UniGen: Universal Domain Generalization for Sentiment Classification via Zero-shot Dataset Generation

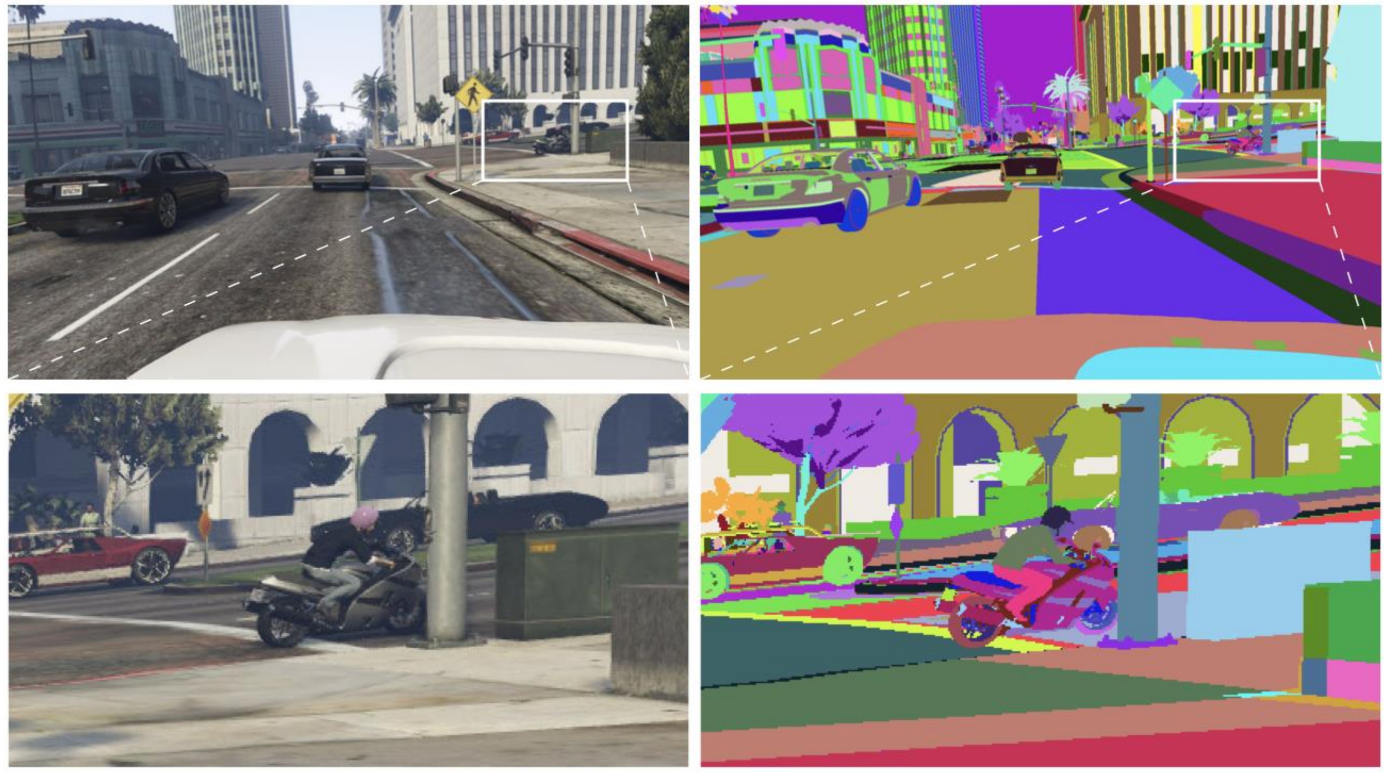
Juhwan Choi, Yeonghwa Kim, Seunguk Yu, Jungmin Yun and Youngbin Kim

Chung-Ang University

Synthetic Data in Deep Learning

Researchers are increasingly exploring the use of **synthetic data** in deep learning

- For example, a scene from a computer game was used as training data for a semantic segmentation task[1]
- In NLP tasks, the data generated by language model was used for data augmentation[2]



Algorithm 1: LAMBADA

Input: Training dataset D_{train}
Classification algorithm $\mathcal{A}$
Language model $\mathcal{G}$
Number to synthesize per class $N_1, \ldots, N_q$

