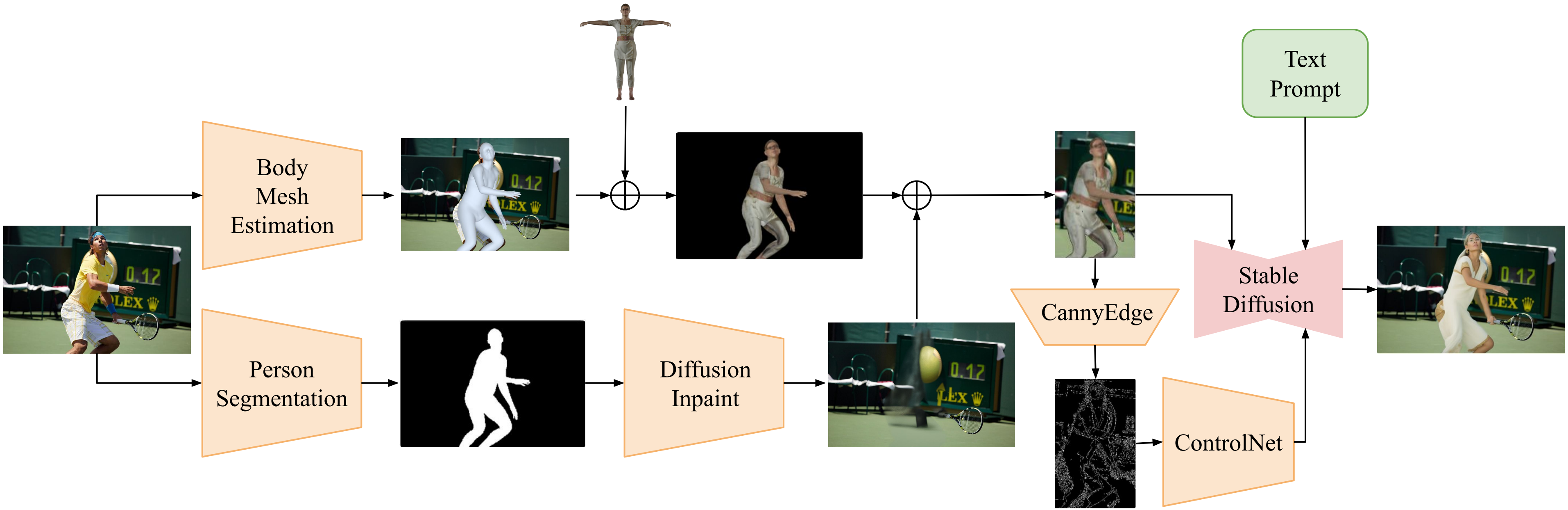
RefSD removed humans completely and replaces them with pose-aligned 3D rendered avatars.



(a) RefSD – Combines 3D human rendering and prompt-based generation



RefSD Pipeline



Prompt Design

Prefix + Attribute Prompt + Suffix

Prefix:

seen from front

seen from behind

Attribute Prompt:

A {age} {ethnicity} {gender} with {body attr}, showing {emotion} emotion.

Suffix:

The image is natural, realistic, sharp focus, high detail, medium format photograph, person, (Nikon DSLR Camera, 8K resolution, Detailed body features).

Prompt Complexity

Basic Prompt:

A {age} person.

A {ethnicity} person.

Simple Prompt:

A {age} {ethnicity} {gender} with {body attr}, showing {emotion} emotion.

Medium Prompt:

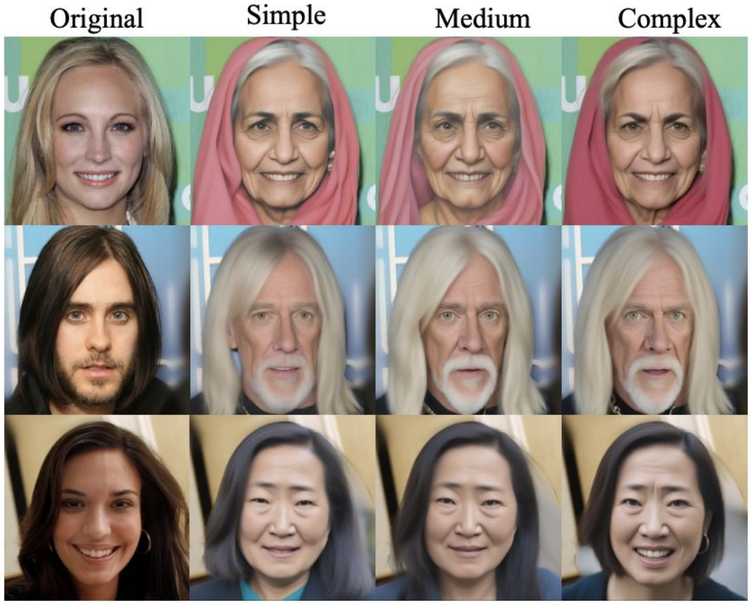
A {age} {ethnicity} {gender} with clearly {body attr}, showing exaggerated {emotion} emotion.

