Inter-University Research Institute Corporation /
Research Organization of Information and Systems
National Institute of Informatics

Interspeech 2026
Area 4 - Oral 9 - 3

Does Fine-tuning by Reinforcement Learning Improve Generalization in Binary Speech Deepfake Detection?

Xin Wang, Wanying Ge, Junichi Yamagishi
National Institute of Informatics, Japan

wangxin@nii.ac.jp

Inter-University Research Institute Corporation /
Research Organization of Information and Systems
National Institute of Informatics

Interspeech 2026
Area 4 - Oral 9 - 3

Does Fine-tuning ~~by~~ Reinforcement Learning in Binary Speech Deepfake Detection?

similarly to

mitigate catastrophic forgetting

Xin Wang, Wanying Ge, Junichi Yamagishi
National Institute of Informatics, Japan

wangxin@nii.ac.jp

Slides by Xin Wang
National Institute of Informatics

© 2026, Xin Wang. All rights reserved.

This work is licensed under the Creative Commons Attribution 3.0 license.
See http://creativecommons.org/ for details.



Note: Speech data is from ASVspoof 2019 database
https://datashare.ed.ac.uk/handle/10283/3336

One-minute summary

- *Context of our experiments*

Pre-training → *Post (or middle) training* → ***Fine-tuning***

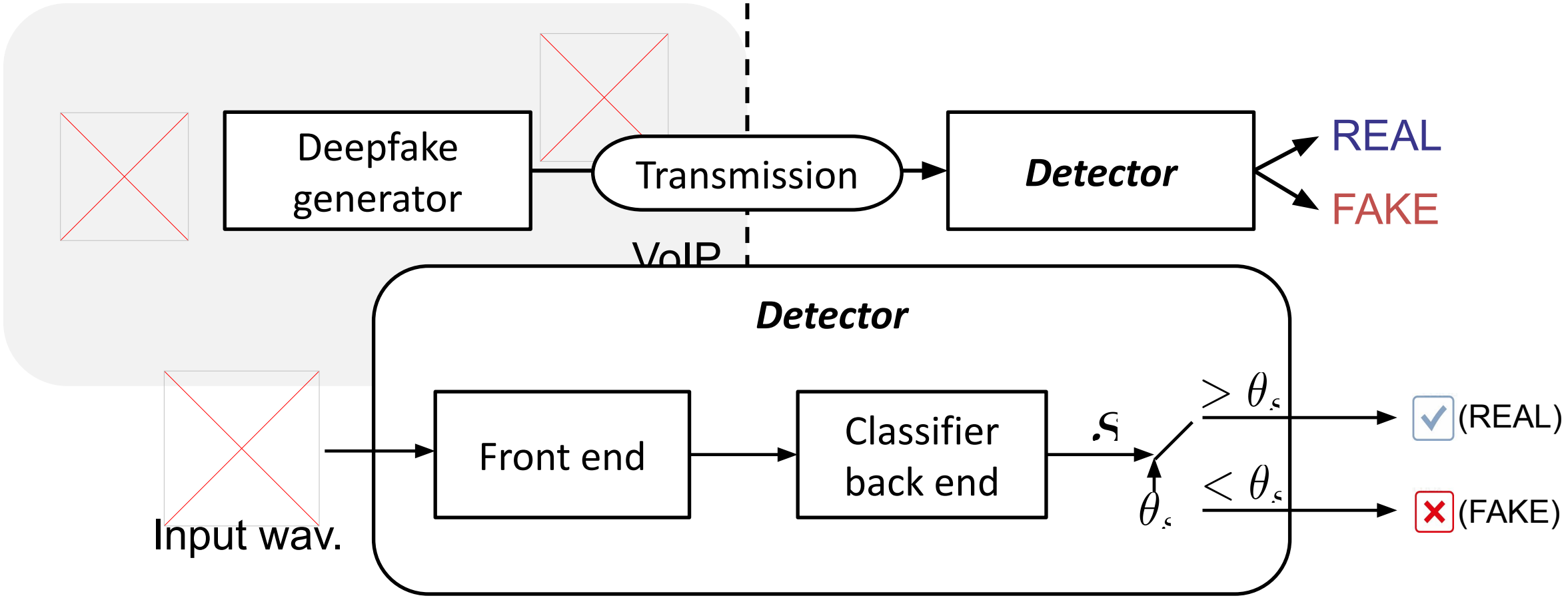
(self-supervised learned front-end) — *(large-scale data, supervised)* — *(small-scale)*

- ***Main findings:***
  - ***Fine-tuning*** *by supervised learning (i.e., cross-entropy) leads to catastrophic forgetting*
  - ***Fine-tuning*** *by Group Relative Policy Optimization (GRPO)-like criterion mitigates catastrophic forgetting*