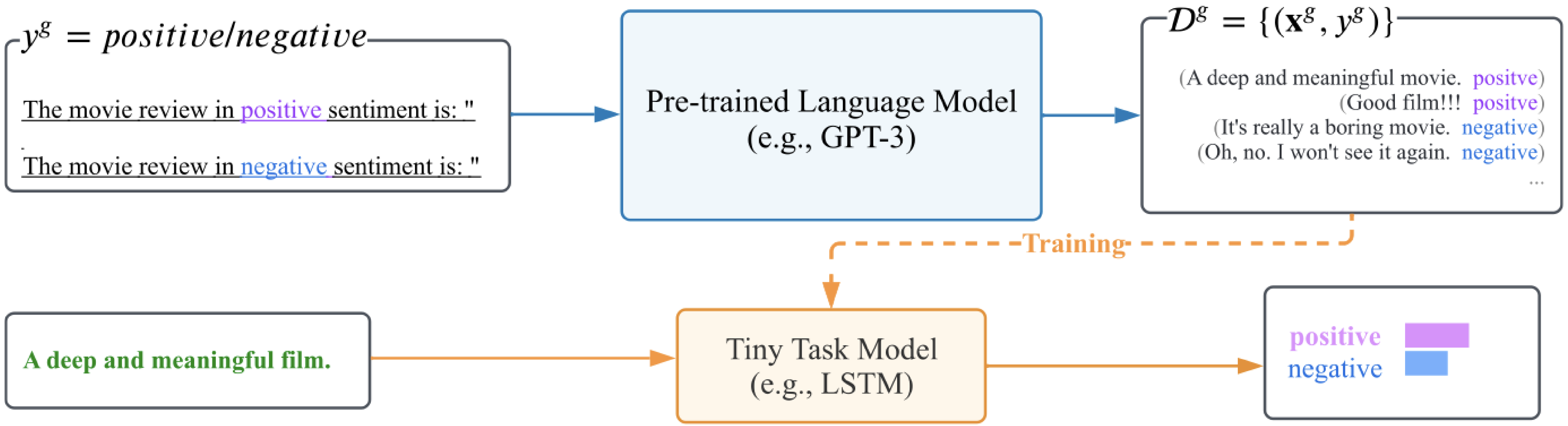
Train a baseline classifier h from D_{train} using $\mathcal{A}$
Fine-tune $\mathcal{G}$ using D_{train} to obtain $\mathcal{G}_{tuned}$
Synthesize a set of labeled sentences D^* using $\mathcal{G}_{tuned}$
Filter D^* using classifier h to obtain $D_{synthesized}$
return $D_{synthesized}$

1. Richter et al., *Playing for Data: Ground Truth from Computer Games*, ECCV 2016.
2. Anaby-Tavor et al., Do Not Have Enough Data? Deep Learning to the Rescue!, AAAI 2020.

ZeroGen: End-to-end Training with Synthetic Data

Recently, **ZeroGen** proposed to solely use synthetic data to train a small model[1]

- This approach begins by generating synthetic data from a pre-trained language model (PLM) with a prompt
- With the generated synthetic data, we train a small model for inference
- ZeroGen enables efficient zero-shot learning, as
  - They use synthetic data generated by PLM and do not require human-annotated data
  - They use the small model at inference and do not require PLM after the generation of synthetic data
- The small model trained with synthetic data is called tiny task model (TAM)

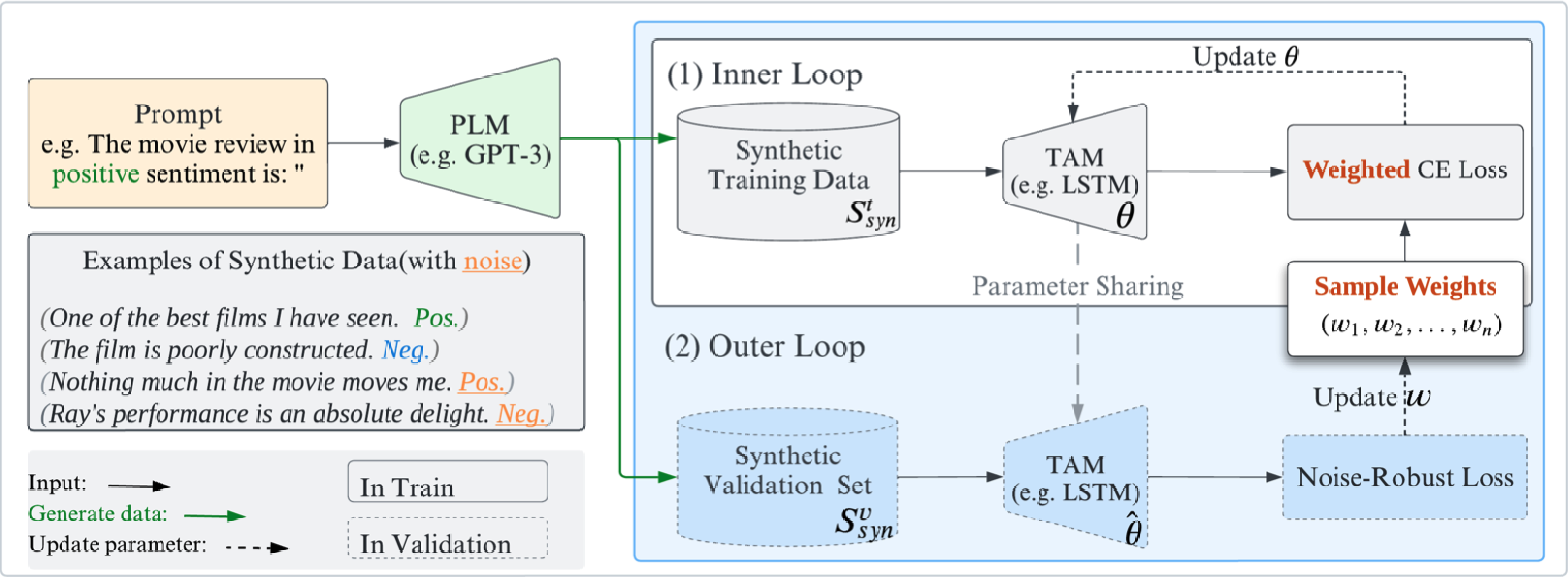


1. Ye et al., [ZeroGen: Efficient Zero-shot Learning via Dataset Generation](), EMNLP 2022.

Drawback of ZeroGen: Noisy Data

However, **ZeroGen approach may generate noisy data**

- These noisy data include data with noisy label, or unrelated data to given prompt
- To mitigate this issue, **SunGen**[1] proposed to learn weights of each synthetic data
- After learning the weights, SunGen selects data with higher weights (i.e., higher quality data)

