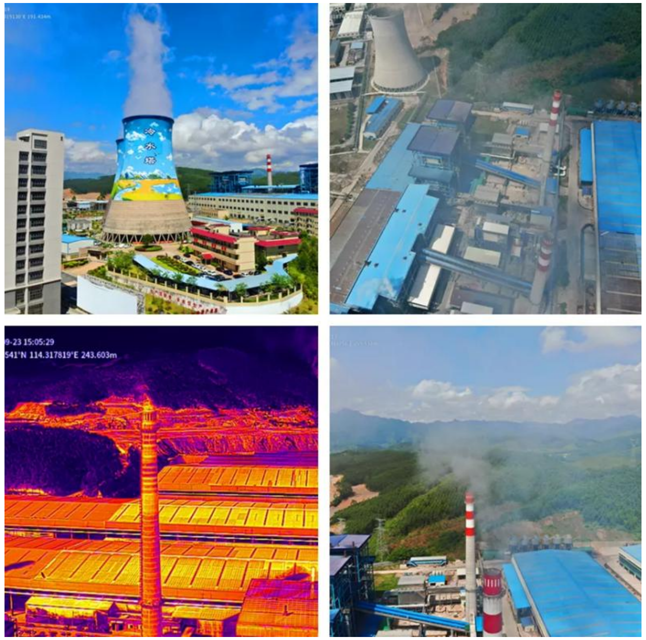
CaliFormer: Leveraging Unlabeled Measurements to Calibrate Sensors with Self-supervised Learning

Haoyang Wang[1], Yuxuan Liu[1], Chenyu Zhao[1], Jiayou HE[2], Wenbo Ding[1], Xinlei Chen[1]

[1]Shenzhen International Graduate School, [2]Hong Kong University of Science and Technology

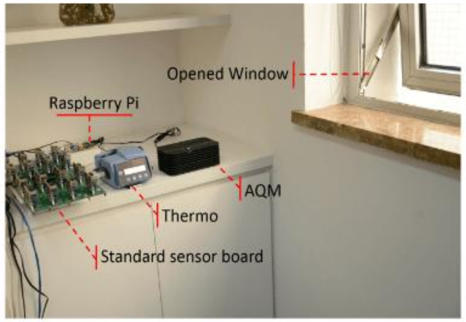
Air Pollution Takes Away a lot of lives

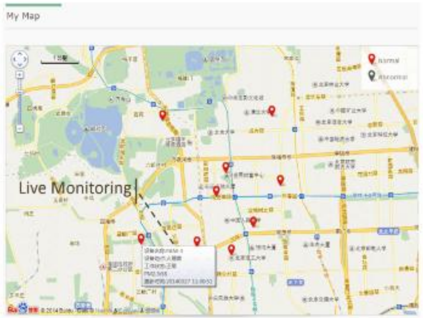


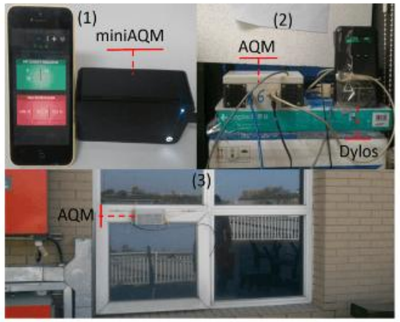
In 2019, air pollution caused 4.2 million deaths worldwide [1]

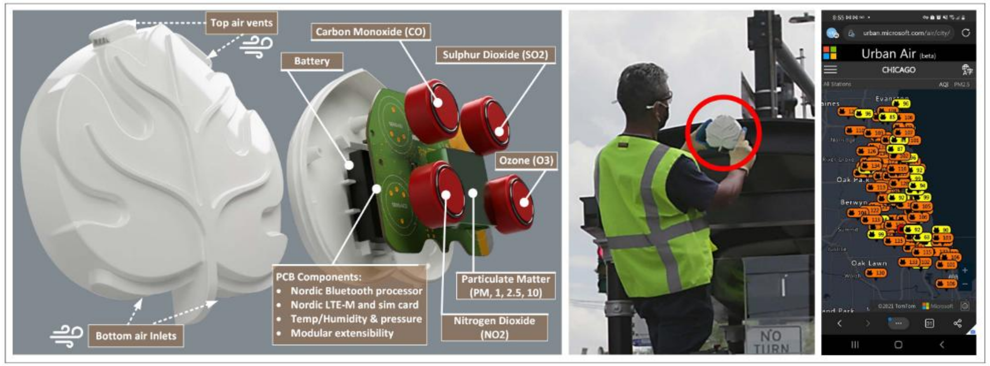
[1] https://www.who.int/zh/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

