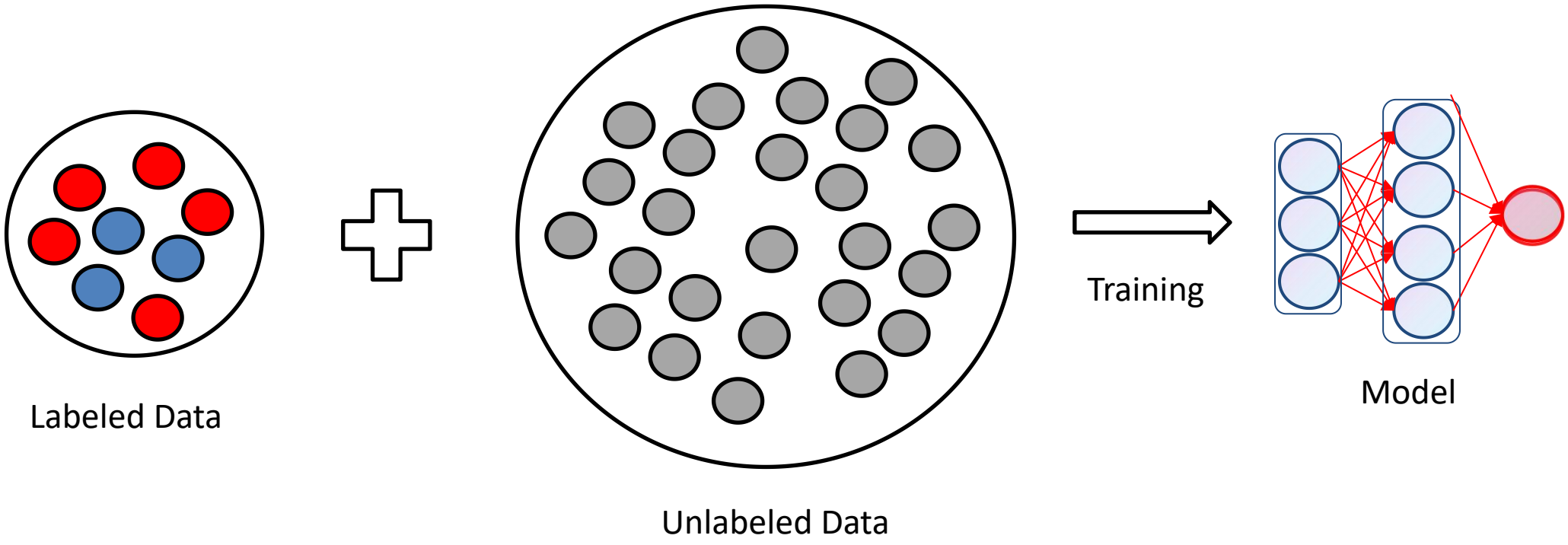
SEMIREWARD: A GENERAL REWARD MODEL FOR SEMI-SUPERVISED LEARNING

Siyuan Li[1,2*] **Weiyang Jin**[2*] **Zedong Wang**[2] **Fang Wu**[2] **Zicheng Liu**[1,2] **Cheng Tan**[1,2] **Stan Z. Li**[2†]
AI Lab, Research Center for Industries of the Future, Hangzhou, China;
[1]Zhejiang University, College of Computer Science and Technology; [2]Westlake University
{lisiyuan; jinweiyang; wangzedong; wufang; liuzicheng; tancheng; stan.zq.li}@westlake.edu.cn