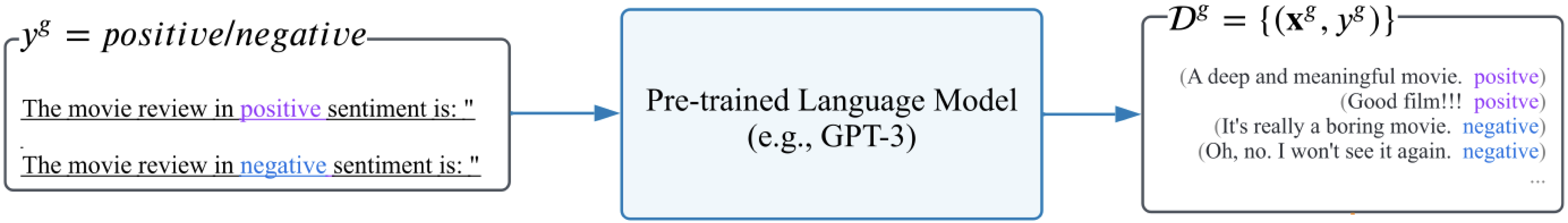


1. Gao et al., *Self-Guided Noise-Free Data Generation for Efficient Zero-Shot Learning*, ICLR 2023.

Drawback of ZeroGen: Domain Limitation

Furthermore, ZeroGen and similar studies generate a TAM tailored to specific domain

- For instance, the example in this figure will lead to a TAM for movie reviews
- This restricts the real-world applicability of methods based on synthetic data
- Unlocking this limitation will enhance the usefulness of synthetic data-based approaches
- In this paper, we aim to effectively distill the domain generalizability of PLMs into TAMs



UniGen: Universal Domain Generalization

We propose **UniGen**, a novel method for enabling domain generalizability for TAMs

- UniGen allows TAMs to achieve domain generalizability, unlike previous methods
- We suggest various components for UniGen to accomplish the domain generalizability
- We maximize the efficiency of synthetic data-based methods by enabling the training of a single TAM that can be universally deployed across multiple domains

	Learning without Human-annotated Data	Domain Generalizability	Light Inference	Handling Noise of Generated Data
Task-specific Fine-tuning	✗	✗	✓	
Previous Domain Generalization (Tan et al., 2022)	✗	✓	✓	
PROMPTING	✓	✓	✗	
ZEROGEN (Ye et al., 2022a)	✓	✗	✓	✗
PROGEN & SUNGEN (Ye et al., 2022b; Gao et al., 2023)	✓	✗	✓	✓
UNIGEN (Ours)	✓	✓	✓	✓