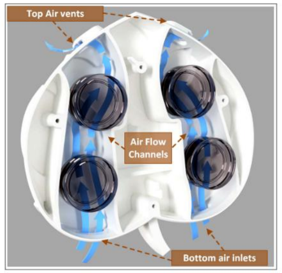
Response: Environment Monitoring









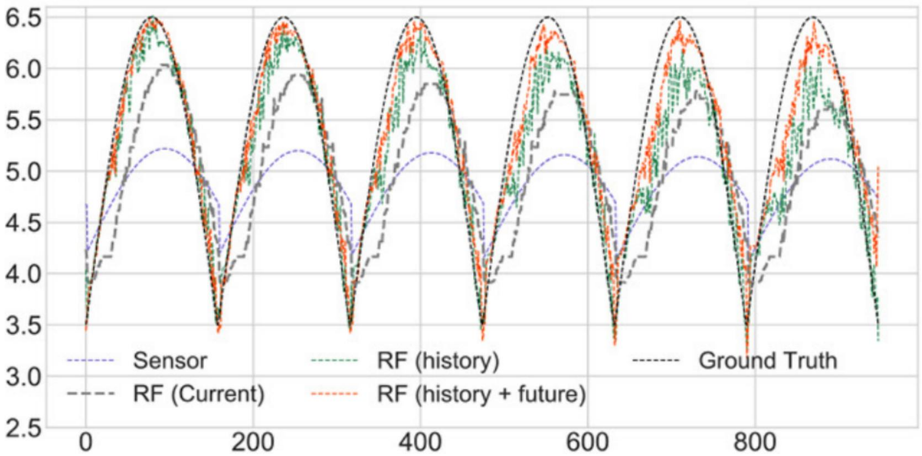


Design specific sensors to monitor the quality of air

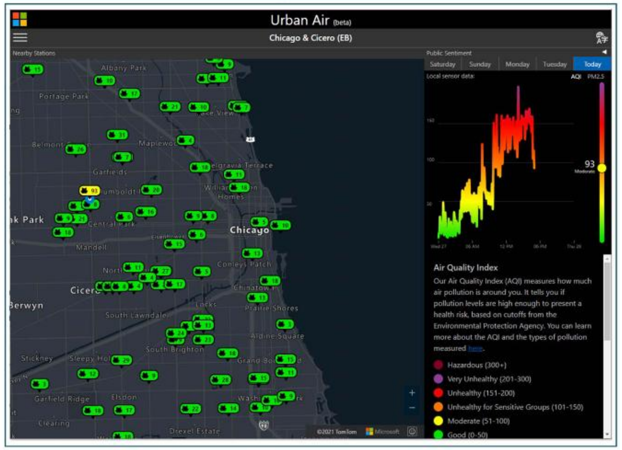
- Cheng Y, Li X, Li Z, et al. AirCloud: A cloud-based air-quality monitoring system for everyone[C]//Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems. 2014: 251-265.
- Daepp M I G, Cabral A, Ranganathan V, et al. Eclipse: an end-to-end platform for low-cost, hyperlocal environmental sensing in cities[C]//2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN). IEEE, 2022: 28-40.

Key issue: the quality of the measurements

Sensor measurements

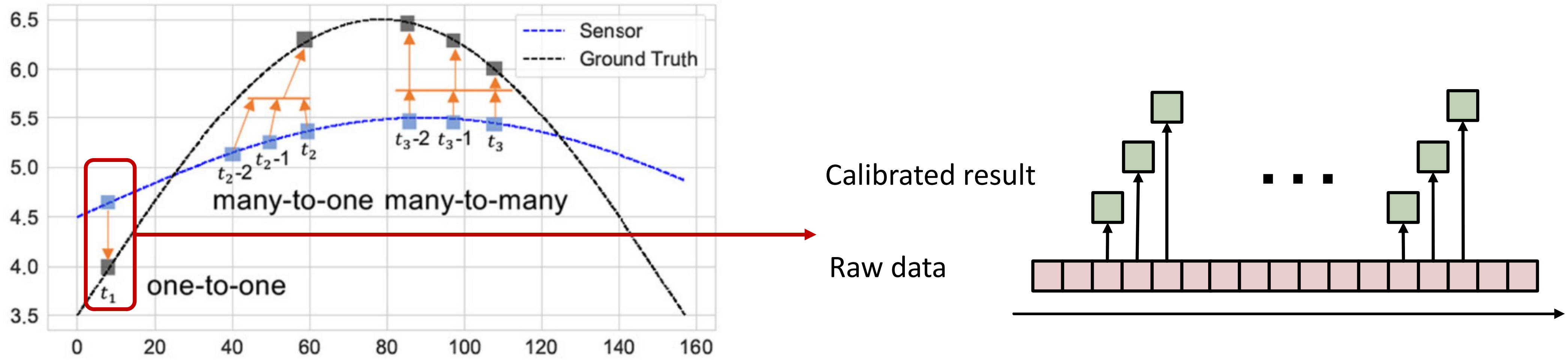


Monitoring system



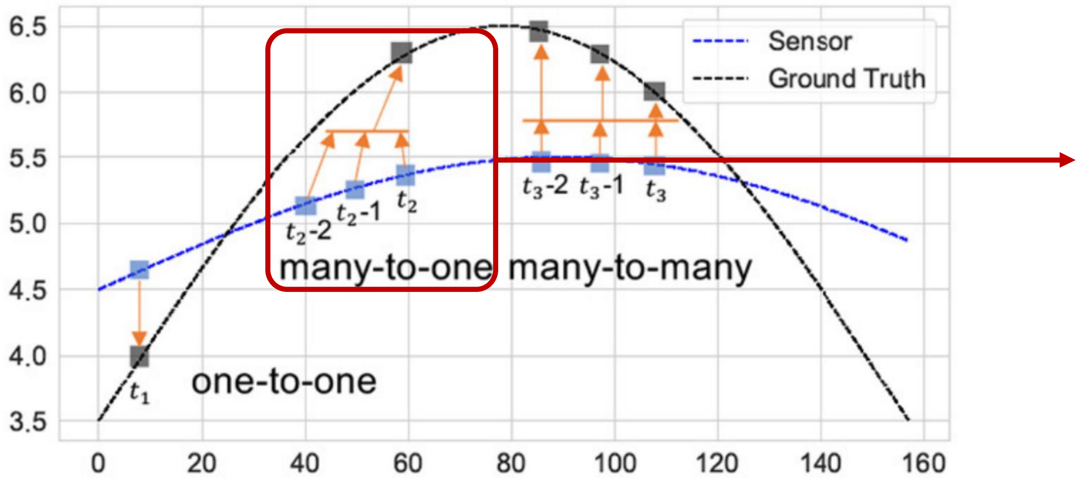
The monitoring systems heavily depends on the quality of measurements

