Empowering SFDA via MLLM-Guided Reliability-Based Curriculum Learning

Dongjie Chen[1*], Kartik Patwari[1*], Xiaoguang Zhu[1], Zhengfeng Lai[1], Samson Cheung[2], Chen-Nee Chuah[1]

[1] University of California, Davis [2] University of Kentucky

WACV 2026 TUCSON, AZ 3/6 – 3/10

TL;DR

Multi-teacher MLLM distillation + reliability-based curriculum learning for SOTA Source-Free Domain Adaptation!

Background and Motivation

Problem

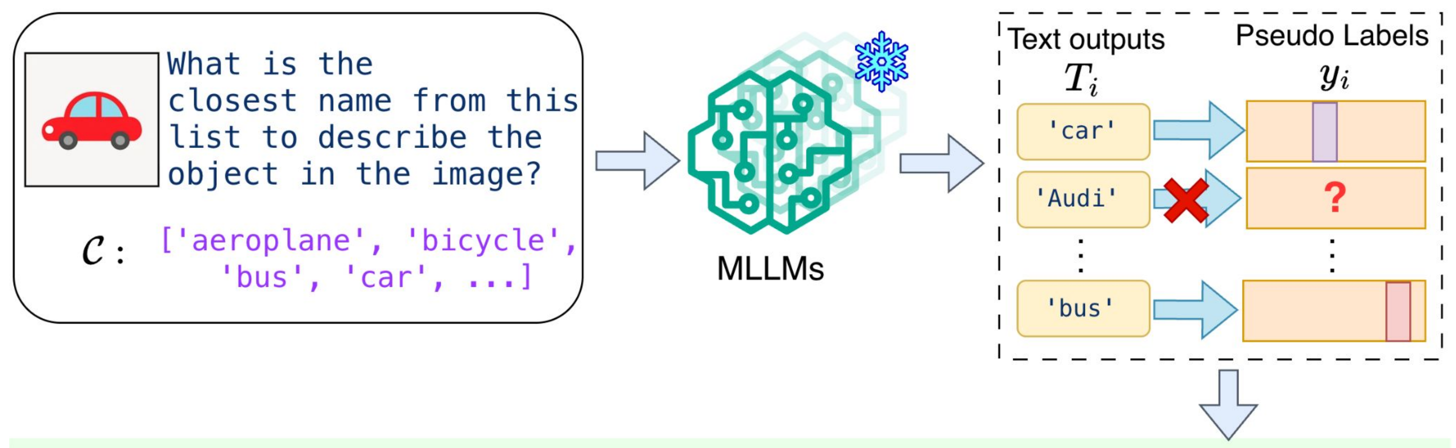
- Source-Free Domain Adaptation (SFDA) adapts models **without access to source data**
- Existing SFDA methods:
  - Single-model pseudo-labels
  - Handcrafted prompts or confidence filtering
- Limited zero-shot performance of multimodal LLMs (MLLMs)
  - Predictions can be inconsistent across samples
  - Outputs are noisy and free-form
  - Direct use is computationally expensive
- Challenge: **use strong teachers while controlling label noise**

Key Idea

- Use **multiple frozen MLLMs** as supervision sources
- Measure **teacher agreement** to estimate pseudo-label **reliability**
- Train progressively using a **reliability-based curriculum**
- Learn a compact student model for efficient deployment

Pseudo-labeling with STS

- MLLM outputs are free-form and not label-aligned
- STS maps outputs and classes into a shared embedding space
- Pseudo-labels are assigned via maximum semantic similarity