Introduction

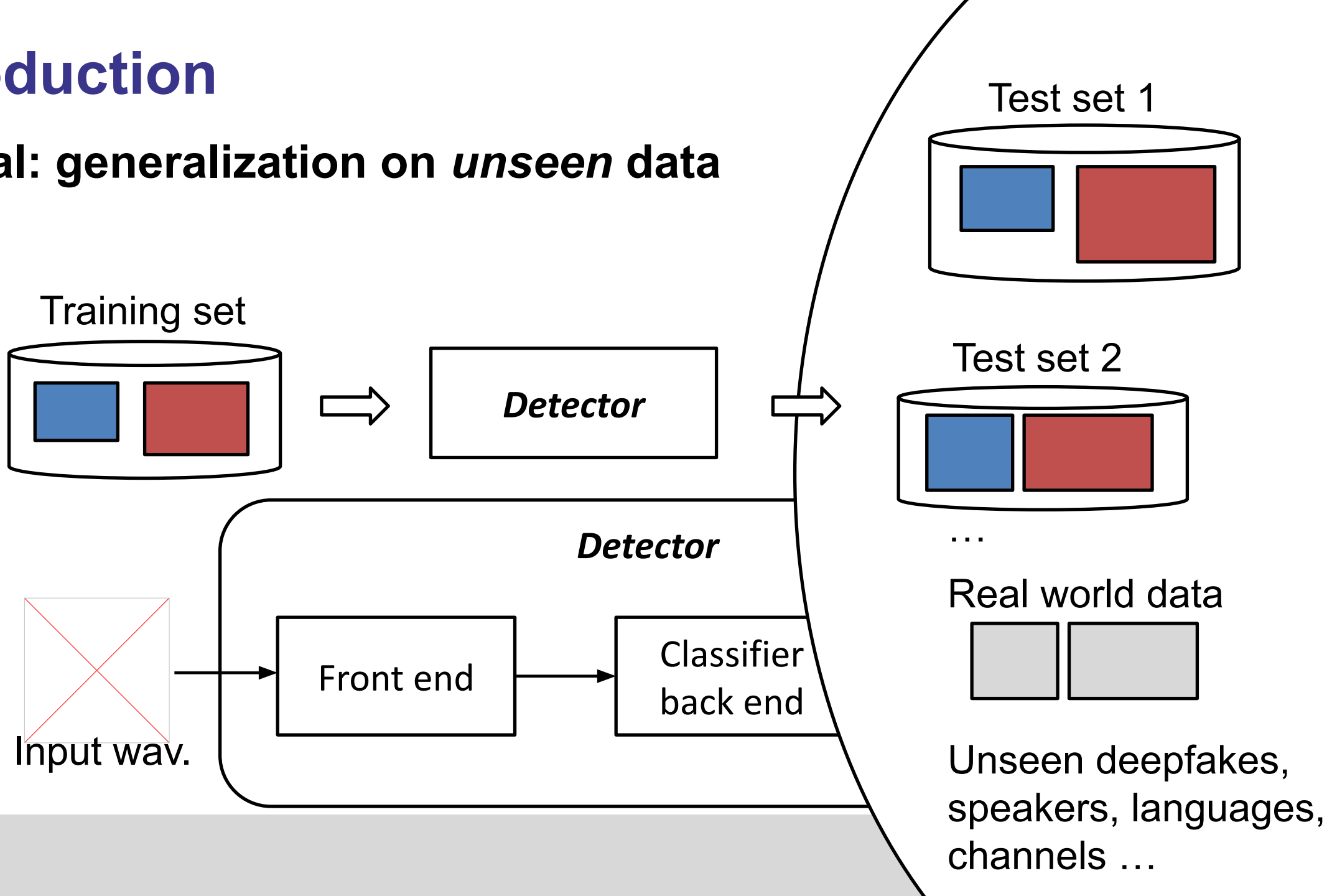
Introduction

- **Task: binary classification**



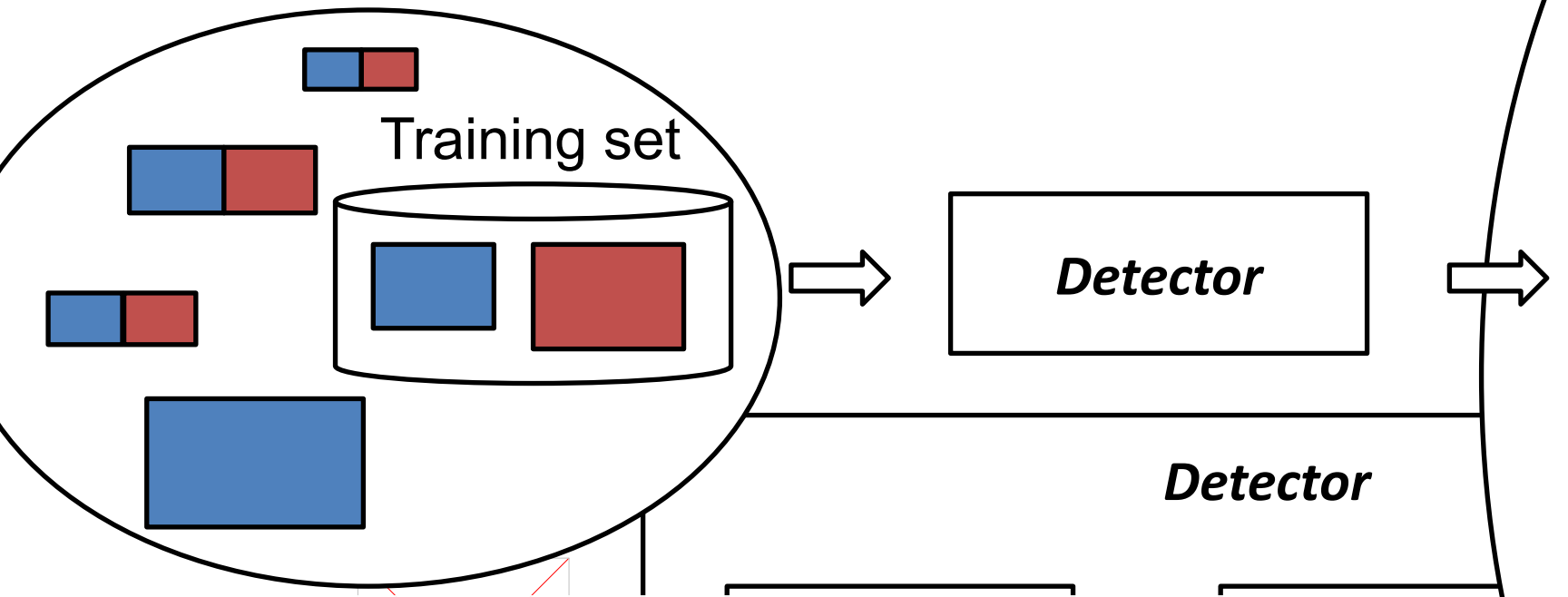
Introduction

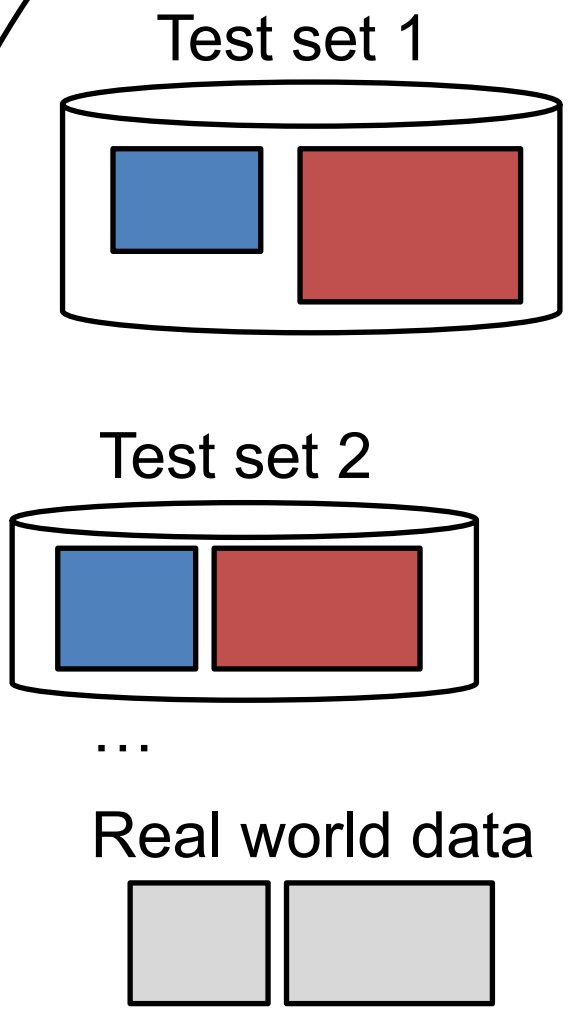
- **Goal: generalization on *unseen* data**



Introduction

- **Promising solution: large-scale training**



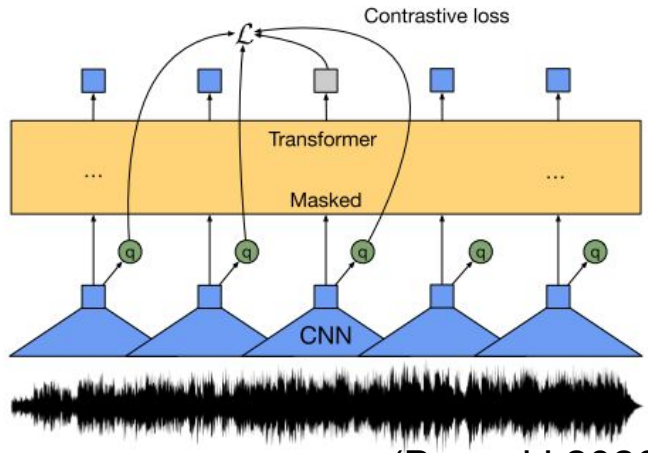


- self-supervised learning (SSL) front-end (Martin-Donas 2022, Wang 2022, Tak 2022)
- (large-scale) supervised training (Ge 2025, Zhu 2026)

Unseen deepfakes, speakers, languages, channels …

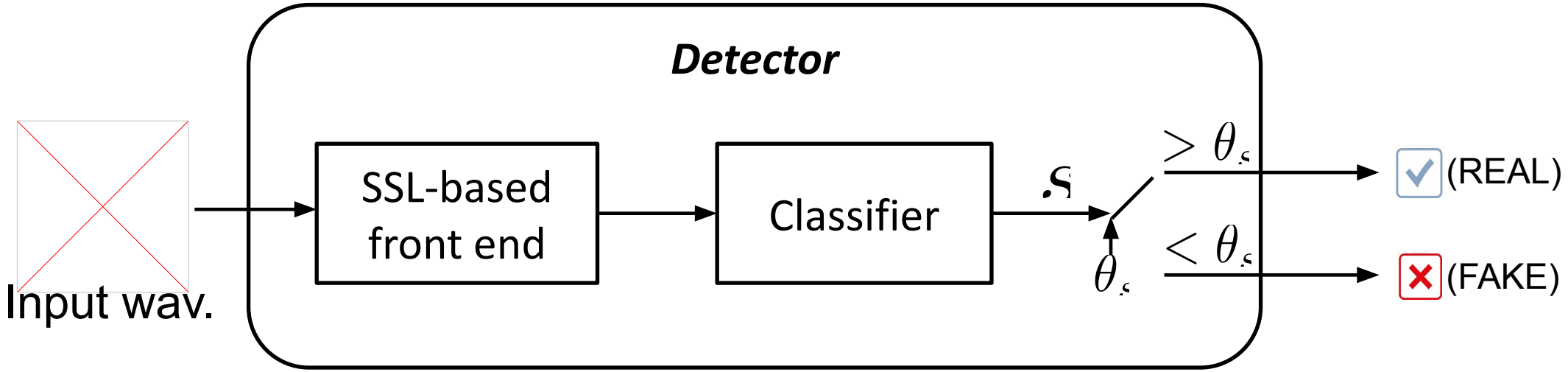
Juan M. Martin-Donas and Aitor Alvarez. 2022. The Vicomtech Audio Deepfake Detection System Based on Wav2vec2 for the 2022 ADD Challenge. In *Proc. ICASSP*, May 23, 2022. 9241–9245.
Xin Wang and Junichi Yamagishi. 2022. Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures. In *Proc. Odyssey*, 2022. 100–106.
Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *Proc*
Wanying Ge, Xin Wang, Xuechen Liu, and Junichi Yamagishi. 2025. Post-training for Deepfake Speech Detection. In *Proc. ASRU*, December 06, 2025.
Yi Zhu, Brahmi Dwivedi, Jayaram Raghuram, and Surya Koppisetti. 2026. Alethia: A Foundational Encoder for Voice Deepfakes. In *Proc. ICML*, 2026.