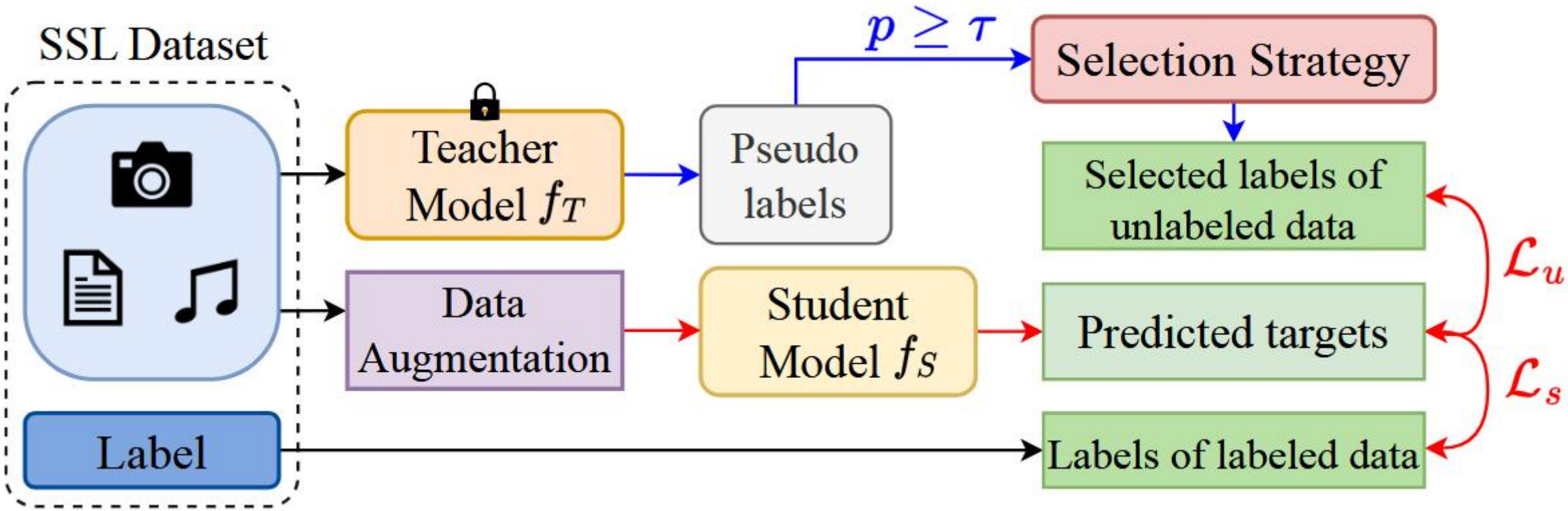
ICLR 2024

Semi-supervised Learning

Leverages both labeled and unlabeled data to improve model performance, especially when labeled data is scarce or expensive to obtain.



Consistency Regularization Framework



Improving Quality of Pseudo-labeling

FixMatch (2020) relies on a fixed threshold but limits usage of more unlabeled data and leads to imbalanced pseudo-labels.

FlexMatch (2021) employs class-specific thresholds to alleviate class imbalance by reducing thresholds for challenging classes.

SoftMatch (2022) explores a trade-off between pseudo-label quantity and quality with a truncated Gaussian function to weigh sample confidence.

ShrinkMatch (2023) applies contrastive learning through adaptive contraction of the class space.

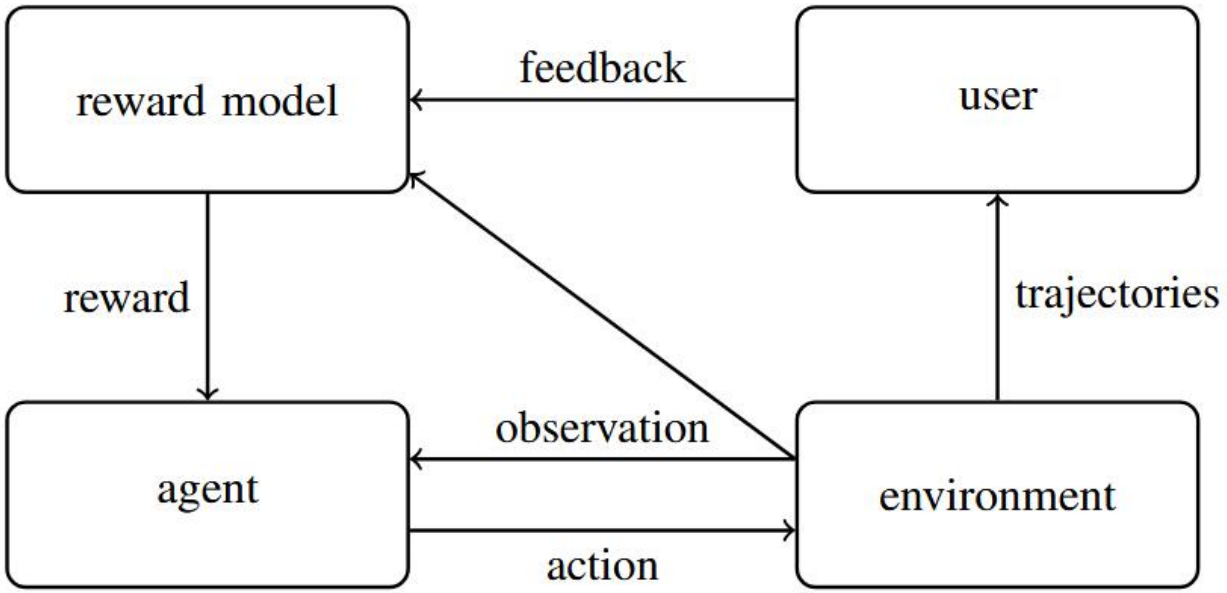
Improving Tolerance of Inaccurate Labels

Π model (2015) introduces dual perturbations to input samples.

Temporal Ensembling (2017) maintains an EMA of label predictions for each training example.