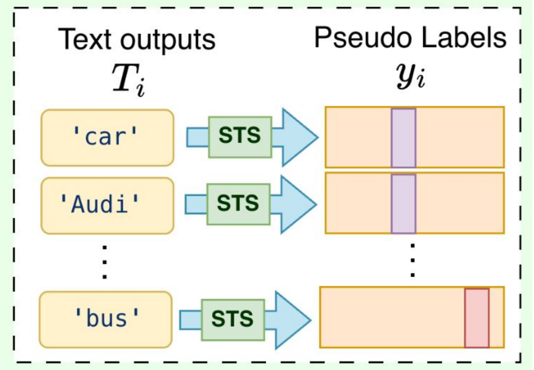


Semantic Text Similarity (STS)

$$\mathbf{v} = f_{\text{STS}}(T)$$

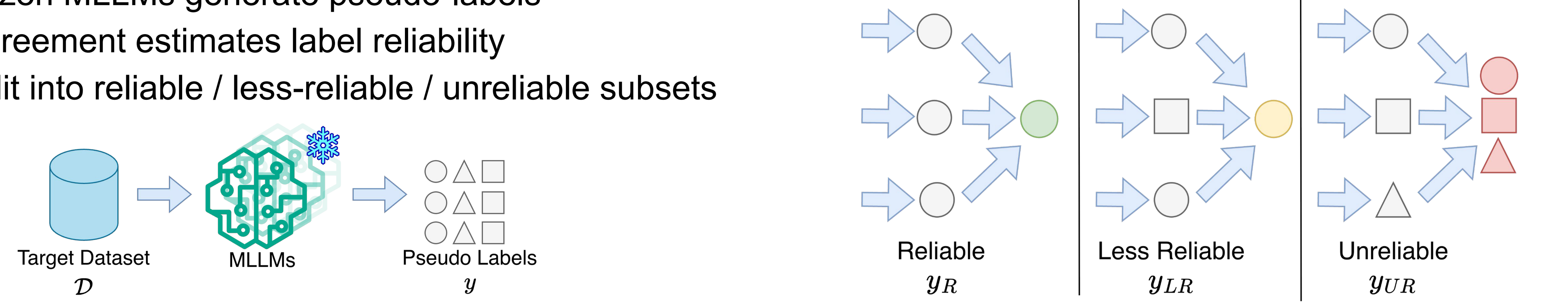
$$\text{STS}(T_1, T_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|_2 \|\mathbf{v}_2\|_2}$$