Introduction – Base method

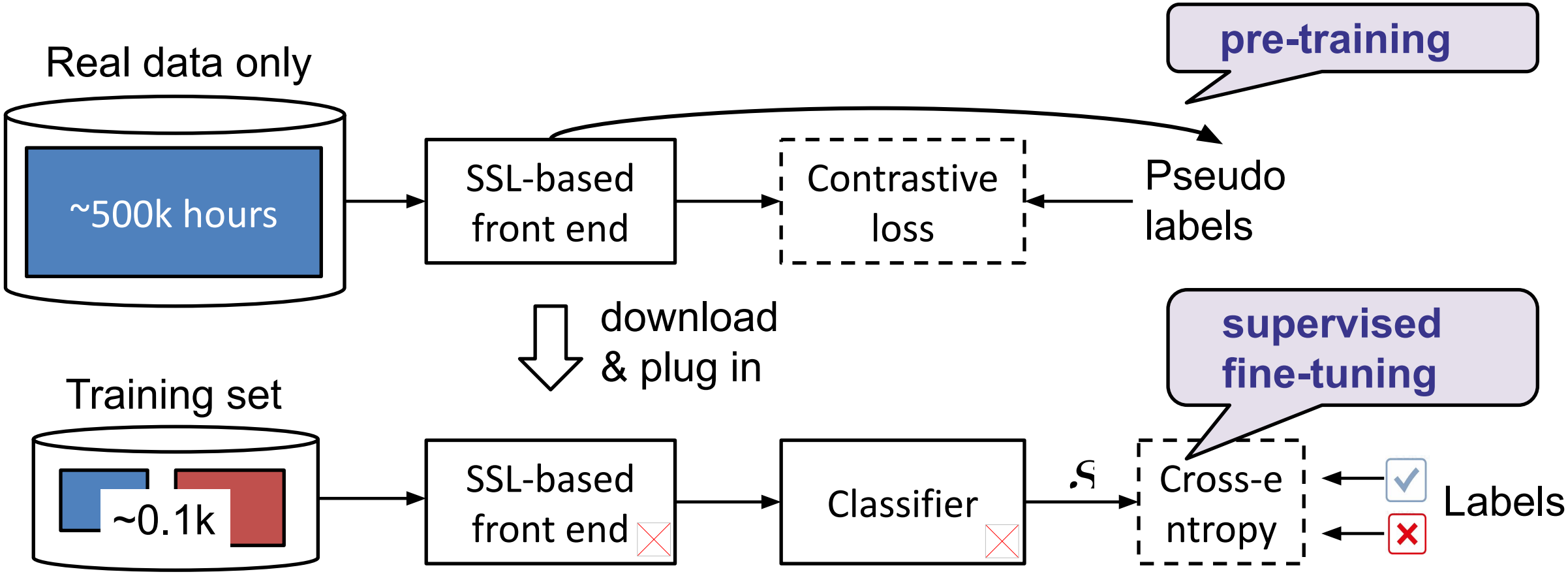


Wav2vec 2.0 (Baevski 2020)