Mean Teacher (2017) averages model weights to reduce label dependency during training.

Reward Modeling in Reinforcement Learning



Most reward models are supervised by classification losses, e.g., ranking loss, on constructed preference datasets from users.

Measurement of Label Quality

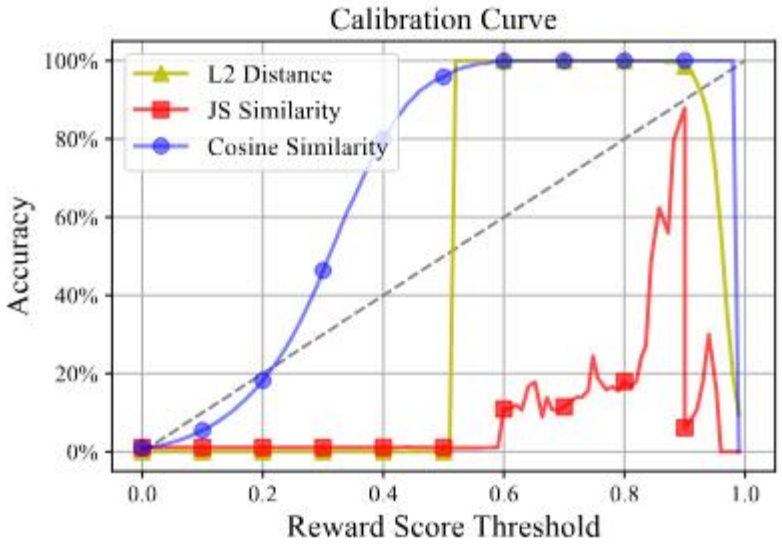
Define a continuous metric of pseudo-label quality based on label similarity.

Reward score:

$$r(y^u, y^l) = \mathcal{S}(y^u, y^l) \simeq \mathcal{R}(x, y^u) \in [0, 1].$$

Label similarity:

$$\mathcal{S}(y_i, y_j) = \frac{y_i \cdot y_j}{2 \left\| y_i \right\| \left\| y_j \right\|} + 0.5 \in [0, 1].$$



The ideal reward score should satisfy monotonicity and smoothness (not increasing dramatically) and strive to meet the trend of calibration curve, where a lower reward confidence indicates poorer label quality.

Network Structures

Rewarder：

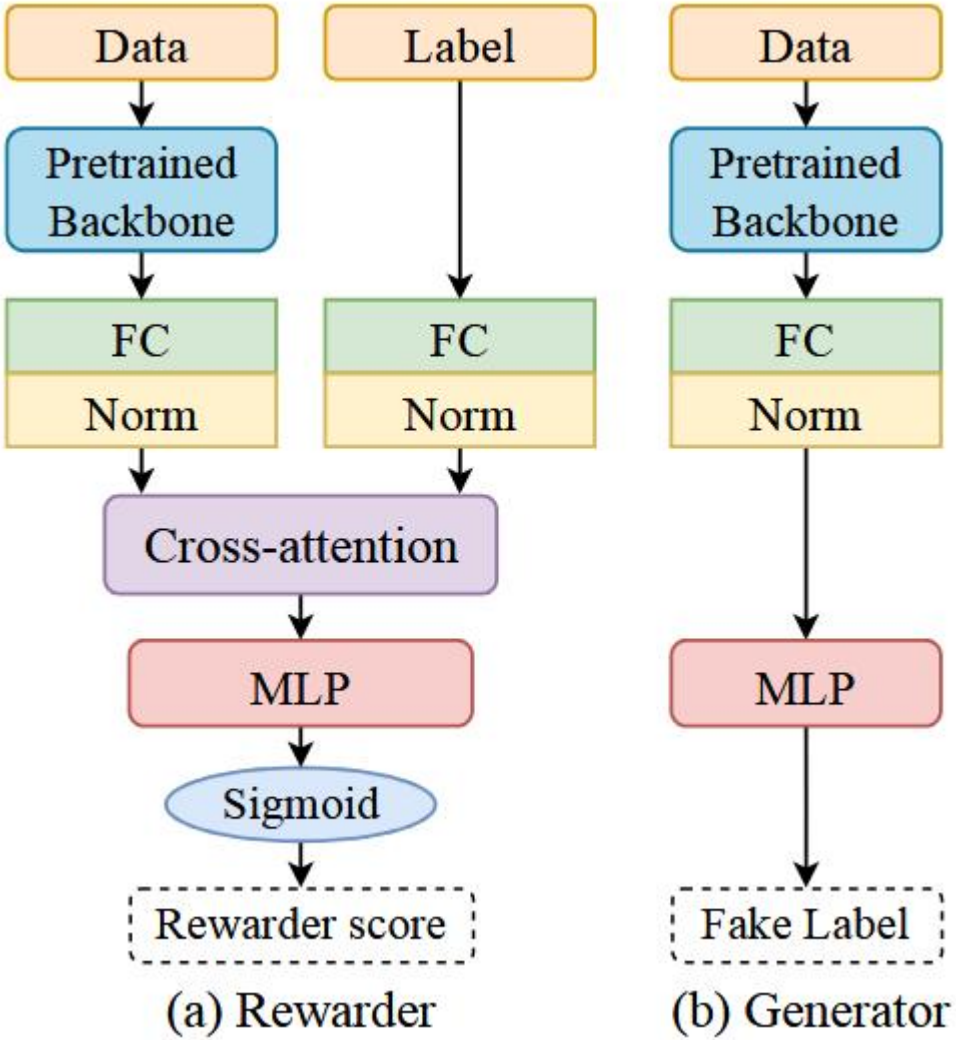
$$\mathcal{R}(x^u, y^u) = \text{Sigmoid}\Big(\text{MLP}\Big(\text{CA}\big(\text{Emb}(f(x^u)), \text{Emb}(y^u)\big)\Big)\Big),$$