Complex Prompt:

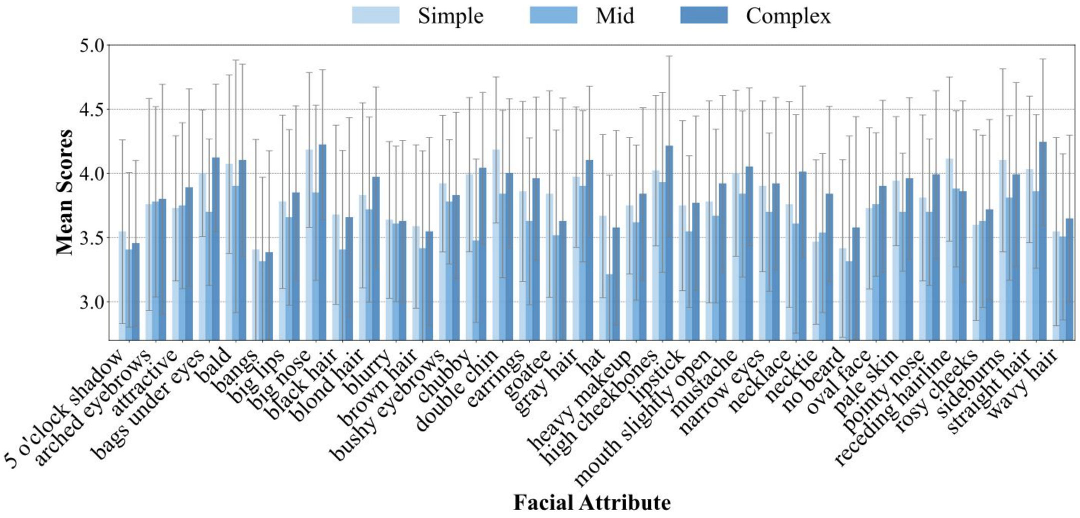
A {age} {ethnicity} {gender} with clearly {body attr}, showing exaggerated {emotion} emotion. + Suffix

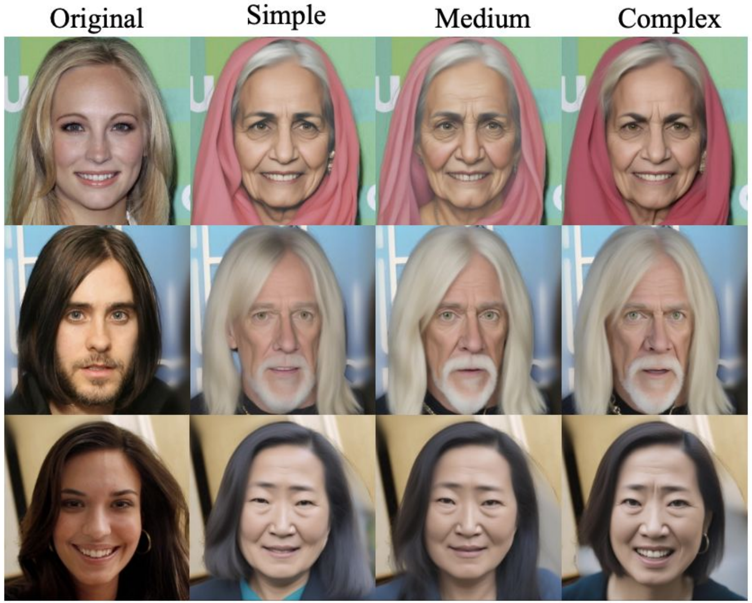
Prompt Complexity

Prompt Complexity	FID ↓	FID_{CLIP} ↓	CLIPScore ↑
Basic	22.5	18.2	0.72
Simple	19.8	16.5	0.78
Medium	21.0	17.3	0.75
Complex	20.1	16.9	0.76