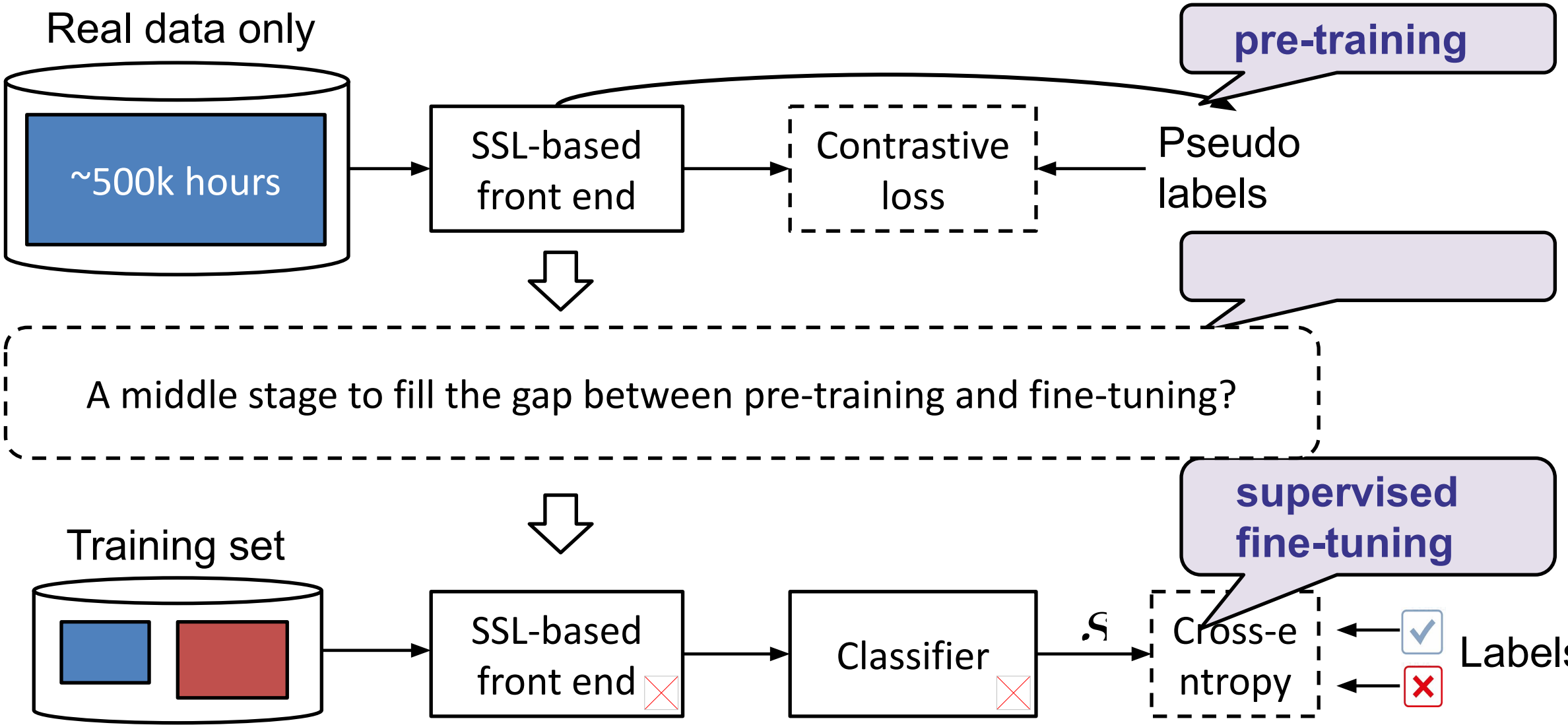
A linear layer,
or a DNN …



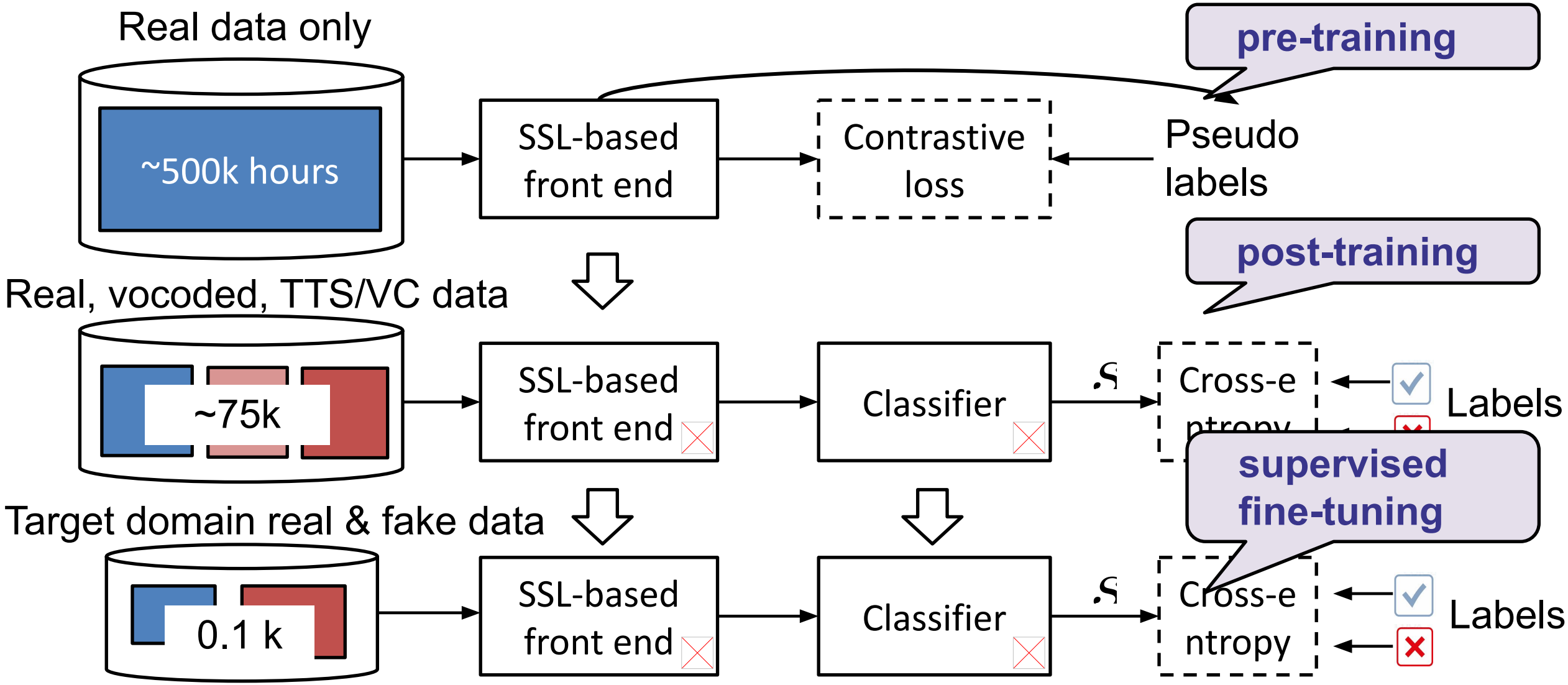
Introduction – Base method



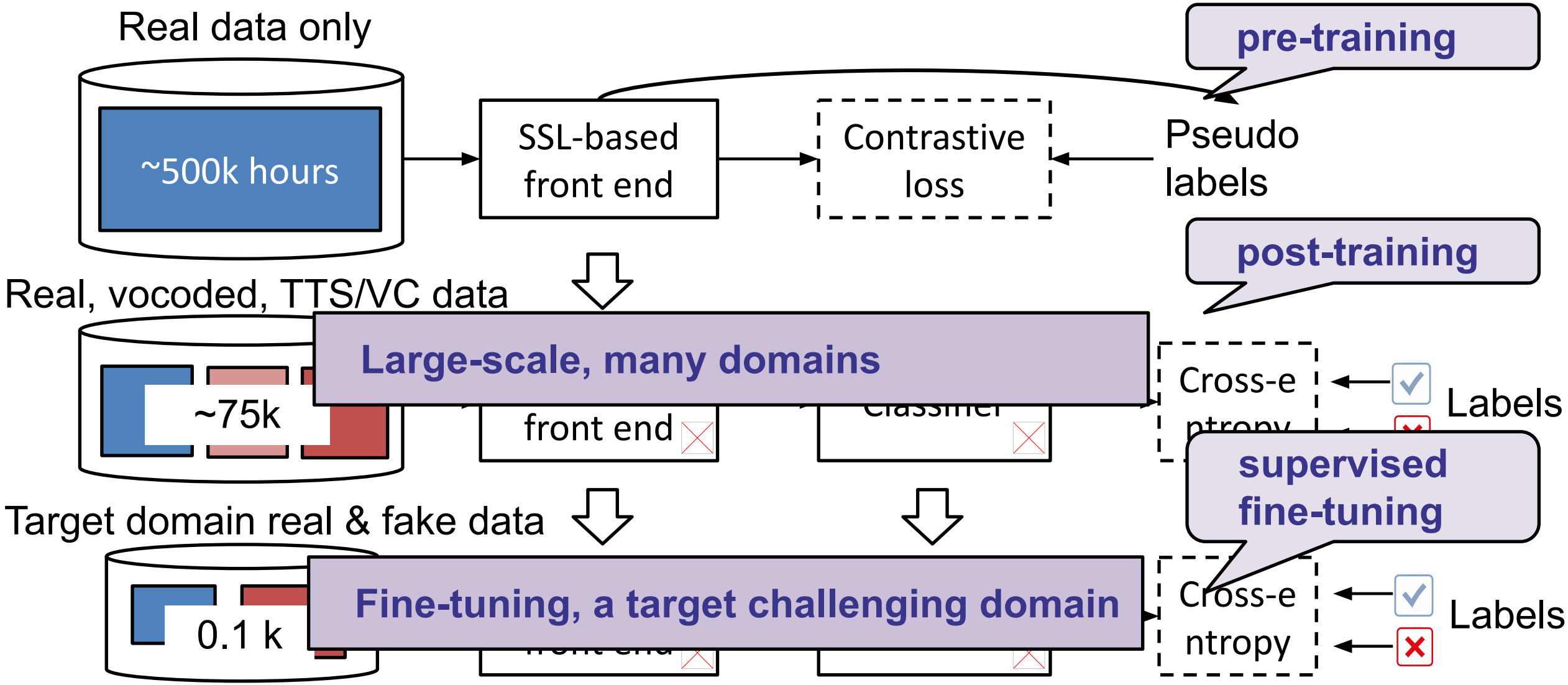
Introduction – Base method II



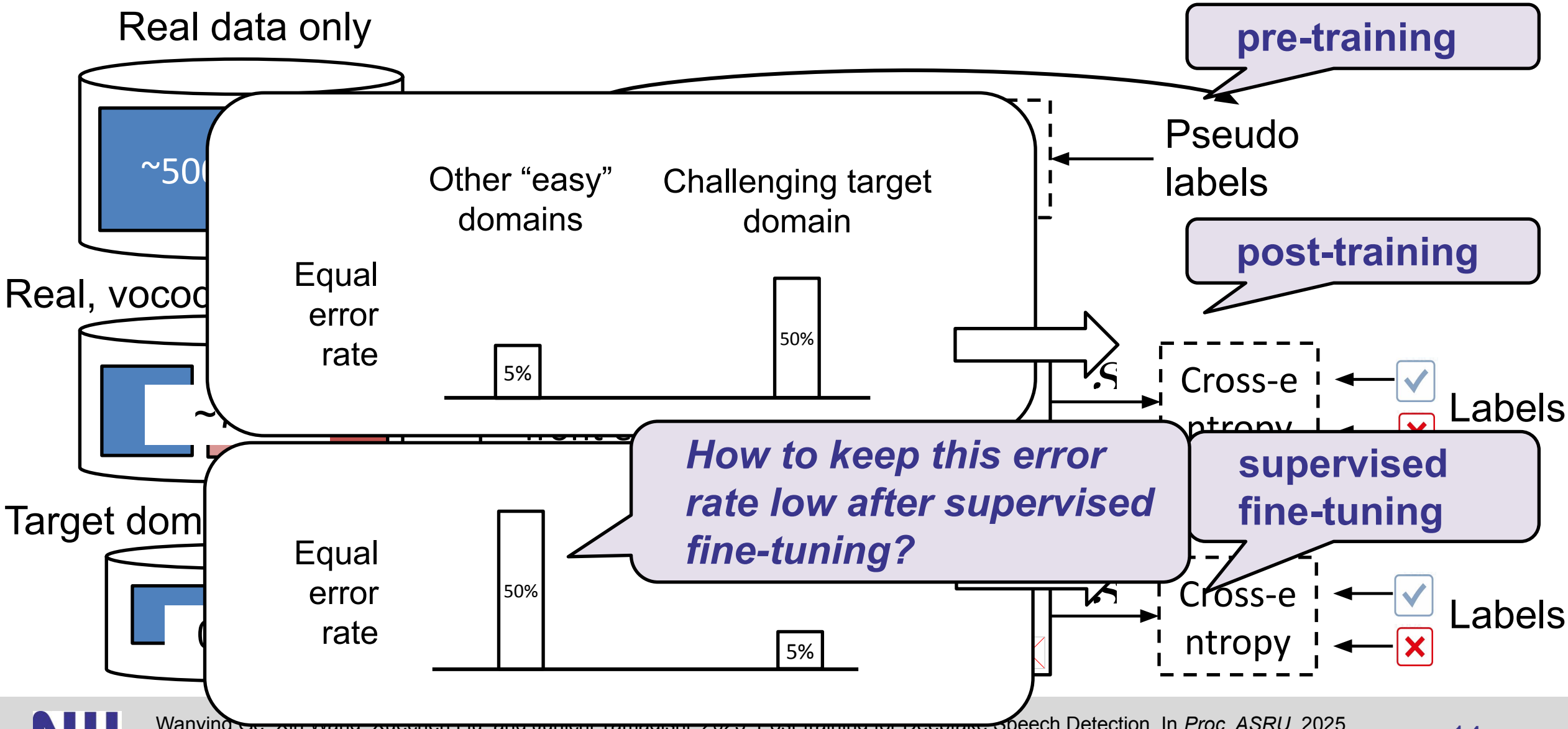
Introduction – Base method II



Introduction – Base method II



Introduction – Issue of catastrophic forgetting

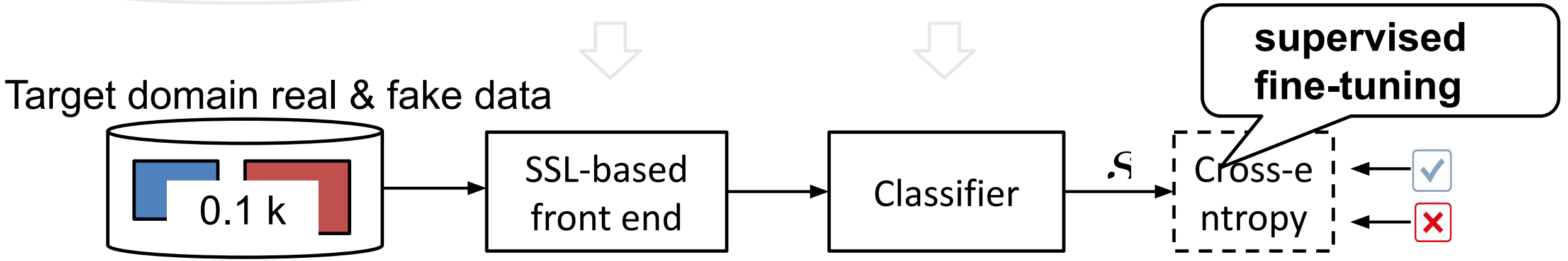


Wanying Ge, Xin Wang, Xuechen Liu, and Junichi Yamagishi. 2025. Post-training for Deepfake Speech Detection. In *Proc. ASRU*, 2025.