extracts semantic information of y^l from $x^u$and tell the similarity between x^u and y^uaccording to their semantic correlation.

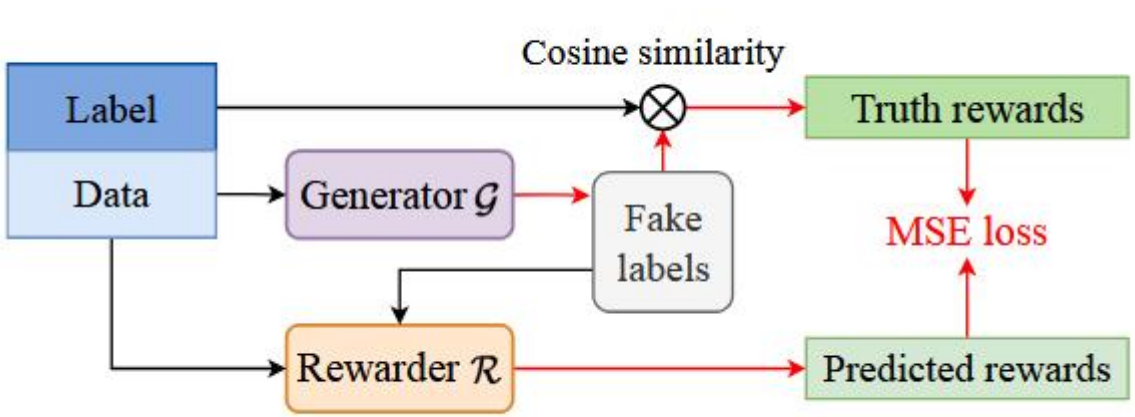
Generator：

$$\mathcal{G}(x^u) = y^f \in \mathbb{R}^C$$

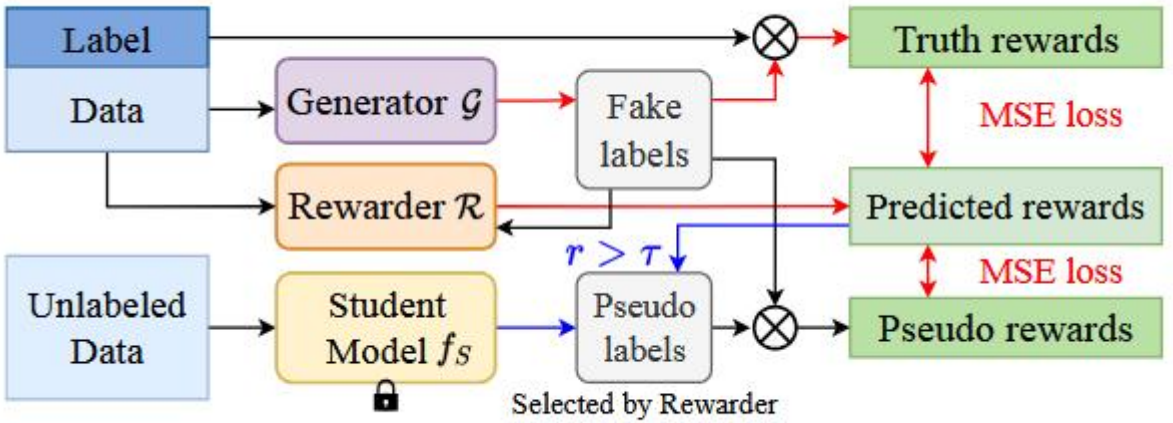
decouples the training of f_S and $\mathcal{R}$ to avoid confirmation bias.