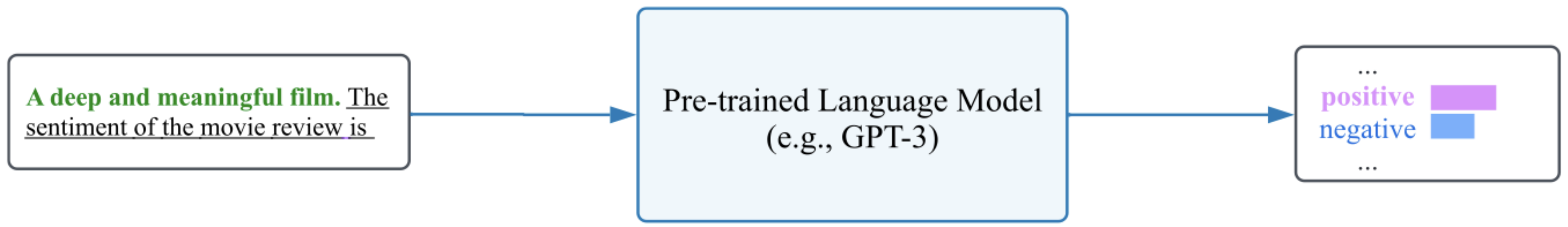
Preliminary: Prompting and ZeroGen

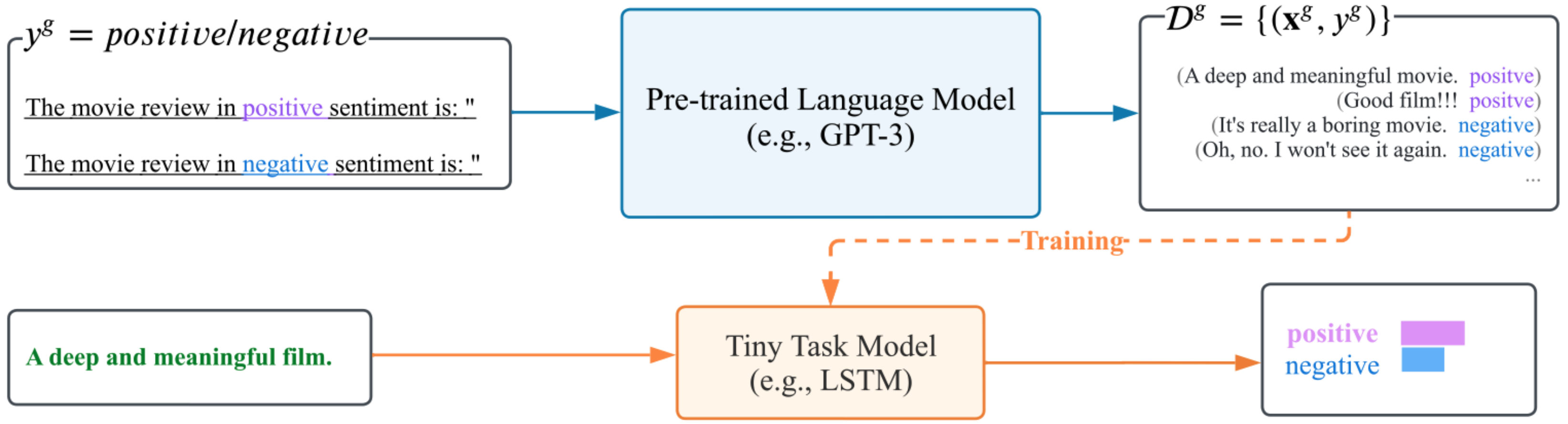
Prompting uses PLM to directly infer the label of input text based on prompt

- The probability of each label y_i is represented as $p(y_i|\boldsymbol{x}_i) = \mathcal{P}(\mathcal{M}(y_i)|\mathcal{T}(\boldsymbol{x}_i))$
- Where $\mathcal{P}$, $\mathcal{M}$, and $\mathcal{T}$ denote PLM, verbalizer, and prompt

ZeroGen guides PLM to generate synthetic data $\boldsymbol{x}_{syn}$ based on given prompt and label

- This synthetic data generation process is denoted as $\boldsymbol{x}_{syn} \sim \mathcal{P}(\cdot\,|\mathcal{T}_{task}(y_{syn}))$
- $\mathcal{T}_{task}$ denotes the prompt to guide the generation process, which specifies the domain
- We use these generated $(\boldsymbol{x}_{syn},\ y_{syn})$ to train TAMs





Proposed Method: Universal Prompt

First, we transform the prompt to generate synthetic data

- Previous methods used $\mathcal{T}_{task}$ such as "The *movie review* in positive sentiment is:"
  - This restricts the generated synthetic data to be specified for movie review
- Instead, we suggest to use **universal prompt** $\mathcal{T}_{uni}$, "The *text* in positive sentiment is:"
  - The generation process is modified as $\boldsymbol{x}_{syn} \sim \mathcal{P}(\cdot \mid \mathcal{T}_{uni}(y_{syn}))$
  - This allows the generation of synthetic data without any specific domain
- We train a single TAM based on synthetic data generated by this universal prompt

Domain	Prompt
Movie	The *movie review* in [positive/negative] sentiment is:
Products	The *product review* in [positive/negative] sentiment is:
Restaurant	The *restaurant review* in [positive/negative] sentiment is:
Electronics	The *electronics product review* in [positive/negative] sentiment is:
Tweet	The *tweet* in [positive/negative] sentiment is:
UNIGEN & PROMPTING	The *text* in [positive/negative] sentiment is:

Proposed Method: Pseudo-relabeling and Filtering

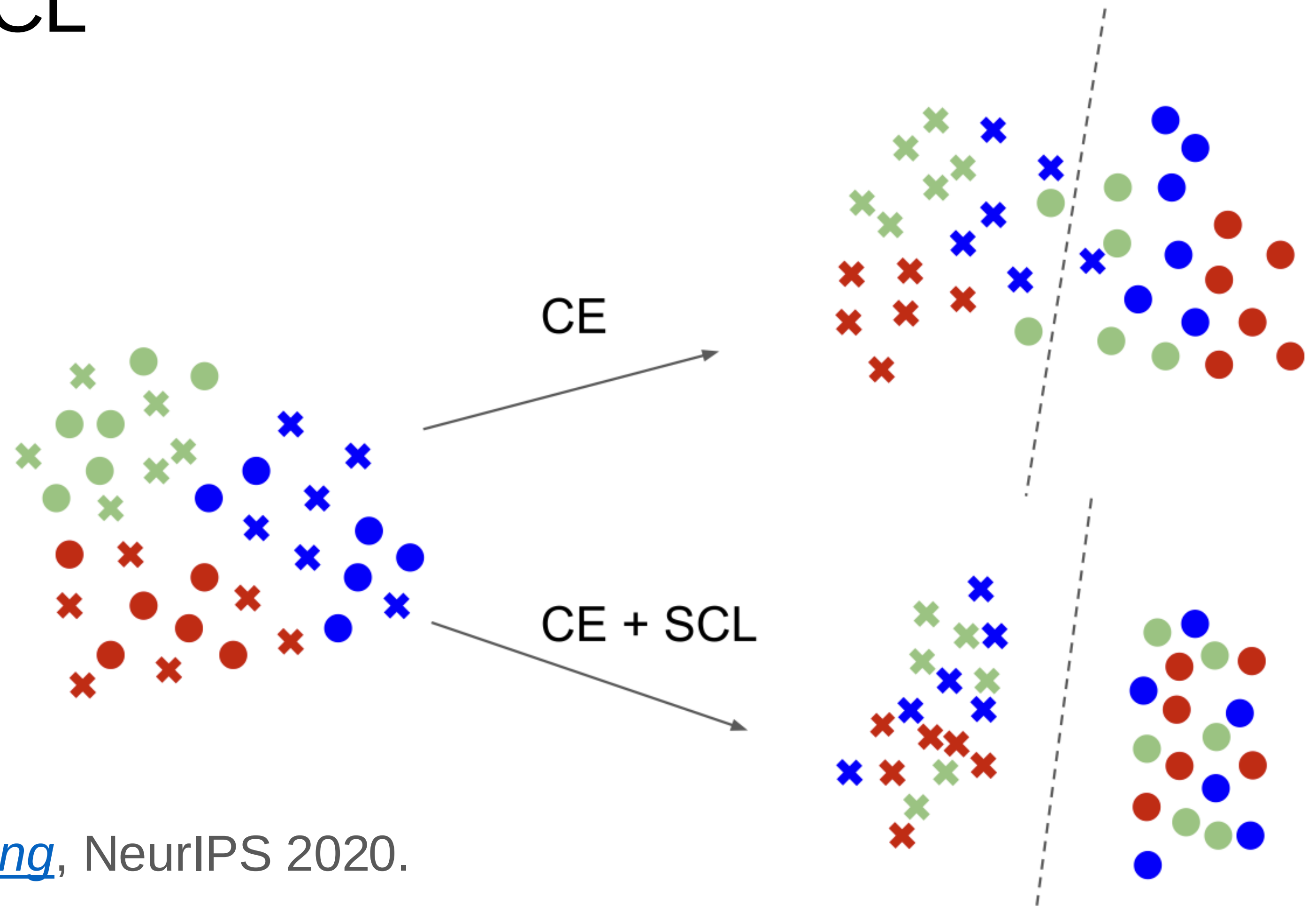
We propose **pseudo-relabeling** procedure to prevent the generation of noisy data

- For each generated synthetic data, we use PLM to acquire its soft label
- We first obtain the logits of each y_i for $\boldsymbol{x}_{syn}$ as $\ell\left(y_i \middle| \boldsymbol{x}_{syn}\right) = \mathcal{P}(\mathcal{M}(y_i)|\mathcal{T}_{uni}(\boldsymbol{x}_{syn}))$
- Next, we acquire pseudo-label $\hat{y}_i = p\left(y_i \middle| x_{syn}\right) = \frac{\exp(\ell\left(y_i|x_{syn}\right)/\tau_{RE})}{\sum_j \exp(\ell\left(y_j|x_{syn}\right)/\tau_{RE})}$
  - τ_{RE} denotes temperature for softmax function
- We use $\hat{y}_i$ for training TAMs instead of original y_i
- Additionally, we suggest two filtering strategies using $\hat{y}_i$
  - We remove data with $\hat{y}_i$ that differ from designated y_i, filtering out data with noisy label
  - We exclude data if $\hat{y}_i$ does not exceed a threshold T_{RE}, eliminating ambiguous data

Proposed Method: Supervised Contrastive Learning

We use **supervised contrastive learning**[1] to enhance domain generalizability of TAMs[2]

- The SCL loss is defined as $\mathcal{L}_{SCL} = -\sum_{\boldsymbol{z}_i \in B} \frac{1}{|P(i)|} log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau_{SCL})}{\sum_{\boldsymbol{z}_a \in A(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau_{SCL})}$
- The usage of SCL helps TAMs to learn domain-agnostic features
- Additionally, we adopt memory bank[3] and momentum encoder[4] to improve the effectiveness of SCL