Method

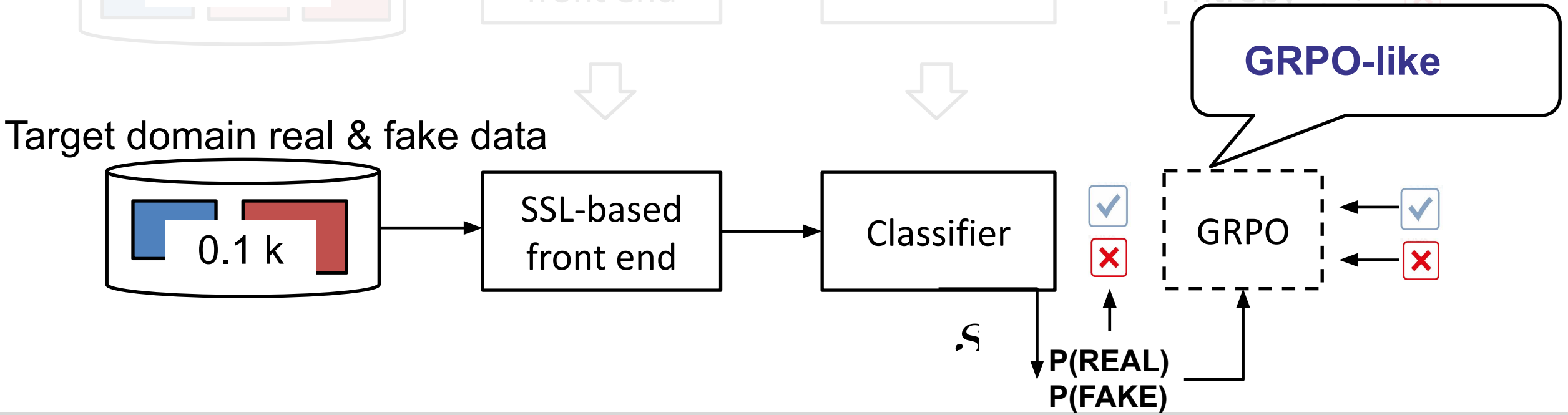
Method

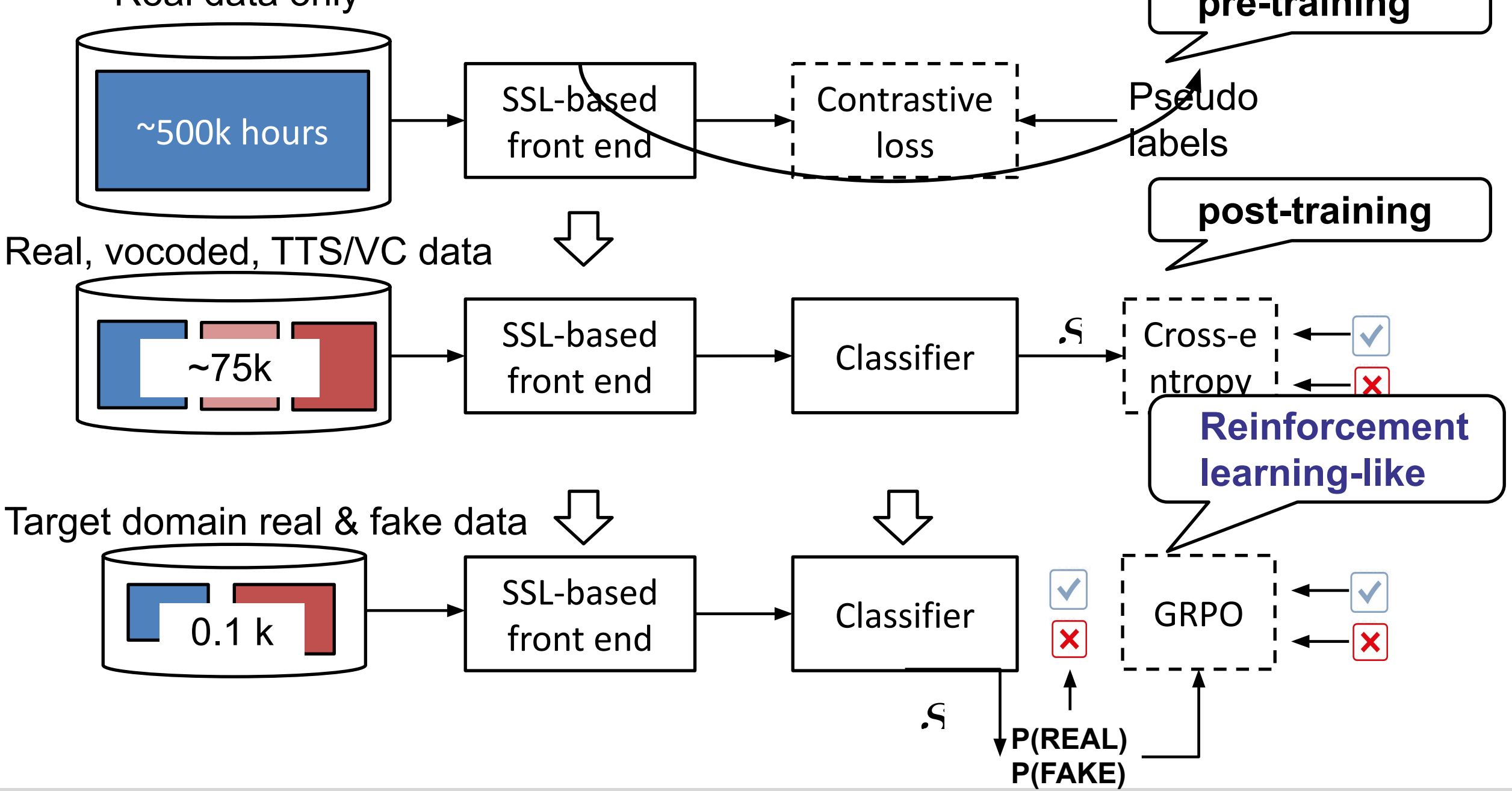
- **The same pre-training / post-training pipeline**



Method

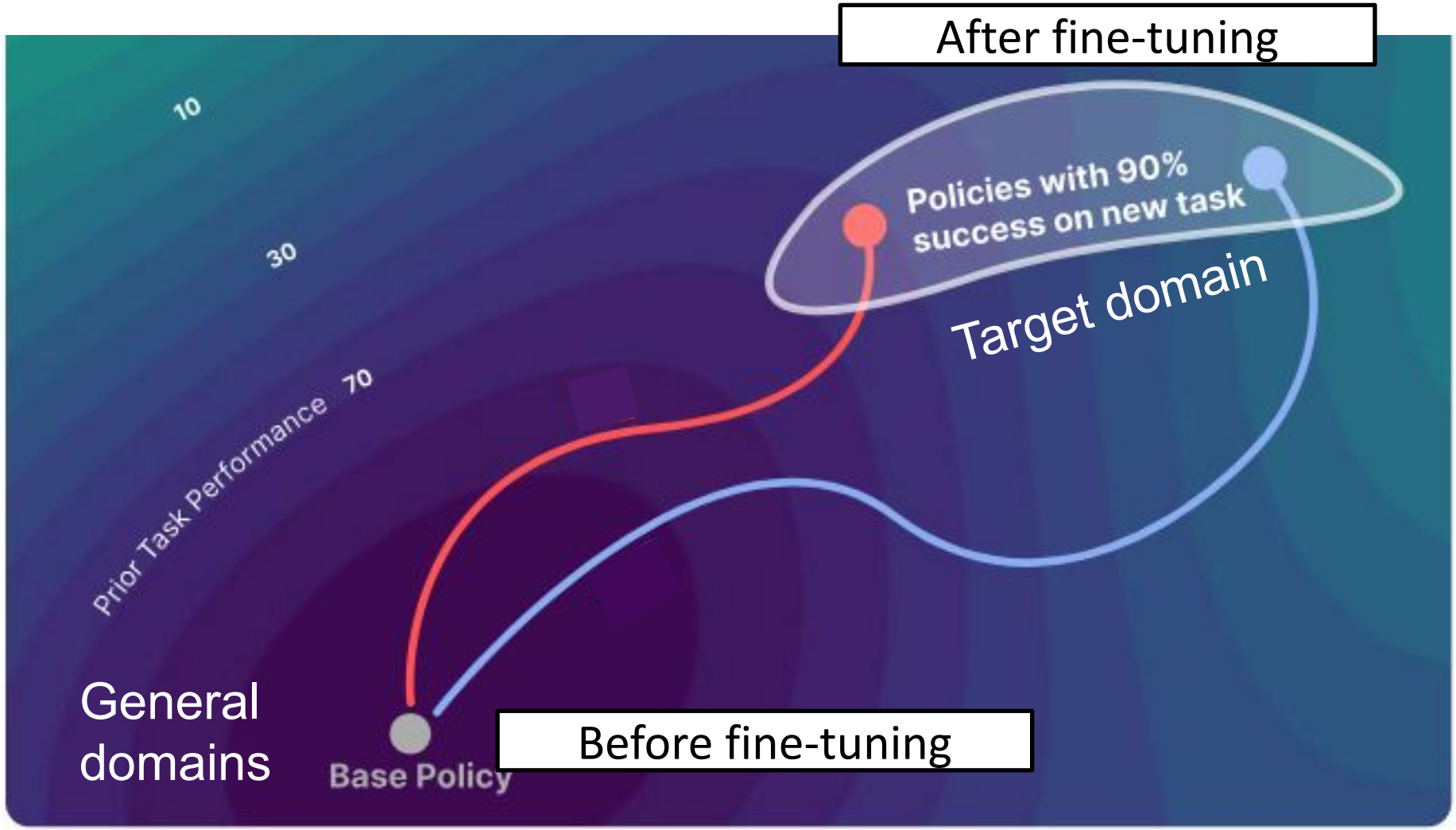
- **The same pre-training / post-training pipeline**
- **Replace the fine-tuning criterion**
  - ~~cross-entropy supervised fine-tuning (SFT)~~
  - similar to Group Relative Policy Optimization (GRPO) (Guo 2025)





Daya Guo, Dejian Yang, et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645, 8081 (September 2025), 633–638.