Two-stage Training Paradigm



(a) Stage 1: Pre-training with labeled data



(b) Stage 2: Semi-supervised training with $\mathcal{D}_R$

Pre-training Rewarder

In the first stage, our main optimization goal is to approximate the ground truth reward scores with a wide range of fake labels without affecting the training of f_S .

$$\mathcal{L}_{\mathcal{R}} = \frac{1}{B_R}\sum_{i=1}^{B_R}\ell_2\Big(\mathcal{R}\big(x_i^r, \overline{\mathcal{G}}(x_i^r)\big), \mathcal{S}\big(y_i^r, \overline{\mathcal{G}}(x_i^r)\big)\Big),$$

$$\mathcal{L}_{\mathcal{G}} = \frac{1}{B_R}\sum_{i=1}^{B_R}\ell_2\Big(\overline{\mathcal{R}}\big(x_i^r, \mathcal{G}(x_i^r)\big), 1\Big),$$

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\mathcal{R}} + \mathcal{L}_{\mathcal{G}}$$

Semi-supervised Training Rewarder

In the second stage, the core objective is to optimize f_S using $\mathcal{R}$ to filter high-quality labels.

$$\mathcal{L} = \underbrace{\frac{1}{B_L}\sum_{i=1}^{B_L}\mathcal{H}\Big(y_i^l, f_S\big(\omega(x_i)\big)\Big)}_{\mathcal{L}_L} + \underbrace{\frac{1}{B_U}\sum_{j=1}^{B_U}\mathbb{I}(\mathcal{R}(x_j^u, y_j^u) > \tau)\mathcal{H}\Big(\hat{y}_j^u, f_S\big(\omega(x_j^u)\big)\Big)}_{\mathcal{L}_U} + \mathcal{L}_{\text{aux}},$$