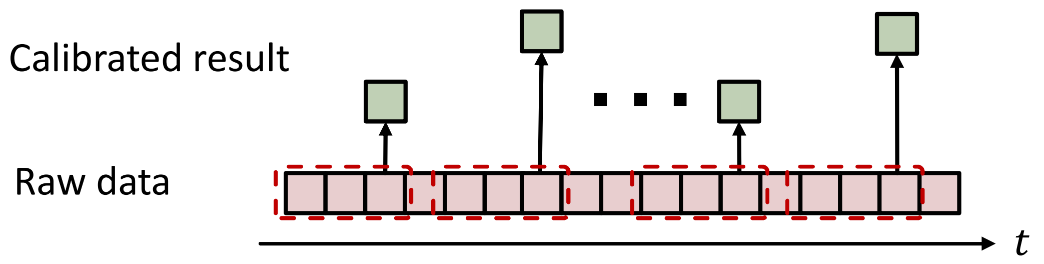
One-to-one calibration



Utilize measurements **at a single time step** to predict a calibrated sensor value **for the same time step**

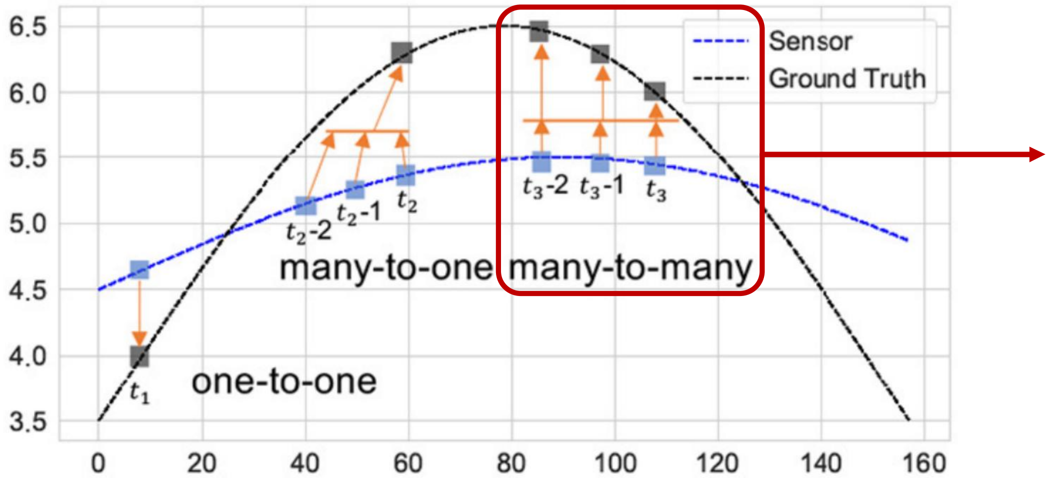
Many-to-one calibration

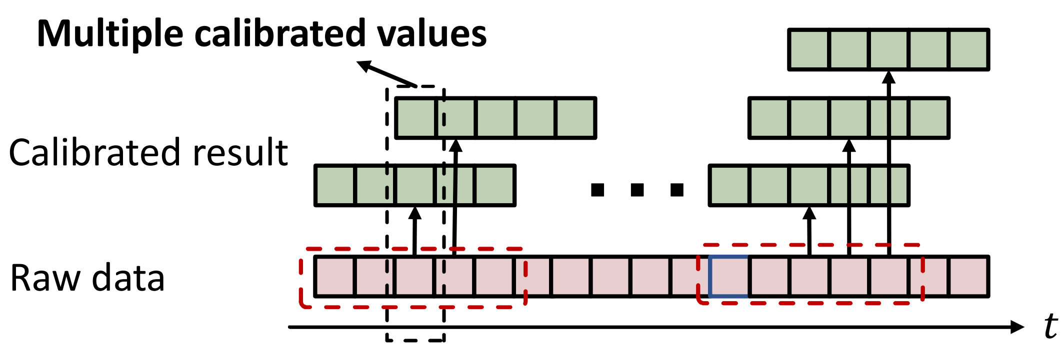




Utilizes measurements in **recent past** to capture **measured phenomena** and **the temporal dynamics**

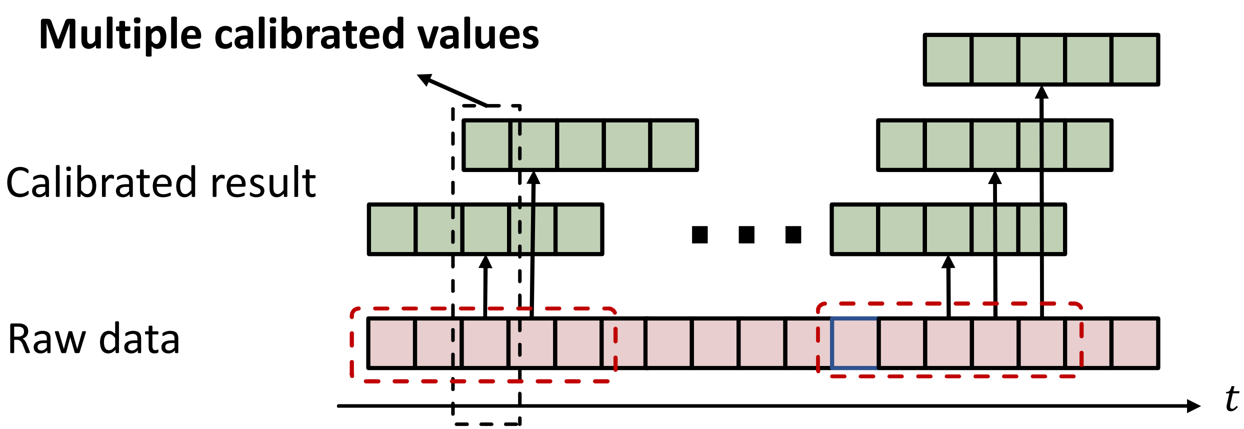
Many-to-Many calibration





Utilizes measurements in **recent past** and **near future** to calibrating low-cost measurements