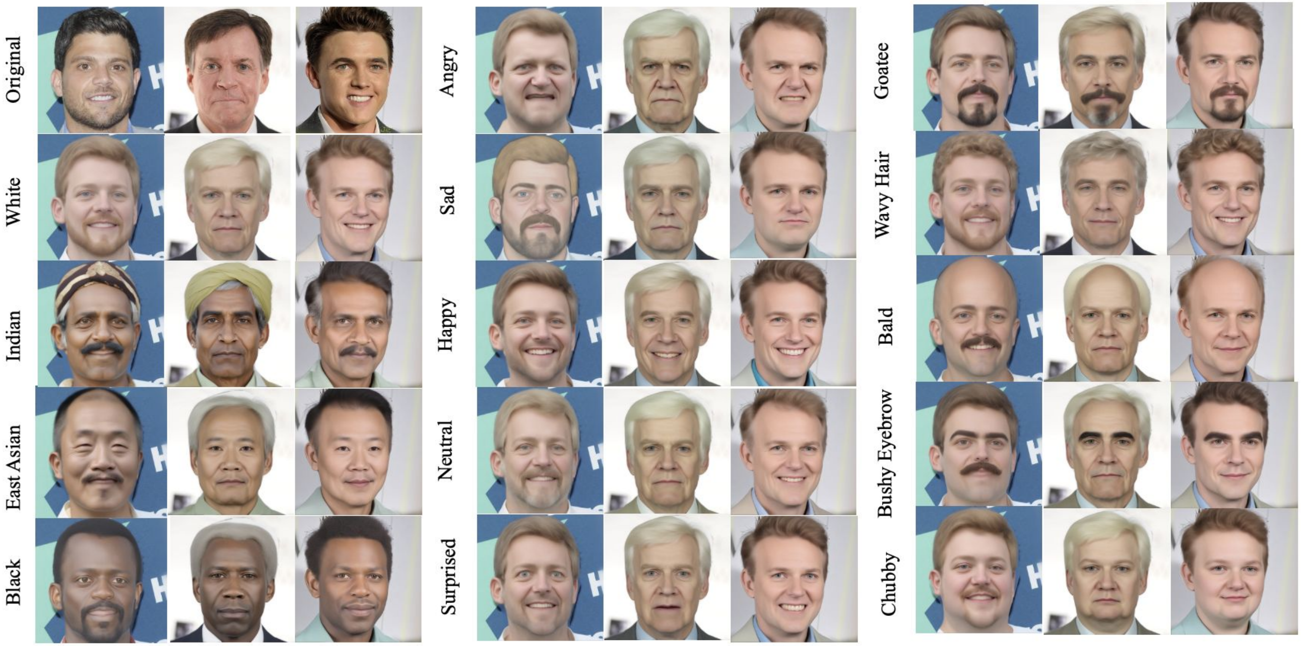


Prompt Complexity

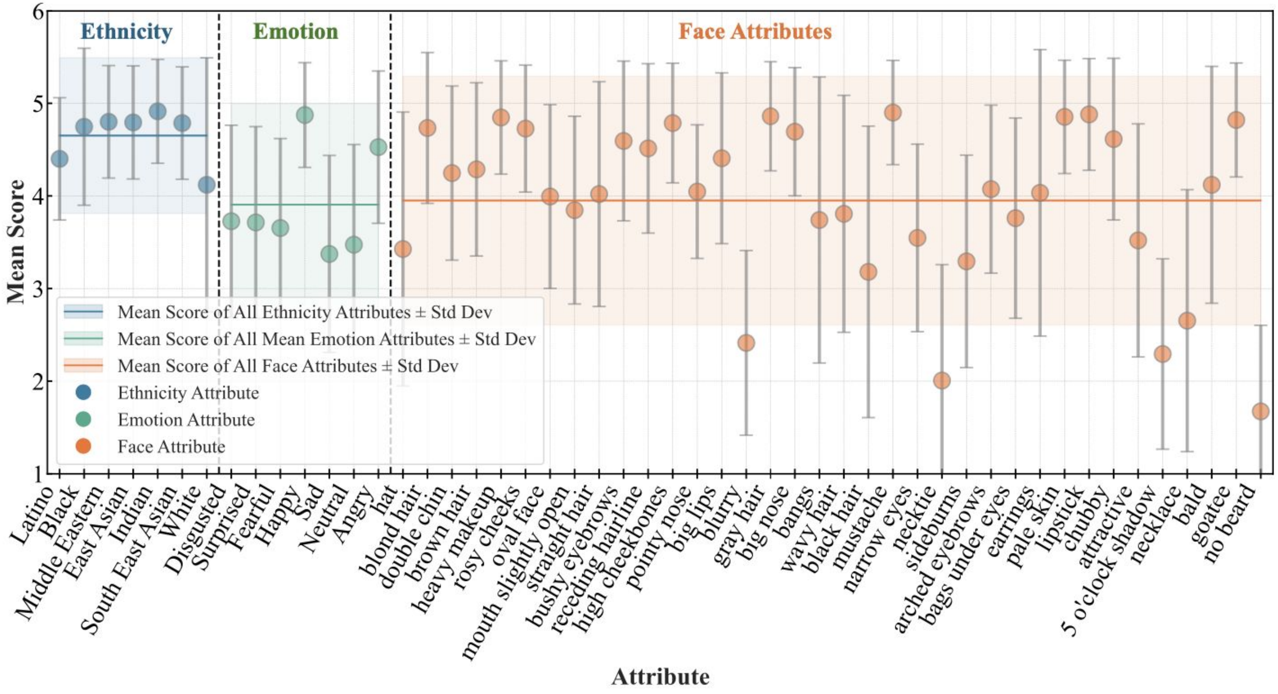




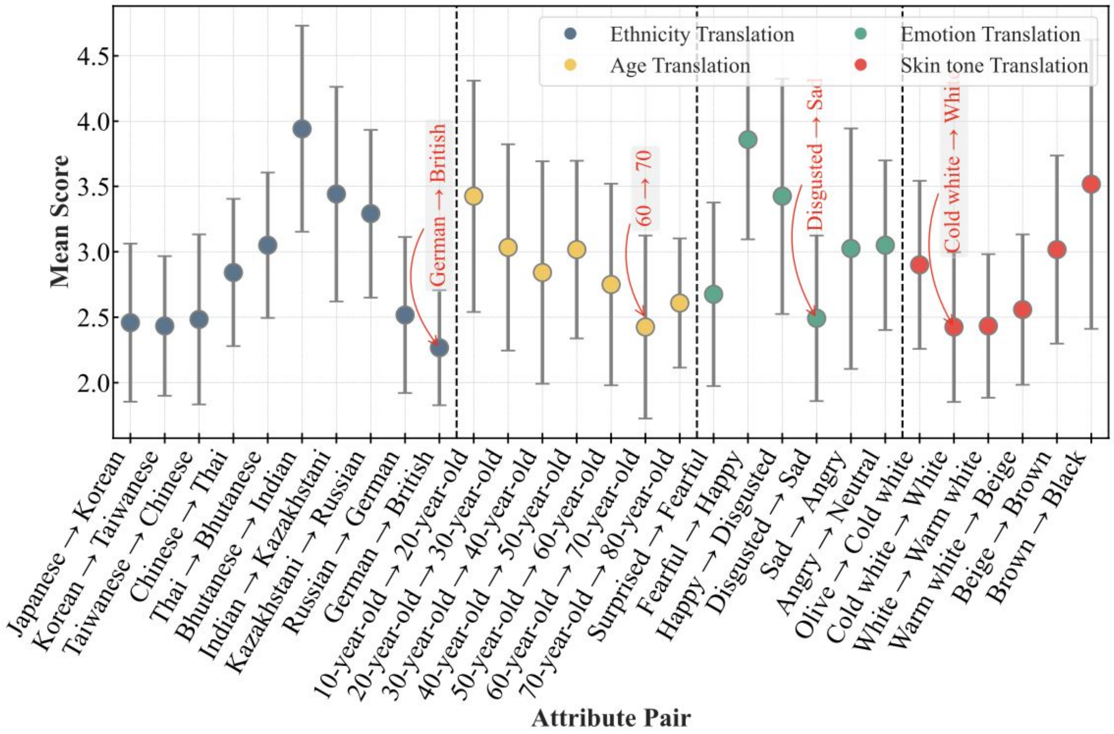
Attribute Fidelity



Attribute Fidelity



Fine-Grain Attribute Translation



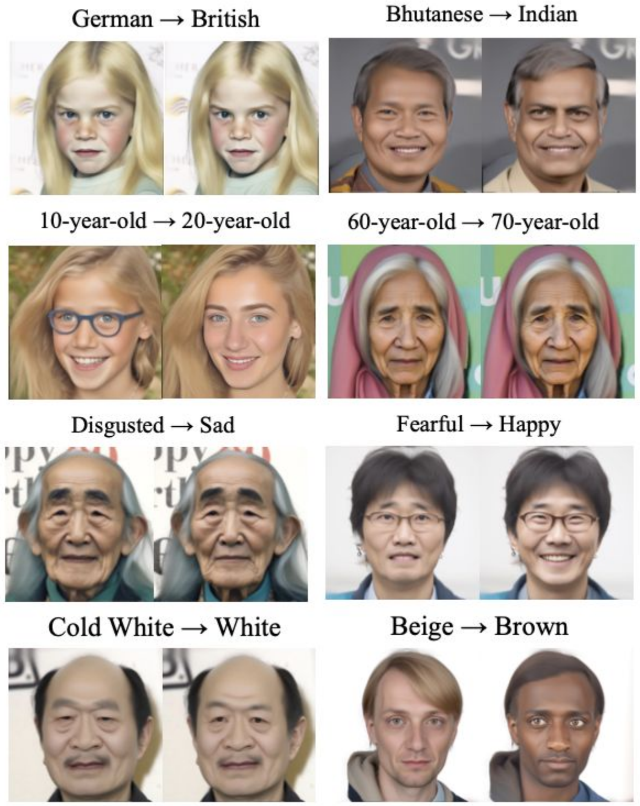


Image Utility: Downstream Training

RefSD consistently improves ML training performance, either used with or as pre-training.

Model	Emotion				Age			