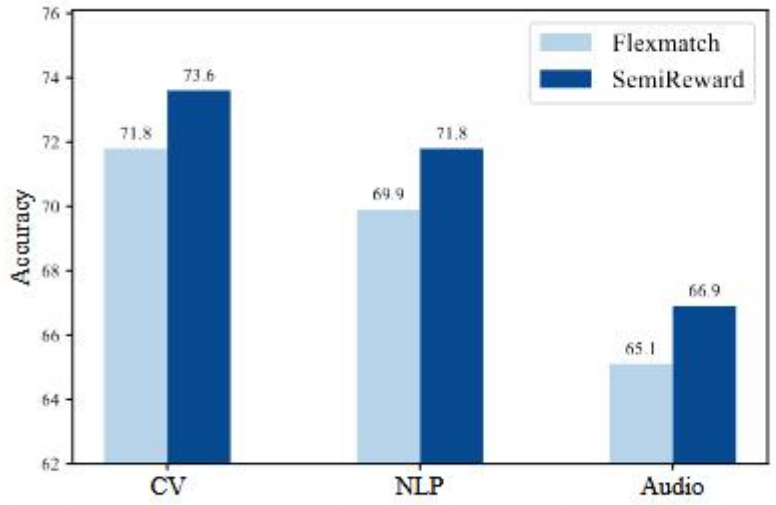
Comparison Results on Classification

Domain	Dataset (Setting)	Pseudo Label		FlexMatch		SoftMatch/FreeMatch		Average	
		Base	+SR	Base	+SR	Base	+SR	Gain	Speed.
Audio	ESC-50 (250)	38.42±0.85	**33.33±0.97**	36.83±0.51	**32.58±0.51**	32.71±0.82	**29.71±0.64**	+4.11	×1.73
	ESC-50 (500)	28.92±0.24	**27.65±0.32**	27.75±0.41	**25.92±0.31**	29.07±1.27	**25.98±0.49**	+2.06	×2.07
	FSDnoisy18k (1773)	34.60±0.55	**33.24±0.82**	26.29±0.17	**25.63±0.28**	29.39±1.83	**26.10±0.83**	+1.77	×1.30
	UrbanSound8k (100)	37.74±0.96	**36.47±0.65**	37.88±0.46	**36.06±0.93**	37.68±1.82	**34.97±1.02**	+1.93	×1.70
	UrbanSound8k (400)	27.45±0.96	**25.27±0.65**	23.78±0.46	**23.45±0.93**	23.78±0.13	**19.39±0.33**	+2.30	×1.08
NLP	AG News (40)	15.19±3.07	**13.90±0.21**	13.08±3.94	**12.60±0.69**	11.69±0.12	**10.67±0.90**	+0.93	×2.77
	AG News (200)	14.69±1.88	**12.10±0.58**	12.08±0.73	**11.05±0.14**	11.75±0.17	**10.02±0.82**	+1.78	×2.30
	Yahoo! Answer (500)	34.87±0.50	**35.08±0.40**	34.73±0.09	**33.64±0.73**	33.02±0.02	**30.92±0.90**	+0.99	×1.80
	Yahoo! Answer (2000)	33.14±0.70	**32.50±0.42**	31.06±0.32	**29.97±0.10**	30.34±0.18	**29.11±0.15**	+0.99	×3.53
	Yelp Review (250)	46.09±0.15	**42.99±0.14**	46.09±0.15	**42.76±0.33**	43.91±0.19	**42.68±0.12**	+2.55	×1.40
	Yelp Review (1000)	44.06±0.14	**42.08±0.15**	40.38±0.33	**37.58±0.19**	40.43±0.12	**38.43±0.14**	+2.26	×1.01
CV	CIFAR-100 (200)	32.78±0.20	**31.94±0.57**	25.72±0.35	**23.74±1.39**	21.07±0.72	**20.06±0.41**	+1.28	×1.04
	CIFAR-100 (400)	25.16±0.67	**23.84±0.20**	17.80±0.57	**17.59±0.35**	15.97±0.24	**15.62±0.71**	+0.63	×1.57
	STL-10 (40)	20.53±0.12	**17.37±0.47**	11.82±0.51	**10.20±1.11**	17.51±0.61	**9.72±0.62**	+4.19	×1.07
	STL-10 (100)	11.25±0.81	**10.88±1.48**	7.13±0.20	**7.59±0.57**	8.10±0.35	**7.10±1.39**	+0.30	×1.11
	Euro-SAT (20)	25.25±0.72	**23.65±0.41**	5.54±0.16	**4.86±1.00**	5.51±0.54	**4.22±0.34**	+1.19	×1.03
	Euro-SAT (40)	12.82±0.81	**8.33±0.33**	4.51±0.24	**3.88±0.69**	5.46±0.34	**3.94±0.71**	+2.21	×1.13

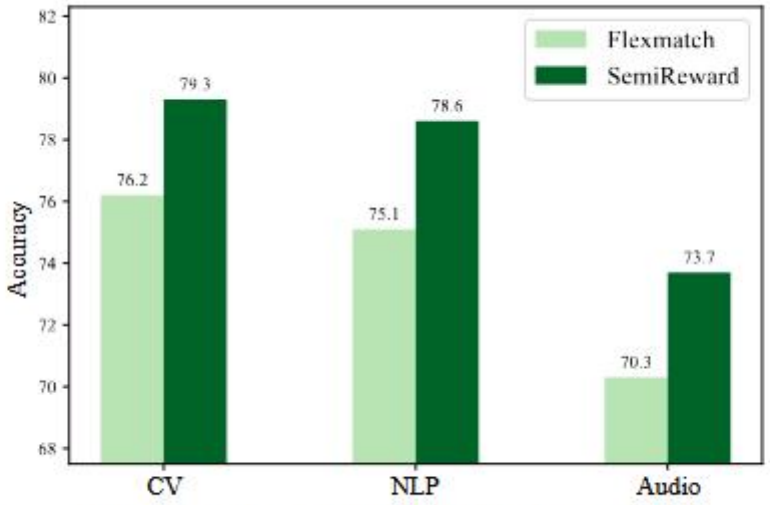
Comparison Results on Regression

Method	RCF-MNIST		IMDB-WIKI		AgeDB	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Supervised	62.02±0.34	22.81±0.07	14.92±0.14	11.52±0.09	14.51±0.13	11.77±0.27
Pseudo Label	62.72±0.11	23.07±0.05	14.90±0.22	11.44±0.53	14.76±0.12	11.71±0.53
Π-Model	63.24±0.63	23.54±0.63	14.80±0.12	11.35±0.12	14.76±0.14	11.92±0.09
MeanTeacher	63.44±0.32	23.25±0.13	15.01±0.64	11.66±0.32	14.99±0.99	12.07±0.48
CRMatch	101.66±0.84	85.45±0.72	22.42±0.23	18.77±0.43	20.42±0.10	17.11±0.49
PseudoLabel+SR	**61.71**±0.34	**22.45**±0.05	**14.80**±0.53	**10.91**±0.12	**14.01**±0.12	**10.77**±0.22
Gain	-0.90	-0.99	-0.10	-0.53	-0.75	-0.94