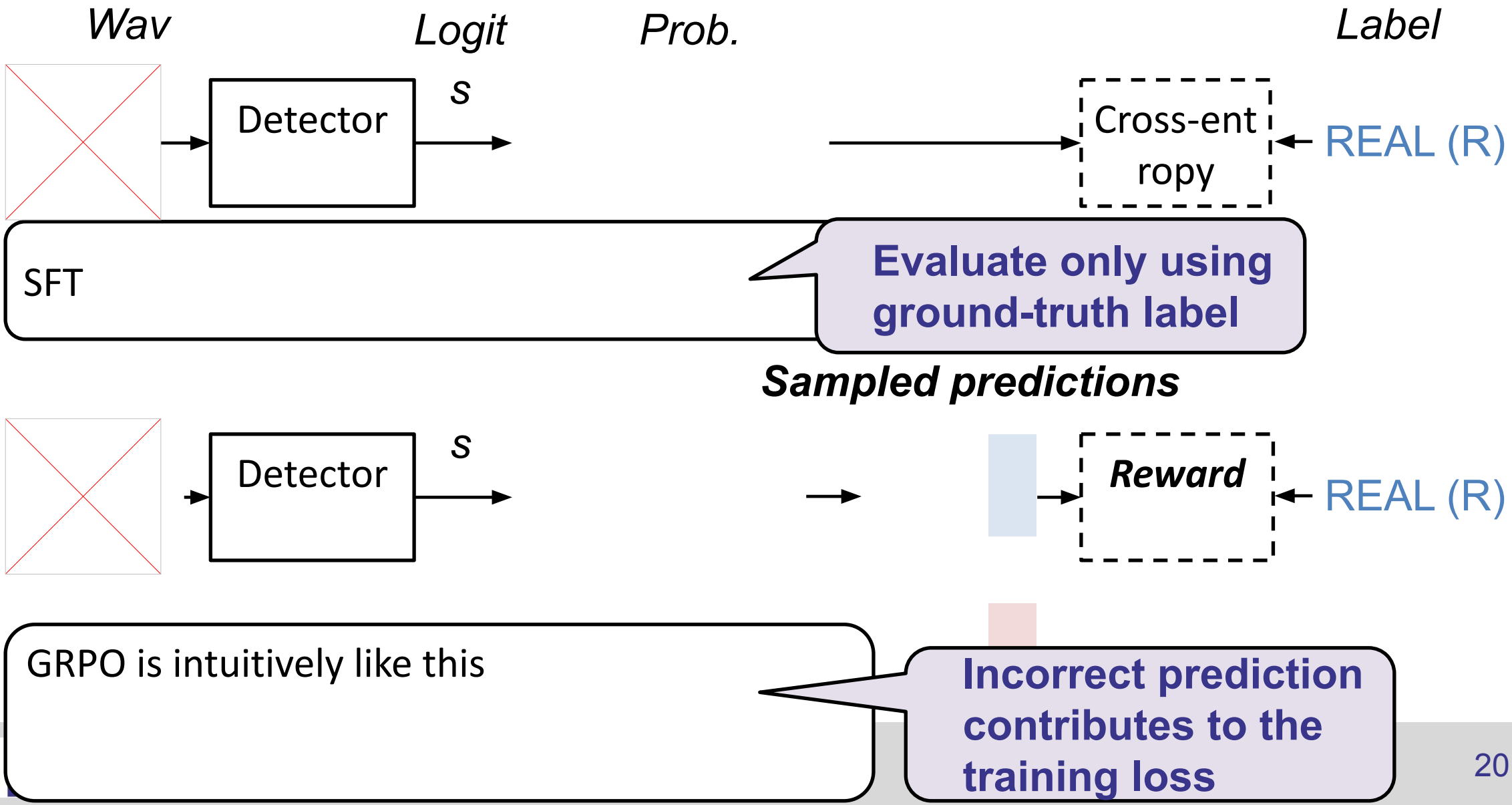
Method – why GRPO-like fine-tuning?



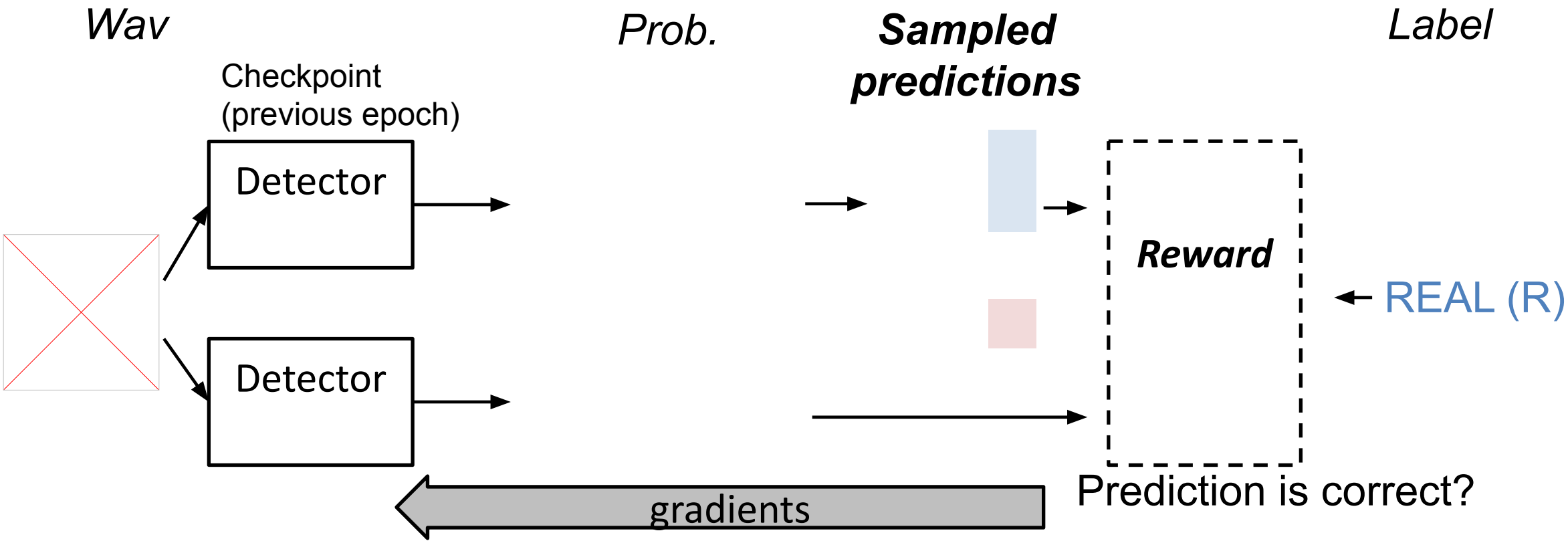
Reinforcement learning (e.g., GRPO)

Supervised training by cross entropy

Method – implementation



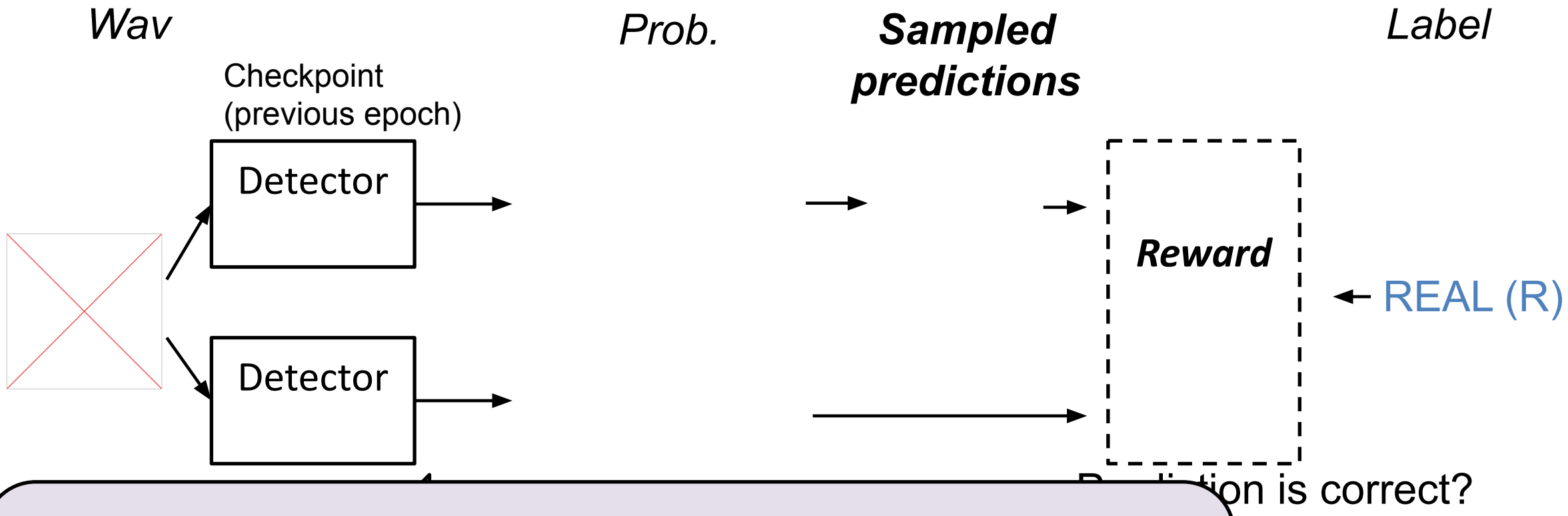
Method – implementation



GRPO is intuitively like this

Incorrect prediction contributes to the training loss

Method – implementation



See paper (appendix on arxiv) & code

1. ***Reward clipping, KLD, …***
2. ***Gradients of GRPO are different from those of SFT***

Experiment

Experiment setup

Research question

- Can GRPO-like fine-tuning mitigate catastrophic forgetting?