Many-to-Many calibration



Providing immediate calibration with its **gradual refinement** as further measurements become available

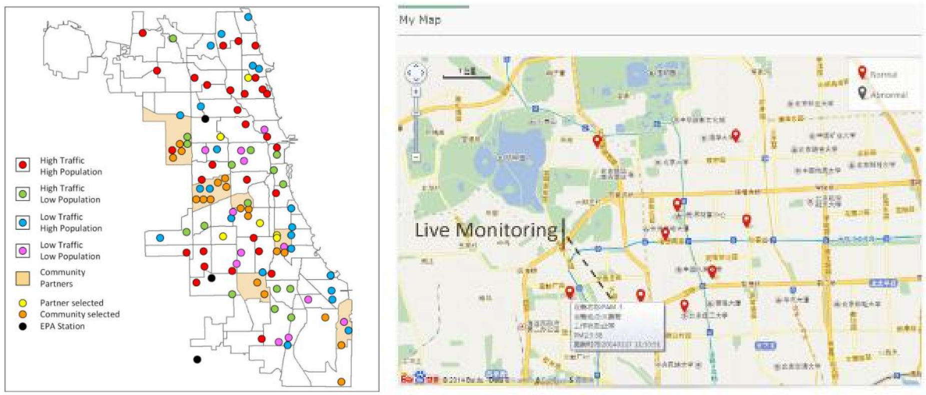
Existing Methods

However, all these methods are data-hungry, and it's hard to collect sufficient data

Reason 1

Sparsely distributed monitoring stations

Reason 2



Various combinations of sensors and usage scenarios

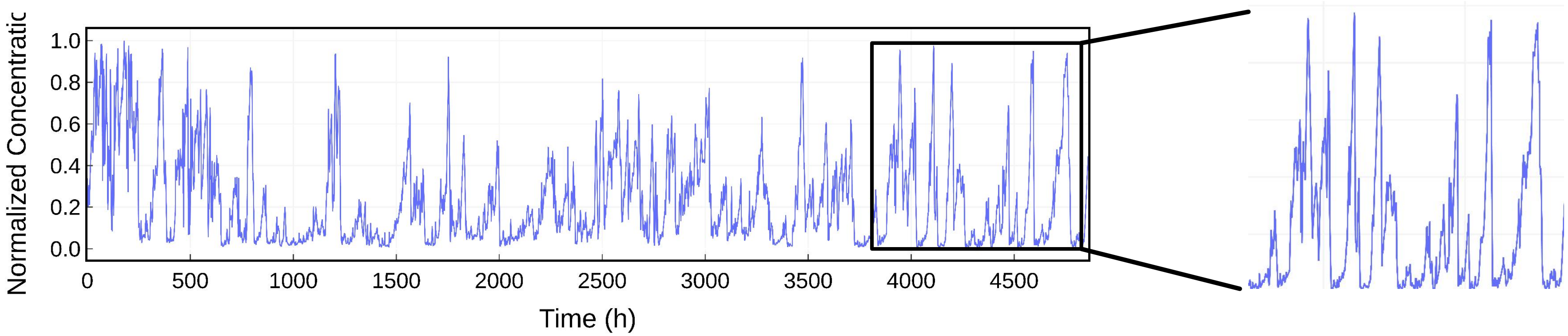
Research problem

How to scale data-driven system to sensing calibration with limited labeled data?

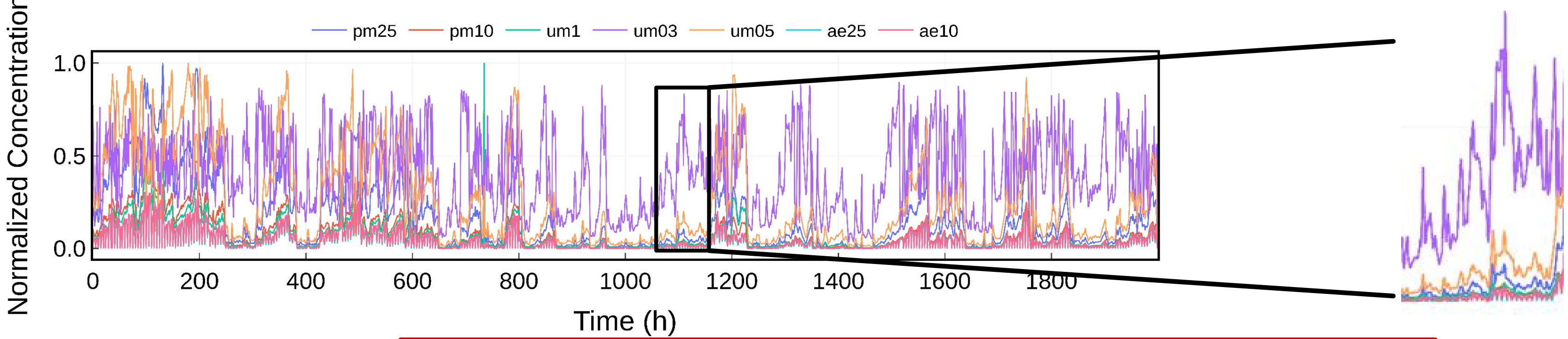
In mathematical form, how to minimize the following loss function with limited labeled data?