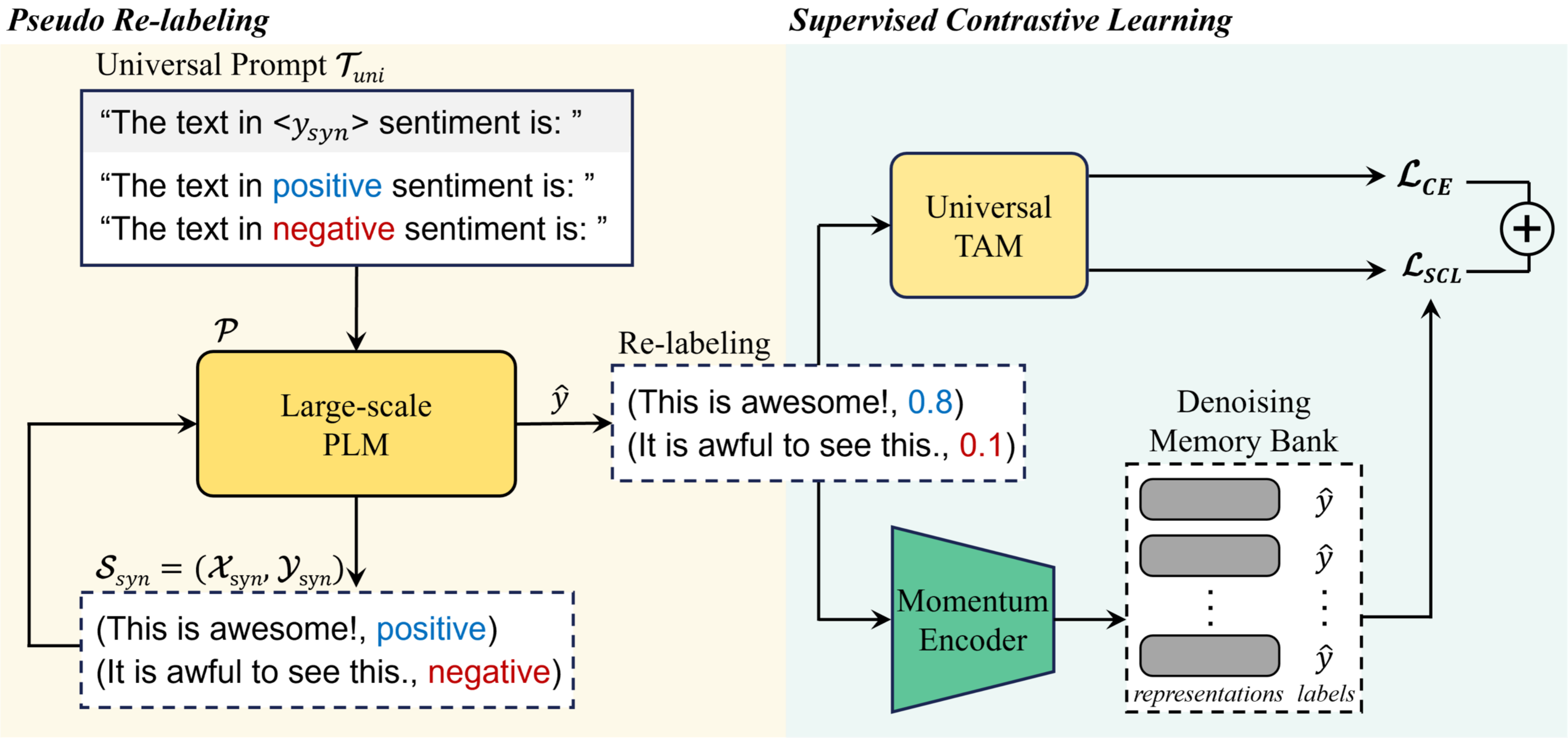


1. Khosla et al., *Supervised Contrastive Learning*, NeurIPS 2020.
2. Tan et al., *Domain Generalization for Text Classification with Memory-Based Supervised Contrastive Learning*, COLING 2022.
3. Wu et al., *Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination, CVPR 2018.*
4. He et al., *Momentum Contrast for Unsupervised Visual Representation Learning*, CVPR 2020.

Proposed Method: Denoising Memory Bank

Based on the usage of SCL for training TAMs, we propose **denoising memory bank**

- We first learn the weight of each synthetic data following the method of SunGen[1]
- Given the weights of data, we only store samples whose weights are larger than T_{MB} to the memory bank
    - T_{MB} denotes the threshold for memory bank
- This ensures the exclusive use of high-quality samples in the memory bank, thereby improving its effectiveness when using synthetic data



1. Gao et al., *Self-Guided Noise-Free Data Generation for Efficient Zero-Shot Learning*, ICLR 2023.

Experiment: Experimental Setup

We used seven different **datasets** across five domains:

- Movie Review: SST-2, IMDB, Rotten Tomatoes
- Product Review: Amazon review dataset
- Restaurant Review: Yelp review dataset
- Electronics Product Review: Customer review dataset
- Tweets from Twitter: Twitter sentiment classification

Models:

- PLMs to generate synthetic data: GPT2-XL (1.5B parameters)
- TAMs to train with synthetic data: LSTM (<7M), DistilBERT (66M), RoBERTa (110M)

Baselines:

- Prompting: Zero-shot classification using PLM
- ZeroGen: Generate 200,000 data for each domain and train different TAMs
- SunGen: Generate 1,000,000 and extract 200,000 data with high quality for each domain

Experiment: Domain Generalizability of UniGen

UniGen TAM performance rapidly improves with increasing of TAM parameter sizes

- Especially, RoBERTa TAM trained with UniGen exceeds PLM prompting in terms of average performance
- This suggests that UniGen can achieve domain generalizability of PLMs using a single TAM, different from previous methods