$$\hat{y}_i = \arg\max_{c \in \mathcal{C}} \text{STS}(T^c, T_i)$$



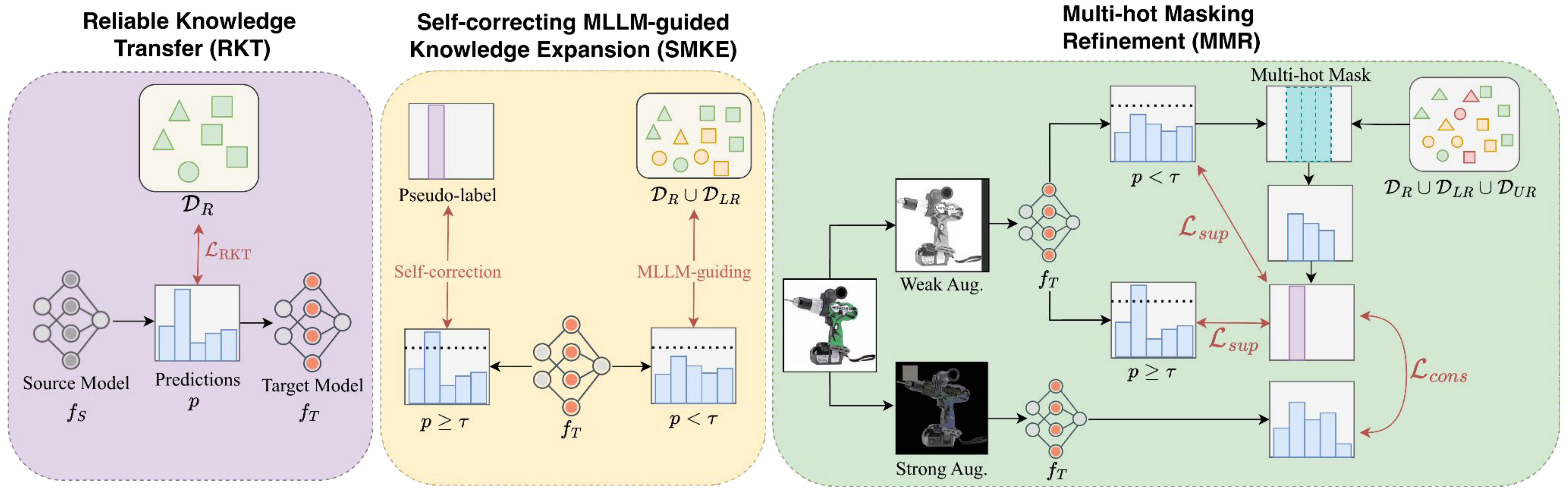
Reliability-Based Curriculum Learning

- Multiple frozen MLLMs generate pseudo-labels
- Teacher agreement estimates label reliability
- Dataset split into reliable / less-reliable / unreliable subsets



Reliability-based Curriculum Learning (RCL)

- Reliable (RKT) → Less reliable (SMKE) → Unreliable (MMR)



Main Results

- Accuracy (%) on SFDA (adaptation from Source → Target domain), and Average across all adaptation tasks.
- **State-of-the-Art on Office-Home, DomainNet, VisDA** across all adaptation tasks

Method	Venue	SF	CP	ViT	Office-Home 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.	DomainNet Avg.	VisDA 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	52.7	45.3
DAPL-RN	NeurIPS'23	✗	✓	✗	54.1	84.3	84.8	74.4	83.7	85.0	74.5	54.6	84.8	75.2	54.7	83.8	74.5	74.8	86.9
PADCLIP-RN	ICCV'23	✗	✓	✗	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	—	88.5
ADCLIP-RN	ICCV'23	✗	✓	✗	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	75.2	88.5
PLUE	CVPR'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	64.7	88.3
C-SFDA	CVPR'23	✓	✗	✗	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	—	87.8
PSAT-GDA	TMM'23	✓	✗	✓	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	—	86.3
TPDS	IJCV'24	✓	✗	✗	59.3	80.3	82.1	70.6	79.4	80.9	69.8	56.8	82.1	74.5	61.2	85.3	73.5	67.1	87.6
DIFO-C-RN	CVPR'24	✓	✓	✗	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	76.7	88.8
DIFO-C-B32	CVPR'24	✓	✓	✓	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	80.0	90.3
LCFD-C-RN	—	✓	✓	✗	60.1	85.6	86.2	77.2	86.0	86.3	76.6	61.0	86.5	77.5	61.4	86.2	77.6	78.0	89.3
LCFD-C-B32	—	✓	✓	✓	72.3	89.8	89.9	81.1	90.3	89.5	80.1	71.5	89.8	81.8	72.7	90.4	83.3	80.0	89.3
LLaVA-34B*	NeurIPS'23	-	✓	✓	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	86.1	92.1
InstBLIP-XXL*	NeurIPS'23	-	✓	✓	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	85.3	86.7
ShrGPT4V-13B*	ECCV'24	-	✓	✓	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	81.7	90.4
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	89.4	93.2
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**	**89.7**	**93.3**