$$argmin_{\theta}\left(\sum_{t=1}^{T} \underbrace{\boxed{\mathcal{L}}}_{\text{Loss function}}\left(\underbrace{\boxed{y_{\langle t-\delta_1, t+\delta_2\rangle}}}_{\text{Ground Truth}}, \underbrace{\boxed{\mathcal{C}_{\theta}}}_{\text{Calibration model}}\left(\underbrace{\boxed{x_{\langle t-\delta_1, t+\delta_2\rangle}}}_{\text{Measurement}}\right)\right)\right)$$

Observation: time- and spatial-invariant knowledge



Temporal dependencies of the measurement series



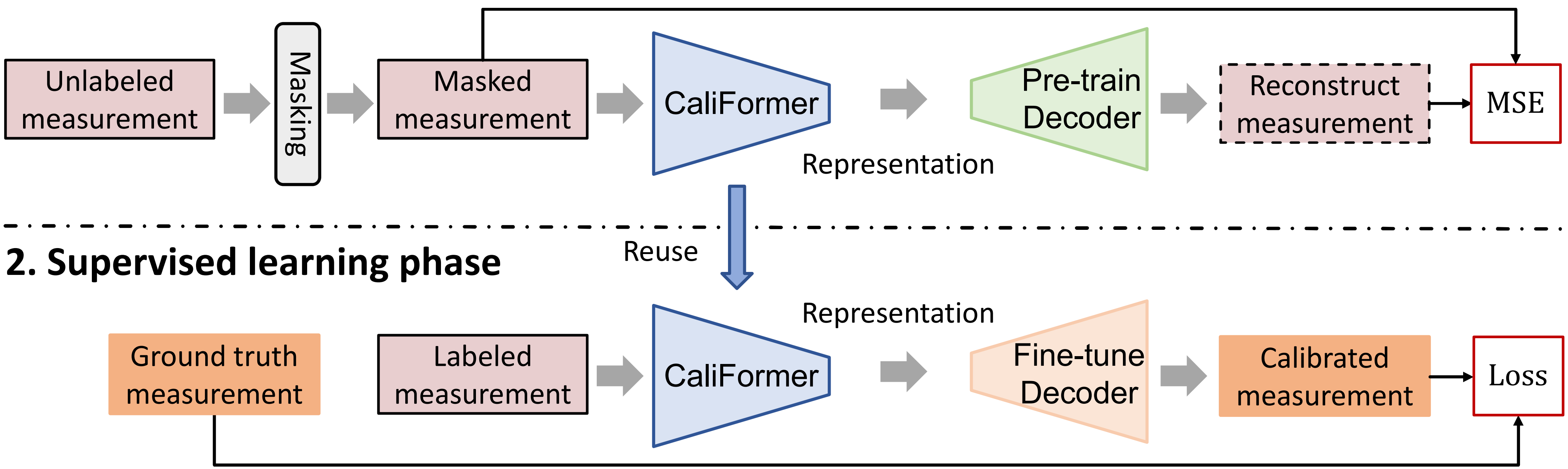
Correlation between multiple pollutants measurements