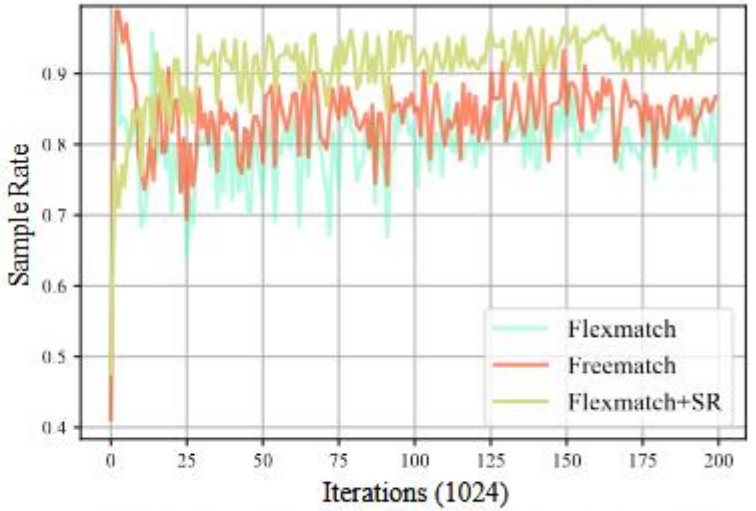
Evaluation of Pseudo-label Quality and Quantity