Data protocol (Ge 2025)

Pre-training → *Post training* → ***Fine-tuning*** → *Evaluation*

			In-domain	*Out-of-domain*
Real data e.g., XLS-R ~500k hours	Real, vocoded, & fake data (Ge 2025) ~75k hours	Deepfake-Eval-2024 train set, real & fake data (Chandra 2025) ~50 hours	Deepfake-Eval-2024 test set	• ADD2023 • FoR • DeepVoice • In-the-wild

We use post-trained checkpoints

Challenging target domain

Wanying Ge, Xin Wang, Xuechen Liu, and Junichi Yamagishi. 2025. Post-training for Deepfake Speech Detection. In *Proc. ASRU*, 2025.
Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, Jongwook Choi, Aerin Kim, and Oren Etzioni. 2025. Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024.

Experiment setup

Research question

- Can GRPO-like fine-tuning mitigate catastrophic forgetting?

Experimental systems

Pre-training → *Post training* → ***No fine-tuning***

Pre-training → *Post training* → ***Fine-tuning*** SFT (cross-entropy)

Pre-training → *Post training* → ***Fine-tuning*** GRPO

We use post-trained checkpoints

Experiment results

Highlighted results

- Model architecture: XLS-R-2B + global average pooling+linear output

Pre-training	*Post training*	***Fine-tuning***	*Average EER (%)*	
			Challenging target domain	*Other easier non-target domains*
v	v	x	22.64	2.69
v	v	SFT	10.26 ⇩	6.28 ⇧
v	v	GRPO		

Out-of-domain performance degraded after SFT

Experiment results

Highlighted results

- Model architecture: XLS-R-2B + global average pooling+linear output

Pre-training	*Post training*	***Fine-tuning***	*Average EER (%)*	
			Challenging target domain	*Other easier non-target domains*
v	v	x	22.64	2.69
v	v	SFT	10.26	6.28
v	v	GRPO	9.23 ⇩	2.69 ⇨

GRPO mitigates catastrophic forgetting

Experiment results

Highlighted results

- Model architecture: XLS-R-2B + global average pooling + linear output

Pre-training	*Post training*	***Fine-tuning***	*Detailed EER (%)*				
			Challenging target domain	*Other easier non-target domains*			
				ADD	FoR	DeepVoice	ItW
v	v	x	22.64	4.67	2.61	2.23	1.24
v	v	SFT	10.26	6.09	3.92	8.75	6.35
v	v	GRPO	9.23	5.34	0.47	2.76	2.19

GRPO mitigates catastrophic forgetting

Experiment results

Highlighted results

- Model architecture: XLS-R-2B + global average pooling + linear output

Pre-training	*Post training*	***Fine-tuning***	*Average EER (%)*	
			In-domain	*Out-of-domain*
v	v	x	22.64	2.69
v	v	SFT	10.26	6.28
v	v	GRPO	9.23	2.69
		w/o sampling	11.41	3.41

Sampling incorrect labels is necessary to GRPO

SSL

Training pipelines

Target domain EERs

Non-target domain

Table 1: *Test set EERs (%). Each EER was average after* three *training-evaluation rounds. Columns 'ave.' list average EERs of in- or out-of-domain. GRPO by default uses* $\beta = 0.04$ *(§3.1). In each column, higher EER is in darker shade.*

	Pre-train	Post-train	Fine-tune	DFE24 eva. set (in-domain) 4 s	10 s	13 s	30 s	50 s	ave.	Other eva. sets (out-of-domain) ADD23	FoR	DV	ItW	ave.	Row ID
XLS-R-2B	✓	✓	×	27.73	23.45	21.74	20.62	19.64	22.64	4.67	2.61	2.23	1.24	2.69	1
			SFT	12.17	10.28	10.11	9.50	9.23	10.26	6.09	3.92	8.75	6.35	6.28	2
			GRPO_s	12.45	11.07	11.06	10.44	10.98	11.20	3.95	0.65	6.72	2.44	3.44	3
			GRPO	11.06	9.99	9.90	9.45	9.27	9.93	5.34	0.47	2.76	2.19	2.69	4
	✓	✓	SFT → GRPO_s	12.17	10.20	10.07	9.36	9.43	10.24	6.42	3.11	8.20	6.24	5.99	5
			SFT → GRPO	11.53	9.94	9.88	8.93	8.76	9.81	5.78	2.75	7.04	5.89	5.37	6
			GRPO ($\beta = 0$)	11.58	10.32	10.00	9.64	9.51	10.21	4.43	0.47	6.50	2.71	3.53	7
			GRPO ($\beta = 1$)	19.37	15.61	15.27	14.63	14.24	15.83	4.02	1.34	1.66	1.13	2.04	8
			GRPO w/o neg.	13.30	11.60	11.23	10.64	10.31	11.41	6.69	0.83	3.09	3.14	3.44	9
	✓	×	SFT	12.24	10.34	10.20	9.63	9.83	10.45	17.81	29.83	16.21	13.49	19.34	10
			GRPO_s	14.33	12.65	12.02	11.27	10.87	12.23	22.88	29.71	23.67	23.03	24.82	11
			GRPO	12.93	11.18	11.06	10.24	10.35	11.15	17.57	25.68	25.19	19.00	21.86	12
MMS-1B	✓	✓	×	27.68	23.49	22.94	22.64	22.59	23.87	9.06	1.23	2.47	1.82	3.64	13
			SFT	11.86	9.87	9.51	9.07	9.08	9.88	4.14	1.35	5.35	6.59	4.36	14
	✓	✓	GRPO_s	12.47	10.90	10.22	9.45	9.47	10.50	3.73	0.73	2.68	4.52	2.92	15
			GRPO	11.87	10.41	9.87	9.21	9.47	10.17	4.25	1.78	2.84	2.98	2.96	16
MMS-300M	✓	✓	×	32.87	30.08	29.38	26.33	27.51	29.23	7.96	1.51	2.35	2.91	3.68	17
			SFT	14.45	11.83	11.55	11.90	11.41	12.23	8.81	3.61	4.37	14.17	7.74	18
	✓	✓	GRPO_s	13.46	11.60	11.16	11.27	10.74	11.65	5.70	1.40	2.91	8.31	4.58	19
			GRPO	14.05	11.73	11.42	11.12	11.37	11.94	4.99	1.57	2.64	5.91	3.78	20

Equal error rate (EER)



Average EER of 3 runs

Experiment results

Other results in the paper

- Similar findings when using other multi-lingual SSL front-ends
  - MMS-1B, MMS-300M

- Other pipelines were sub-optimal

Pre-training → *Post training* → ***Fine-tuning*** SFT then GRPO (Deepseek R1?)

Pre-training ⟶ ***Fine-tuning*** SFT/GRPO

- Ablation study on GRPO
  - KLD regularization is NOT a must

Summary

Summary

- **Does GRPO-like fine-tuning mitigate catastrophic forgetting?**
  - YES, at least in our experiments

Pre-training → *Post training* → ***Fine-tuning*** GRPO

- **Limitations**
  - Not all factors are ablated
    - Size of sampling
    - Time to update old checkpoint
  - Test on other target domains

- **Future work**
  - GRPO for post training?
  - Reward beyond simple label comparison?
  - …

Thank you



Code

github/nii-yamagishilab/AntiDeepfake

Appendix

A.2. Details on GRPO implementation

In the implementation, we follow the GRPO paper and added the clipping function:

$$\mathcal{L}_{\mathrm{RL}}(\Theta) = \mathbb{E}_{x\sim\mathcal{D},\tilde{y}\sim p_{\widehat{\Theta}}(\mathbf{y}|\mathbf{x})}\left[\left(\min\left(\frac{p_{\Theta}(\tilde{y}|\boldsymbol{x})}{p_{\widehat{\Theta}}(\tilde{y}|\boldsymbol{x})}\mathcal{A}(\tilde{y},y), \mathrm{clip}\left(\frac{p_{\Theta}(\tilde{y}|\boldsymbol{x})}{p_{\widehat{\Theta}}(\tilde{y}|\boldsymbol{x})}, 1-\epsilon, 1+\epsilon\right)\mathcal{A}(\tilde{y},y)\right) - \beta\mathcal{L}_{\mathrm{p}}(\dot{\Theta},\Theta)\right)\right], \quad (12)$$

wherein the advantage function has a floor value in the denominator in case the standard deviation is 0.

$$\mathcal{A}(\tilde{y}_i, \{\tilde{y}_j\}_{j=1}^{G}, y) = \frac{r(\tilde{y}_i, y) - \bar{r}(\{\tilde{y}\}_{j=1}^{G}, y)}{\sigma(\{\tilde{y}\}_{j=1}^{G}, y) + \rho}, \quad (13)$$

0 If equal, else 1

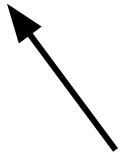
The hyper-parameters are set to:

Table 2: *Hyper-paramters of the implemented GRPO*

Hyper-parameter	Value
ϵ	0.2
ρ	1e-5
#. of training steps before updating $\widehat{\Theta}$	1k
#. of sampled $\tilde{y}$ per $\boldsymbol{x}$	64