清華大學深圳国际研究生院
Tsinghua Shenzhen International Graduate School

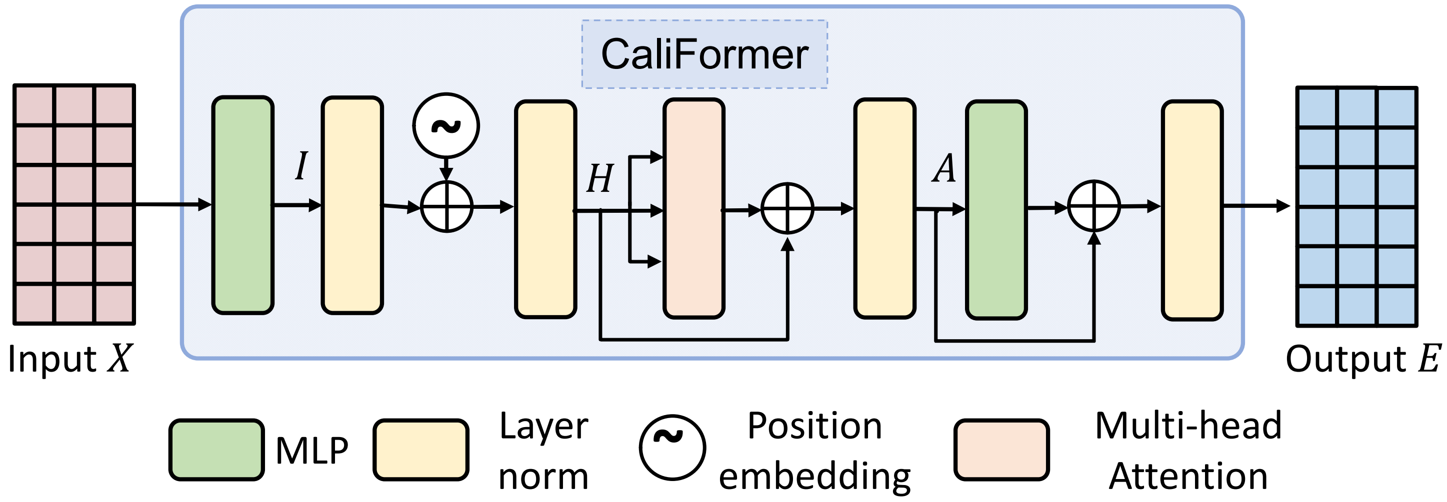
Framework Overview

1. Self-supervised learning phase

2. Supervised learning phase

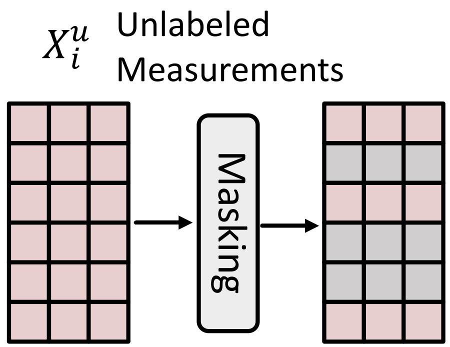


CaliFormer: Representation Learning Model

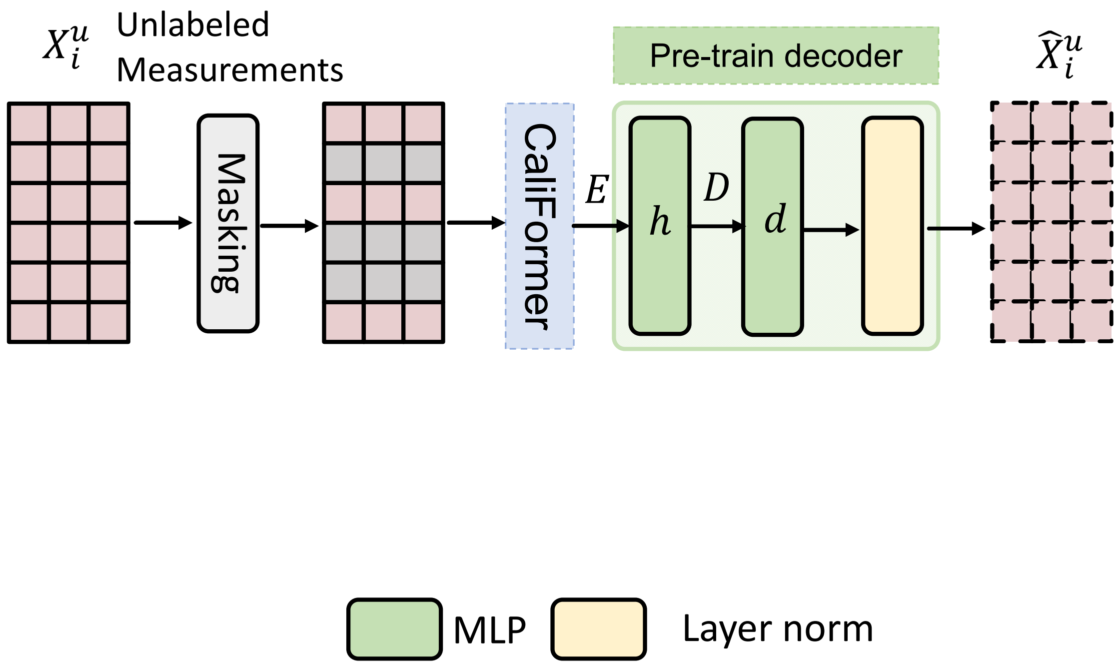


Inspired by the efficiency of the Transformer in sequence-to-sequence prediction tasks