	S	R	S→R	S+R	S	R	S→R	S+R
ViT-Tiny	39.6	41.5	**42.2**	42.0	48.4	57.0	55.7	**58.5**
ViT-Base	36.3	41.5	**45.3**	44.3	48.2	58.4	58.1	**59.9**

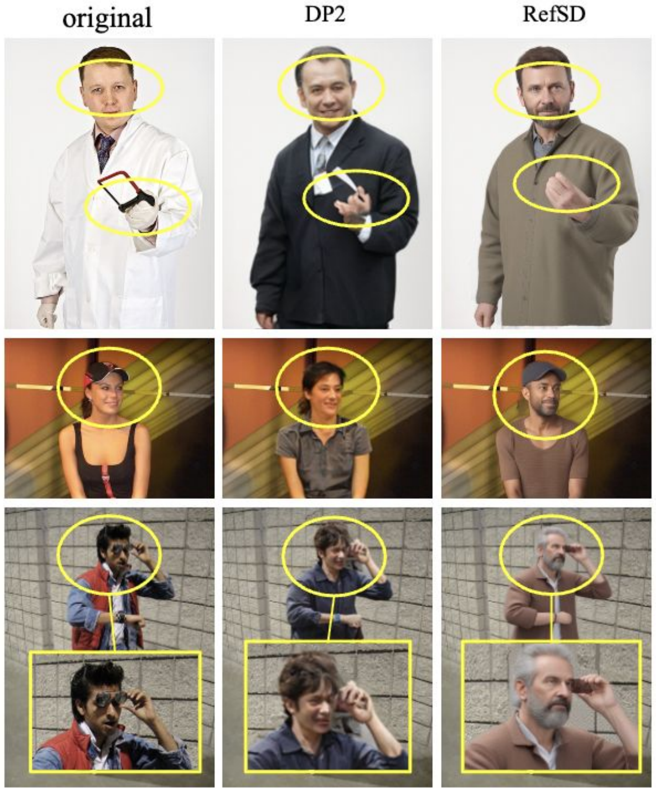
Model	Gender				Ethnicity			
	S	R	S→R	S+R	S	R	S→R	S+R
ViT-Tiny	52.9	60.6	**65.1**	63.4	68.2	77.5	**77.6**	77.5
ViT-Base	53.1	61.9	64.4	**73.0**	67.6	78.2	78.8	**79.9**

Table 4: Classifier training using real (R) and RefSD's synthetic (S) data on RAF-DB dataset

Metric	S	R	S → R
mAP@[.5 : .95] ↑	26.4	25.3	**30.8**
mAP@0.5 ↑	33.2	32.2	**38.8**

Table 4: Detector training using real (R) and RefSD's synthetic (S) data on OpenImages dataset

Comparisons with Recent Anonymization Methods



PPML:

Utilizing Image Datasets With and Without Consent

Table of Contents

PerceptAnon: Exploring the Human Perception of Image Anonymization Beyond Pseudonymization for GDPR