SF: Source-Free ; CP: uses CLIP; ViT: ViT backbone; *Zero-shot with STS

Ablation Studies

Component Ablation

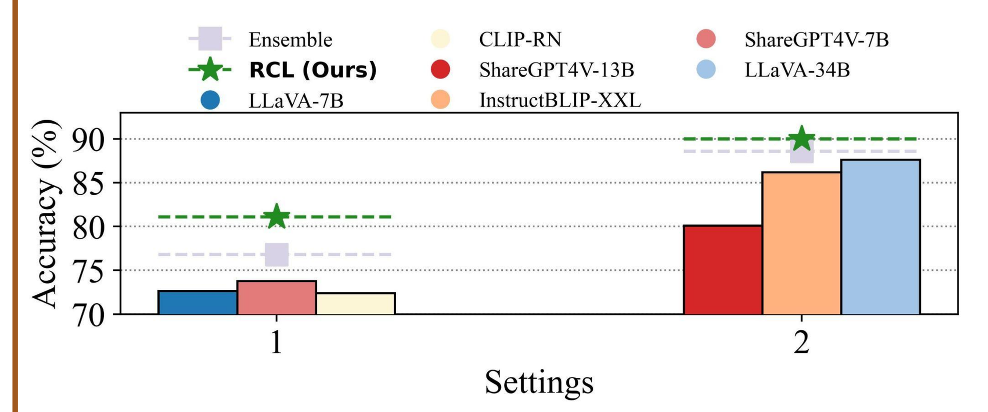
RCL RKT	SMKE	MMR	Office-Home →A	→C	→P	→R	Avg.
✓	✗	✗	82.8	73.3	89.3	88.1	83.3
✓	✗	✓	87.7	80.2	93.3	92.0	88.3
✓	✓	✗	88.5	80.9	95.1	92.5	89.3
✓	✓	✓	**89.3**	**82.3**	**95.3**	**92.9**	**90.0**

RCL Without MLLMs

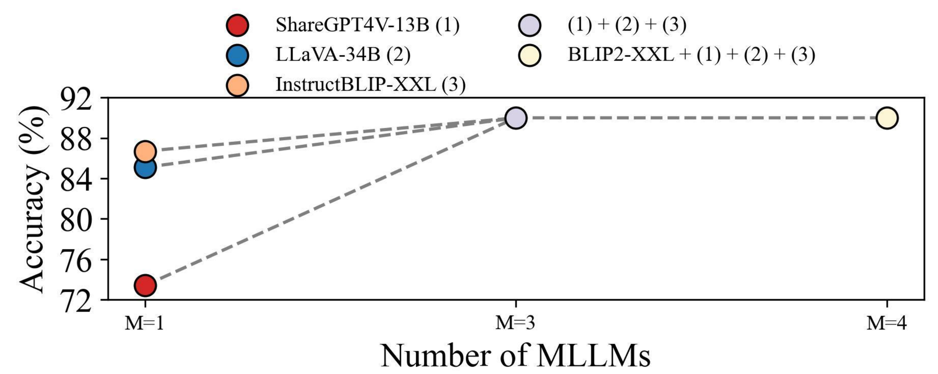
Method	→C	→P	→R	→A	Avg.
TPDS (A)	59.1	81.7	81.7	71.6	73.5
LCFD-C-B32 (B)	72.2	90.2	89.7	81.0	83.3
DIFO-C-B32 (C)	70.4	**90.8**	88.8	**82.3**	83.1
RCL (A, B, C)	**71.9**	90.7	**89.2**	81.7	**83.4**

- Each RCL stage/component improves performance incrementally
- Even without MLLMs, RCL gets competitive performance

Different Teacher MLLMs



Number of Teacher MLLMs



- RCL remains robust across different teacher choices
- Performance improves up to three teachers, and saturates beyond

Backbone Ablation

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**

Prompt Sensitivity

Prompt Template	A→C	A→P	A→R	Avg.
Naive prompt	76.80	92.10	87.45	85.71
Default prompt (D)	**78.35**	**93.78**	**89.58**	**87.19**
D + Domain info	77.54	93.08	88.73	86.25
D + Paraphrased classes	76.20	92.45	87.90	85.20
D + Distractor labels	76.90	92.70	88.25	85.79

- Performance gains are architecture-agnostic
- STS reduces sensitivity to prompt variations for teacher pseudo-label generation

Conclusion / Summary

- Introduce MLLMs as foundation teachers for SFDA task
- Align free-form MLLM outputs using STS
- Propose Multi-teacher consensus to estimate label reliability
- Design RCL, a three-stage curriculum learning framework:
  - Reliable → Less reliable → Unreliable samples
- Achieves SOTA SFDA without foundation model fine-tuning

RCL improves robustness by prioritizing reliable supervision before learning from noisy samples.

Project Page

Code