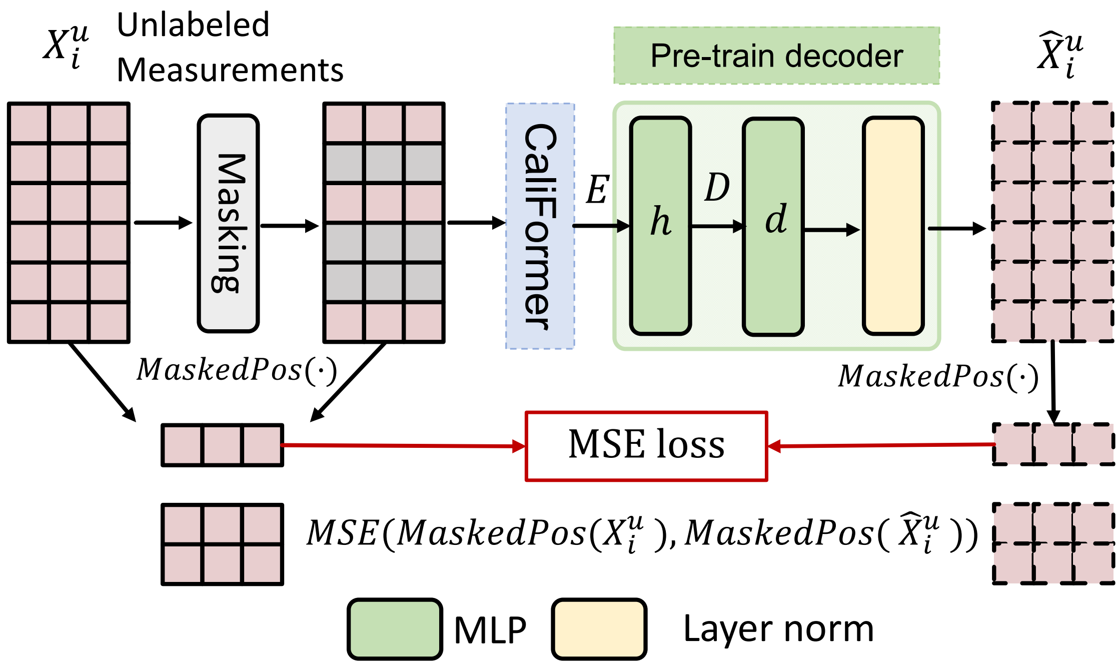
The Self-supervised Learning Phase



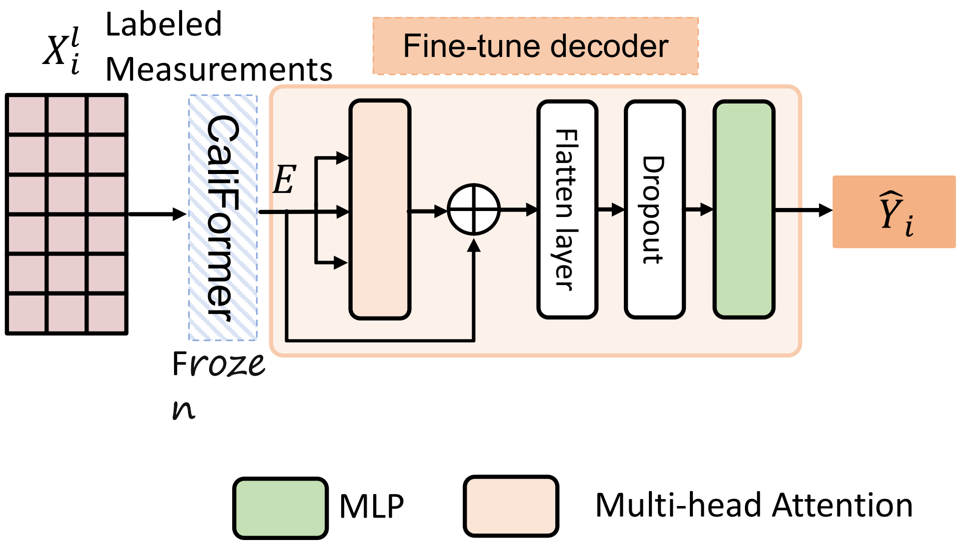
The Self-supervised Learning Phase



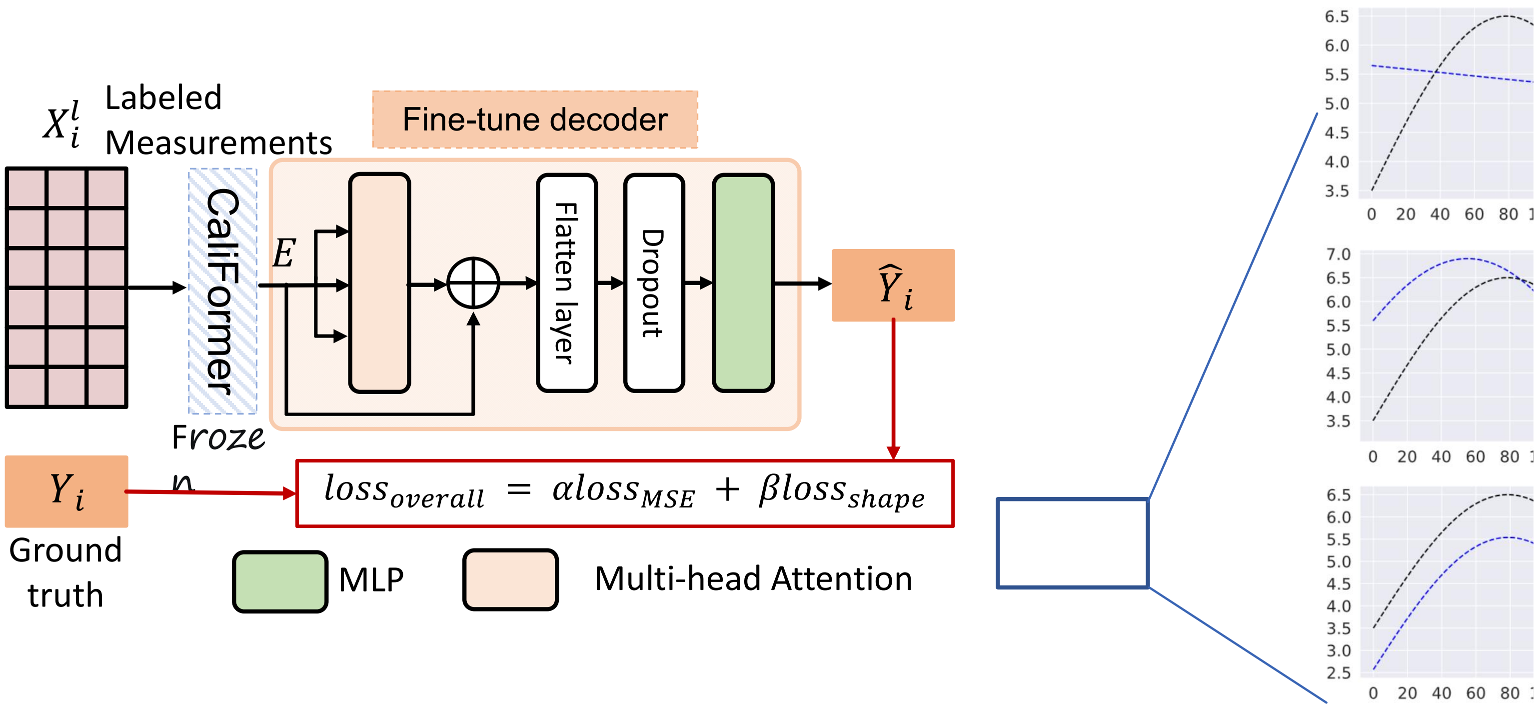
The Self-supervised Learning Phase



The Supervised Learning Phase



The Supervised Learning Phase



α, β: adjustable trade-off coefficient to balance the importance of two parts of the loss function