Model	#Param	Training Domain	Setup	SST-2	IMDB	Rotten	Amazon	Yelp	CR	Tweet	Average
Test Domain					Movie		Products	Restaurant	Electronics	Tweet	
GPT2-XL	1.5B	-	PROMPTING	82.15	70.26	77.56	79.06	78.04	80.30	80.38	78.25
DistilBERT	66M	Movie	ZEROGEN	80.06	69.13	74.73	73.02	72.77	73.59	74.83	74.02
			SUNGEN	**82.43**	**70.59**	**76.37**	74.13	73.56	75.14	75.96	75.45
		Products	ZEROGEN	71.04	64.99	65.57	74.54	71.89	74.57	71.93	70.65
			SUNGEN	72.35	65.95	66.84	**76.92**	74.98	75.84	73.01	72.27
		Restaurant	ZEROGEN	77.32	65.47	68.86	74.01	77.94	74.89	73.74	73.18
			SUNGEN	78.93	67.12	69.92	74.93	**80.67**	76.06	75.28	74.70
		Electronics	ZEROGEN	73.77	66.14	66.78	72.38	73.21	78.82	74.58	72.24
			SUNGEN	74.49	67.19	68.29	73.49	75.34	**80.49**	75.37	73.52
		Tweet	ZEROGEN	73.98	66.58	67.43	72.88	71.86	75.68	80.86	72.75
			SUNGEN	75.12	67.53	69.06	73.64	72.73	78.17	**82.46**	74.10
		-	UNIGEN	77.67	67.81	73.16	75.06	74.81	79.86	81.41	**75.68**
RoBERTa	110M	Movie	ZEROGEN	84.38	73.03	78.38	77.38	76.83	77.36	77.94	77.90
			SUNGEN	**85.24**	**74.09**	**79.19**	78.56	77.61	78.21	79.72	78.95
		Products	ZEROGEN	79.14	71.16	70.92	79.94	75.79	76.35	80.17	76.21
			SUNGEN	81.51	71.28	72.67	**81.50**	77.76	78.55	81.94	77.87
		Restaurant	ZEROGEN	82.87	70.71	69.58	78.61	81.47	76.43	79.51	77.03
			SUNGEN	83.65	71.40	71.05	79.42	**82.72**	77.60	80.92	78.11
		Electronics	ZEROGEN	76.82	69.42	67.89	75.02	76.53	81.24	76.51	74.78
			SUNGEN	77.51	71.23	68.77	76.91	78.33	83.49	79.03	76.47
		Tweet	ZEROGEN	78.43	68.31	72.25	78.09	74.61	79.08	82.96	76.25
			SUNGEN	82.19	70.62	73.21	79.84	76.27	81.46	83.25	78.12
		-	UNIGEN	84.86	72.24	78.82	80.79	79.15	**86.37**	**87.89**	**81.45**

Experiment: Example of Generated Data

The generated data from UniGen shows that:

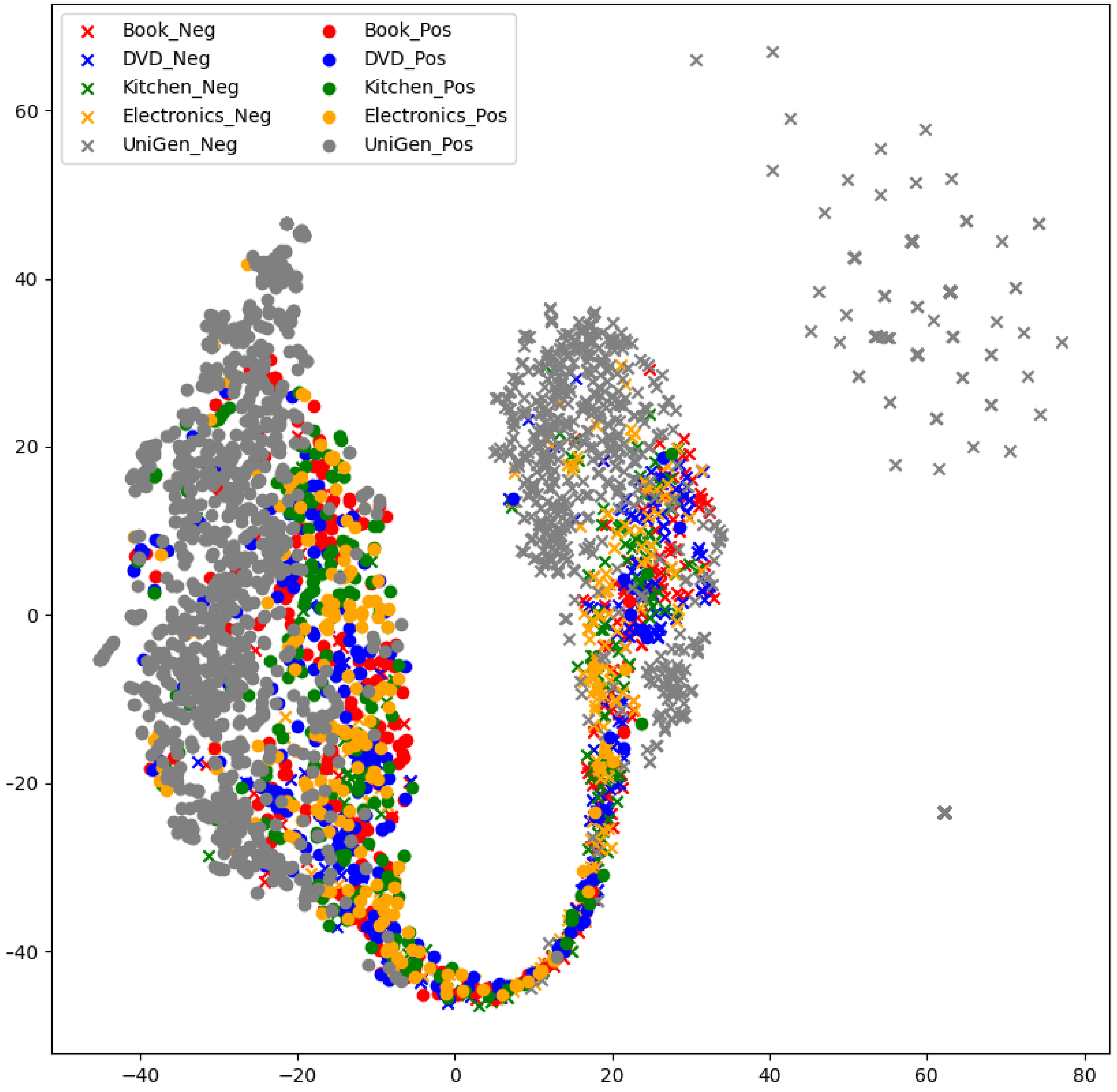
- UniGen can generate domain-agnostic data with soft label
- This domain-agnostic data enables TAMs to generalize across various domains
- The soft label allows to effectively distil the degree of the label from PLMs to TAMs