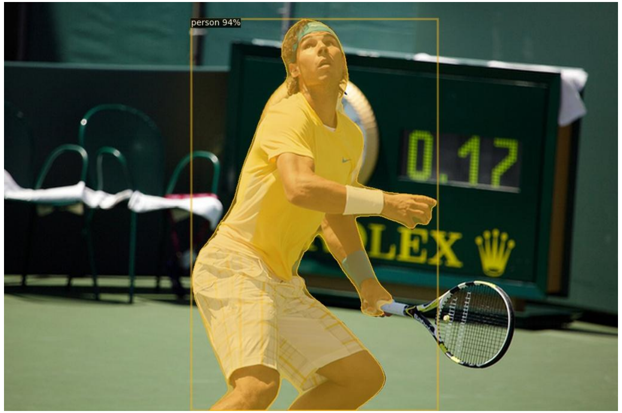
Kartik Patwari*, David Schneider*, Xiaoxiao Sun, Chen-Nee Chuah, Lingjuan Lyu, Vivek Sharma*

ICML 2024

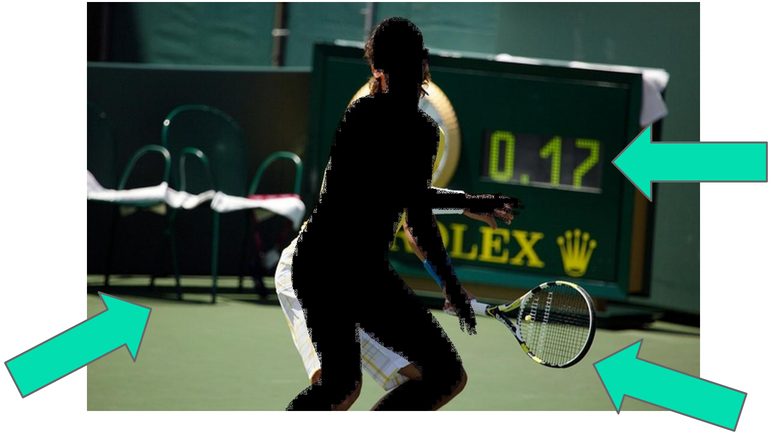
Image Anonymization: Remaining Privacy Cues?

Anonymization is process of removing PIIs

Can background de-anonymize?

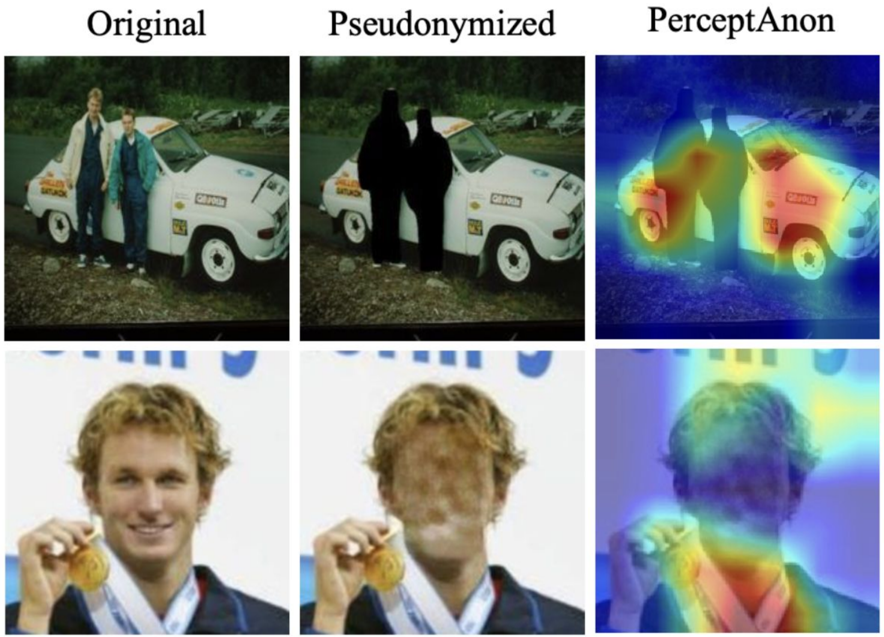


Remove PIIs



PerceptAnon

GDPR: “*the use of additional information can lead to identification of individuals*".



Paper

Thank you!

kpatwari@ucdavis.edu
kartikp7.github.io