Experiments Setup

- **Datasets**: Beijing Data Set, which comprise PM2.5 (particles of diameter less than 2.5μm) measurements at seven locations in Beijing. The sensor reports seven feature measurements at a time which are utilized to train the models. There are 60450 samples used in the experiments.
- **Preprocessing**: We split the sensor measurements into the training (60%), validation (20%), and test (20%) sets. The training set is then split into the labeled set (1%) and unlabeled set(99%).
- **Methods in comparison**: (1) Naïve; (2) SensorFormer (SF) : To the best of our knowledge, SF is the state-of-the-art many-to-many calibration method, which is based on the Transformer model; (3) SensorFormer-mo (SF-mo). SF-mo is the many-to-one version of state-of-the-art calibration method, which means this method do not use lossshape in the loss function. (4) SensorFormer-oo (SF-oo) [7]. SF-oo is the one-to-one version of state-of-the-art calibration method. (5) CaliFormer-FT. CaliFormer-FT stands for CaliFormer with the fine-tune decoder.
- **Metric**: Mean Absolute Error (MAE)

Overall Performance

TABLE I
OVERALL PERFORMANCE WITH 1% LABELED DATA, WHICH IS SHOWN IN MAE($\mu G/m^3$).

Methods	Naive	SF-oo	SF-mo	SF	CaliFormer-FT
MAE	31.25	25.20	24.91	24.08	**18.20**

TABLE II
PERFORMANCE WITH DIFFERENT LABELING RATES, WHICH IS SHOWN IN MAE($\mu G/m^3$).

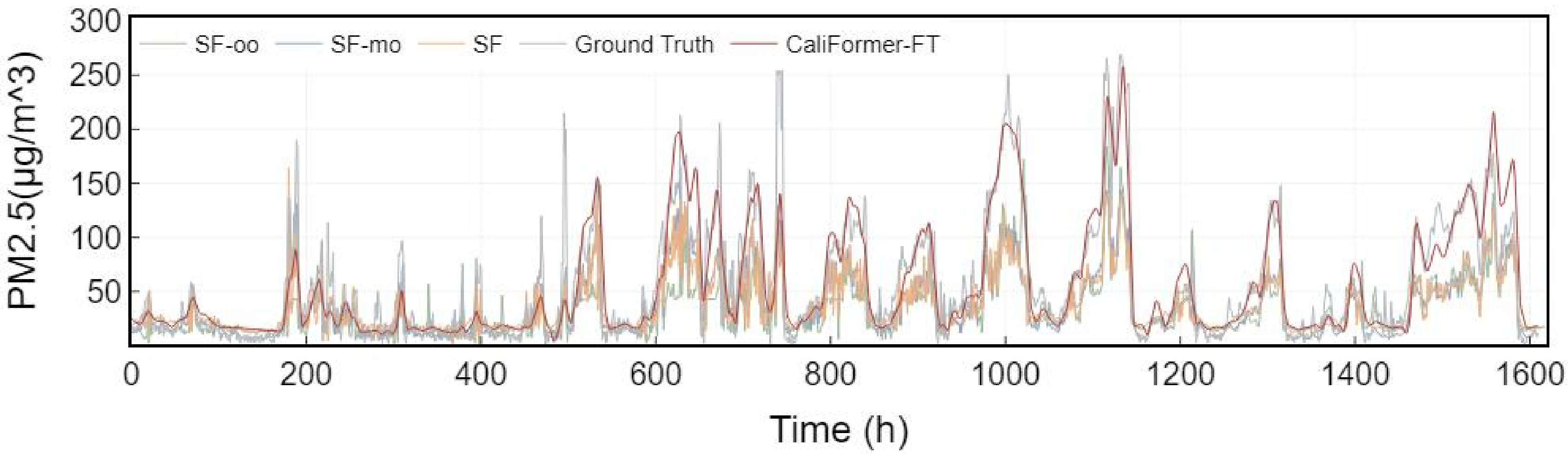
Labeling rate	0.5%	1%	2%	5%	10%	Average
Naive	31.25	31.25	31.25	31.25	31.25	31.25
SF-oo	29.89	25.20	24.68	21.93	21.78	24.70
SF-mo	29.72	24.91	22.33	21.77	20.86	23.92
SF	29.74	24.08	22.11	21.70	20.37	23.60
CaliFormer-FT	**19.91**	**18.20**	**15.20**	**14.84**	**14.57**	**16.54**

Overall performances: with only 1% of the labeled dataset. The CaliFormer-FT outperforms the state-of-the-art method SF by 25%. This is because the CaliFormer extracts *effective representation* from unlabeled data.

Varying labeling rate: the performance with different labeling rates, varying from 0.5% to 10%. The fine-tune decoder achieve better performance with learned CaliFormer. *The gain is significant with low labeling rate.*

清華大学深圳国际研究生院
Tsinghua Shenzhen International Graduate School

Calibration results



Compared to other baselines, CaliFormer-FT performs better, especially during the peak periods

Observations and Contributions

- To the best of our knowledge, CaliFormer is the **first attempt** to incorporate **self-supervised learning** into sensor calibration. Compared to prior research, the CaliFormer **necessitates significantly less labeled data**, which constitutes a tangible advancement toward **practical in-field sensor calibration**.
- Drawing inspiration from the **Transformer** architecture, we develop the CaliFormer to process **multi-modal sensor data**. Additionally, we propose **a set of enhancements** in pre-training methodology and model architecture to facilitate **the effective training** of the calibration model.
- A prototype system is developed and experimentally compared with **state-of-the-art methods**. Extensive evaluation results demonstrate the **effectiveness** of the CaliFormer based calibration system.

Thank you for listening

Haoyang Wang, Yuhan Cheng, Baining Zhao

Lab 2C

The Self-supervised Learning Phase