(a) Pseudo-label quality after stage 1

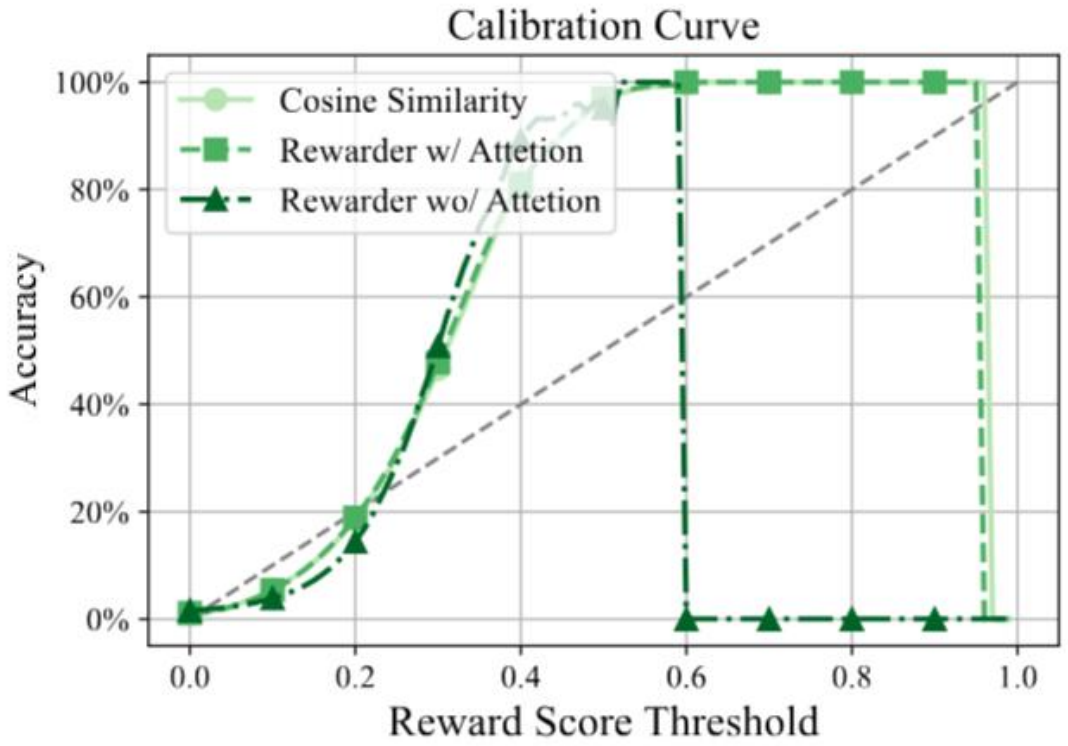


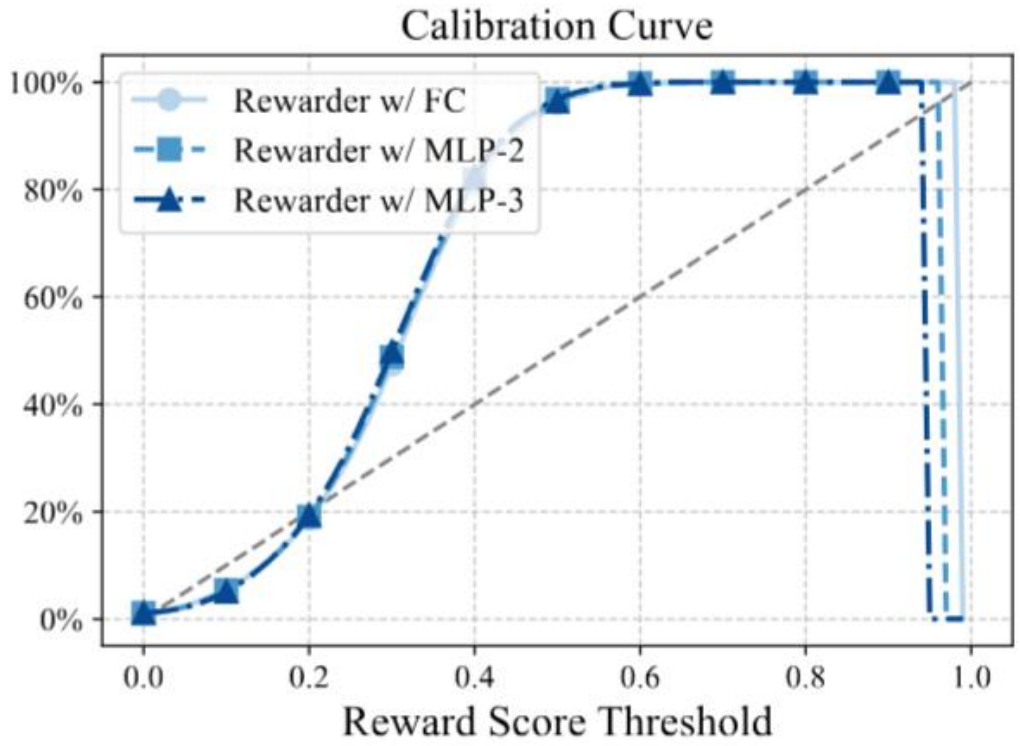
(b) Final pseudo-label quality



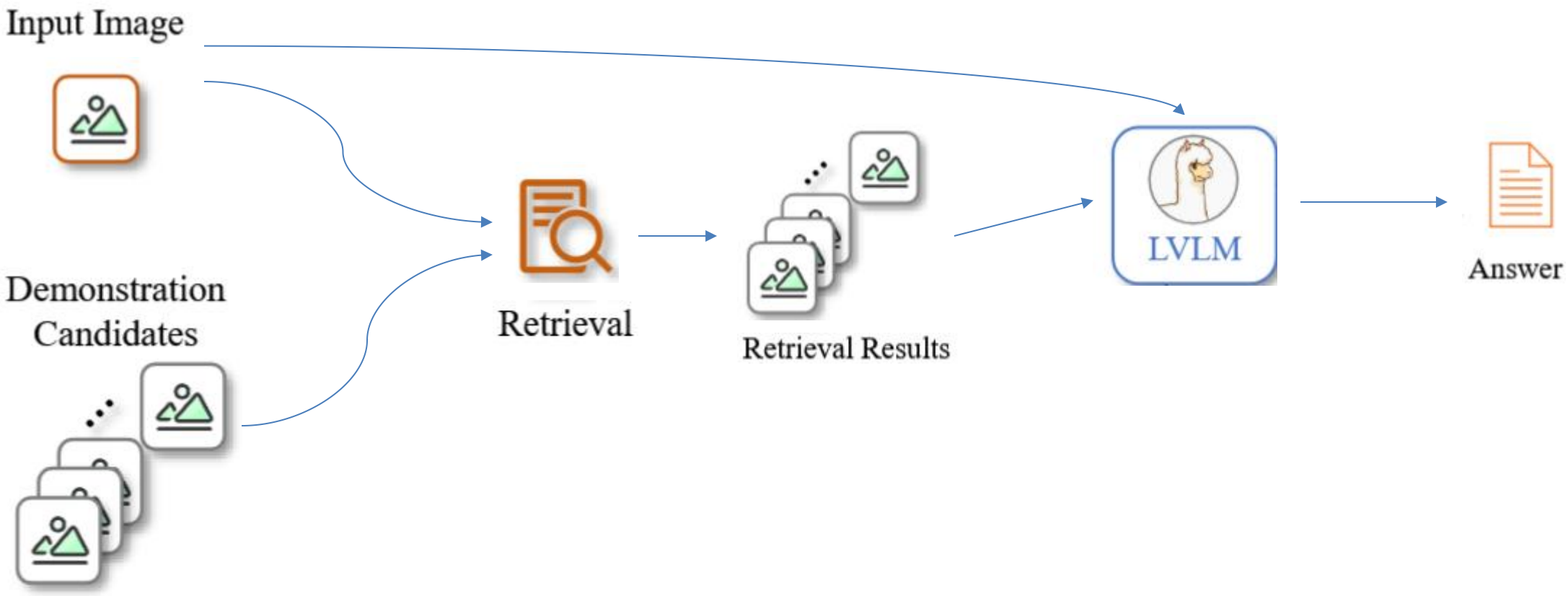
(c) Sampling rate *v.s.* training steps

Ablation Study





Visual In-context Learning



Thanks