Compare gradients

without group-mean / var
without clipping, KLD …



GRPO-like

Cross entropy

(appendix https://arxiv.org/abs/2603.02914)

Compare gradients

without group-mean / var
without clipping, KLD …

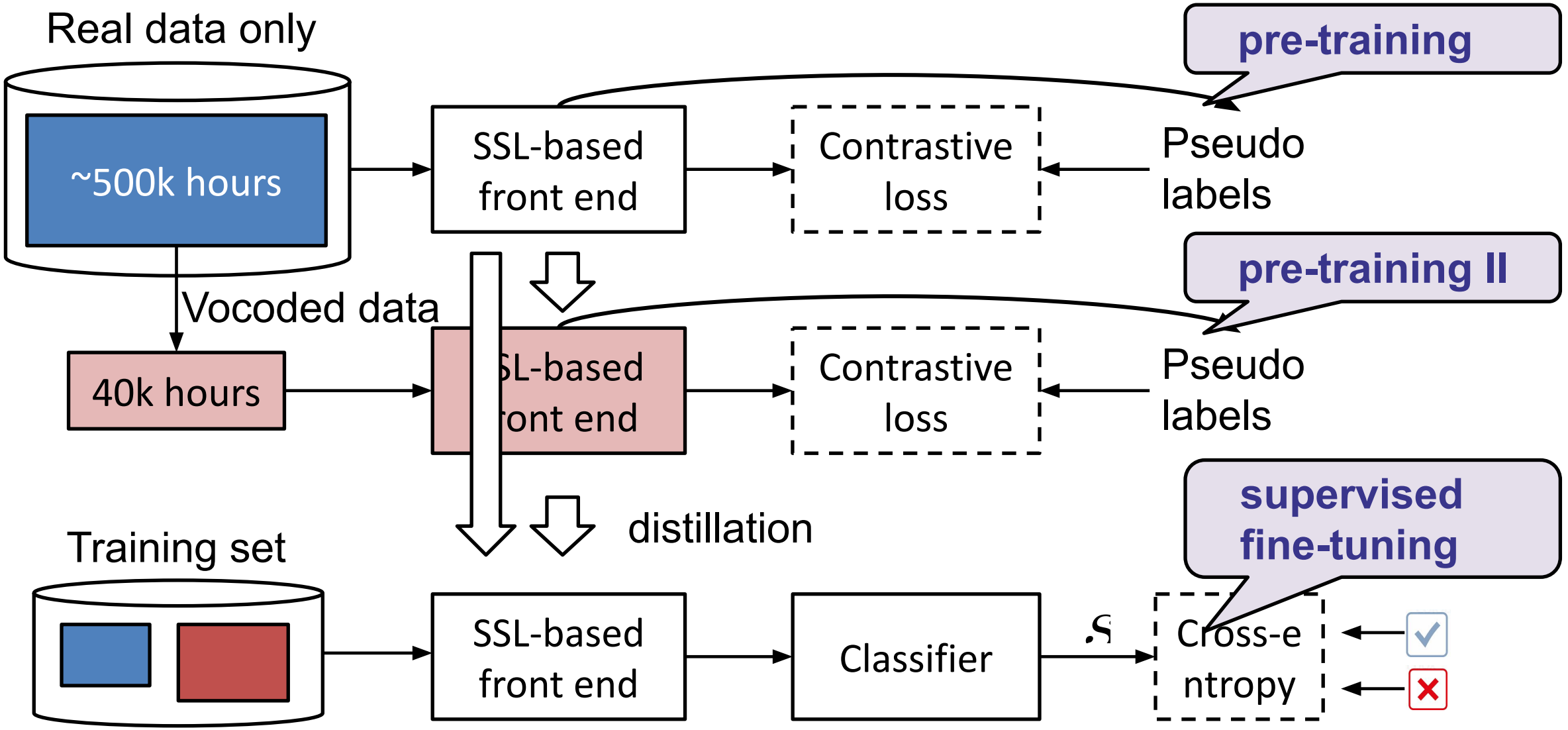
GRPO-like

Mixup

$$\nabla_\Theta \mathcal{L}_{\text{Mixup}} \propto -\left(\frac{\lambda}{p_\Theta(y_i|\tilde{\mathbf{x}})} - \frac{1-\lambda}{1-p_\Theta(y_i|\tilde{\mathbf{x}})}\right)\nabla_\Theta p_\Theta$$

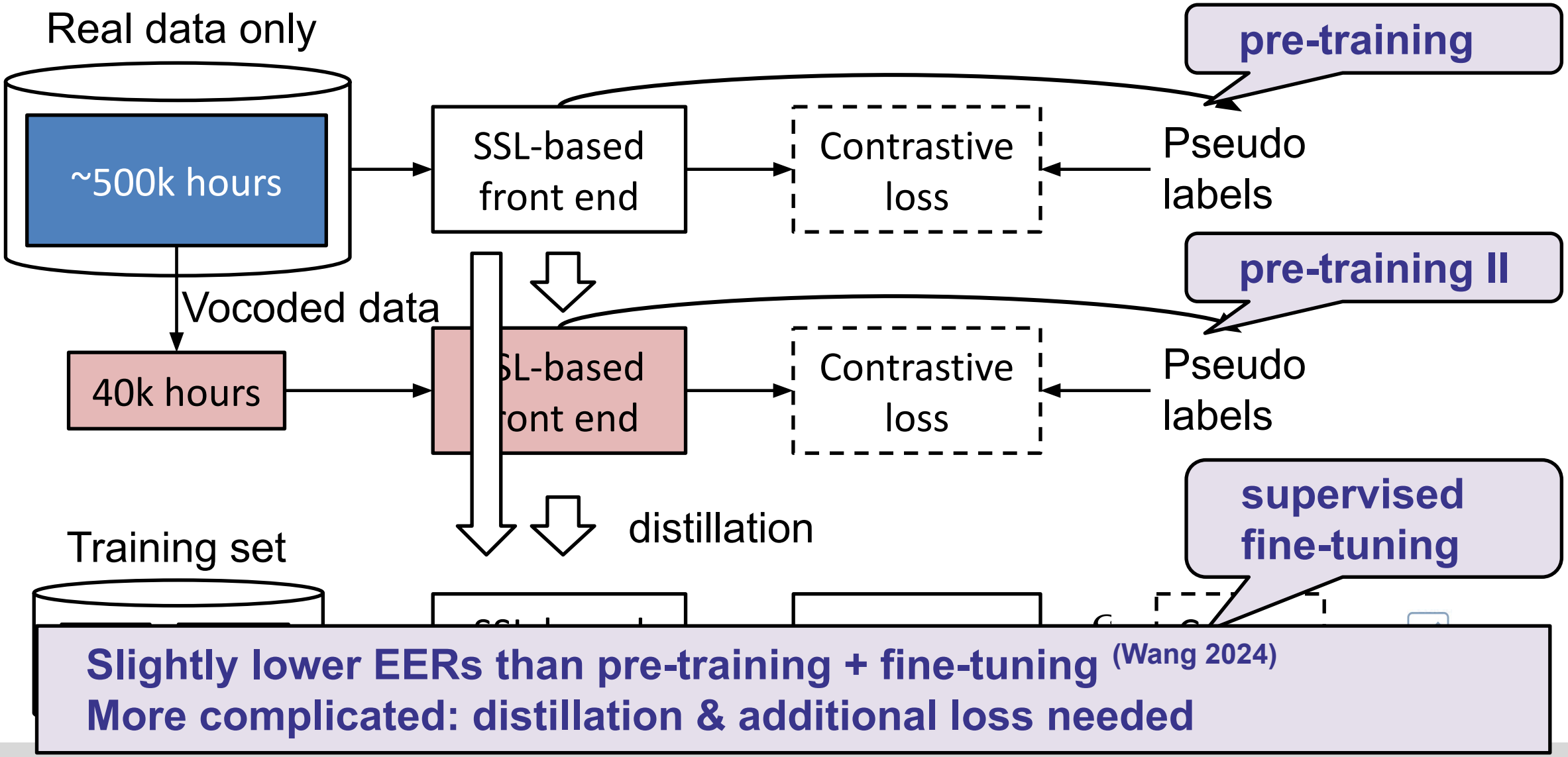
(appendix https://arxiv.org/abs/2603.02914)

Introduction – Base method



Xin Wang and Junichi Yamagishi. 2023. Spoofed Training Data for Speech Spoofing Countermeasure Can Be Efficiently Created Using Neural Vocoders. In *Proc. ICASSP*, 2023.
Xin Wang and Junichi Yamagishi. 2024. Can Large-Scale Vocoded Spoofed Data Improve Speech Spoofing Countermeasure with a Self-Supervised Front End? In *Proc. ICASSP*, 2024. 10311–10315.

Introduction – Base method



Slightly lower EERs than pre-training + fine-tuning (Wang 2024)
More complicated: distillation & additional loss needed

Xin Wang and Junichi Yamagishi. 2023. Spoofed Training Data for Speech Spoofing Countermeasure Can Be Efficiently Created Using Neural Vocoders. In *Proc. ICASSP*, 2023.
Xin Wang and Junichi Yamagishi. 2024. Can Large-Scale Vocoded Spoofed Data Improve Speech Spoofing Countermeasure with a Self-Supervised Front End? In *Proc. ICASSP*, 2024. 10311–10315.