Positive Examples	Labels
You are a person who is hardworking, honest, and reliable. You have a good sense of humor, and you love being in charge.	$[0.19, 0.81]$
You are beautiful, you are powerful, you are amazing.	$[0.29, 0.71]$
In a city full of great ideas and creativity, I've met a few people who have done things you wouldn't believe.	$[0.26, 0.74]$
The American Dream is alive in this great city. As a new generation of American heroes begins to realize their own American Dream.	$[0.24, 0.76]$
Negative Examples	**Labels**
No one likes it. Nobody wants it. It is a disgrace.	$[0.7, 0.3]$
The company is no longer in business and has ceased operations.	$[0.71, 0.29]$
Please don't use this feature to communicate with customers	$[0.74, 0.26]$
Do not buy from this seller.	$[0.79, 0.21]$

Experiment: Visualization of Generated Data

We performed T-SNE visualization using TAMs trained with UniGen

- The TAM is only trained on synthetic data (gray), generated by UniGen framework
- This TAM effectively classifies data from various domains



Experiment: Ablation Study

We performed ablation study to verify the effectiveness of our proposed components:

- The experimental results show that the usage of soft label from pseudo-relabeling, denoising memory bank, and supervised contrastive learning is beneficial
- Additionally, UniGen outperformed TAMs trained with task-specific data gathered from each domain, demonstrating its superiority

DistilBERT	SST-2	IMDB	Rotten	Amazon	Yelp	CR	Tweet	Average
UniGen	77.67	67.81	73.16	75.06	74.81	79.86	81.41	**75.68**
UniGen w/ Hard Relabeling	77.18	67.18	72.37	72.91	72.95	78.14	80.39	74.45
UniGen w/o Relabeling	76.34	66.58	71.78	70.63	70.97	76.59	79.62	73.22
UniGen w/o Denoising MB	77.06	67.13	72.04	74.69	73.66	78.47	80.84	74.84
UniGen w/o SCL	75.53	66.10	69.63	71.43	69.58	77.22	79.31	72.69
Combined Prompts	74.19	63.16	71.08	73.62	72.93	78.05	78.02	73.01

Experiment: Comparison between Various PLMs

We compared the differences of TAMs trained with synthetic data from different PLMs:

- We used Gemma-2b, Qwen2-1.5B, and Phi-1.5
- The experimental results show that GPT2-XL excels in terms of average performance
- However, it should be noted that optimal prompt design may vary for each PLMs
- We plan to explore methods to effectively optimize prompts and hyperparameters

DistilBERT	SST-2	IMDB	Rotten	Amazon	Yelp	CR	Tweet	Average
UNIGEN w/ GPT2-XL	77.67	67.81	73.16	75.06	74.81	79.86	81.41	**75.68**
UNIGEN w/ Gemma-2b	71.50	69.40	67.04	76.48	76.89	77.24	52.03	70.08
UNIGEN w/ Qwen2-1.5B	66.37	63.19	63.76	71.69	72.44	66.06	63.49	66.71
UNIGEN w/ Phi-1.5	74.98	68.35	70.82	73.86	75.11	71.82	84.01	74.13

Experiment: Extensibility of Relabeling Strategy

We examined our pseudo-relabeling approach can be generalized to other methods:

- We applied the pseudo-relabeling approach to ZeroGen
  - Soft relabeling: Original method suggested in our study
  - Hard relabeling: Alternative method that assigns hard label instead of directly leveraging soft label
- The results suggest that pseudo-relabeling can enhance ZeroGen, not just UniGen
- We plan to investigate the broad application of pseudo-relabeling to improve other methods based on synthetic data

DistilBERT	SST-2	IMDB	Rotten	Amazon	Yelp	CR	Tweet	Average
ZeroGen	80.06	69.13	74.73	73.02	72.77	73.59	74.83	74.02
ZeroGen w/ Hard Relabeling	80.72	69.25	73.98	73.41	73.18	73.76	74.91	74.17
ZeroGen w/ Soft Relabeling	81.79	70.40	75.32	73.65	73.31	74.72	75.14	**74.90**

Conclusion

We proposed:

- UniGen, a novel method to improve domain generalizability of methods based on synthetic data

We found that:

- UniGen can achieve domain generalizability using only a single small model, surpassing the performance of PLM used to generate synthetic data
- This enables the usage of single, lightweight model during inference, improving usefulness of synthetic data

We plan to:

- Further improve performance of UniGen on each domain
- Leverage small task-specific samples to optimize TAMs trained with UniGen

Thank You!