Transformer Networks
Intuition, Theory, and Real-world Deployment

Daniel Romero

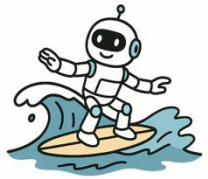
Dept. of Information and Communication Technology

University of Agder, Norway

Acknowledgements:

Pham Q. Viet

Research Council of Norway (IKTPLUSS grant 311994)



The AI wave…

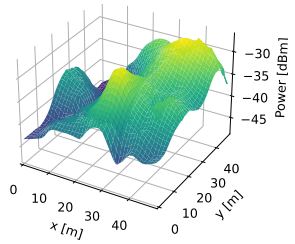
- Social focus on AI mainly due to advances in large language models (LLMs)

Gemini Grok Claude deepseek LLaMA

- Goal of this talk:
  - ➢ Understand the **intuition** and **theoretical principles** in Attention
  - ➢ Learn to apply these ideas to **inference problems** beyond LLMs

Outline of this talk

Attention in inference problems



Attention in LLMs

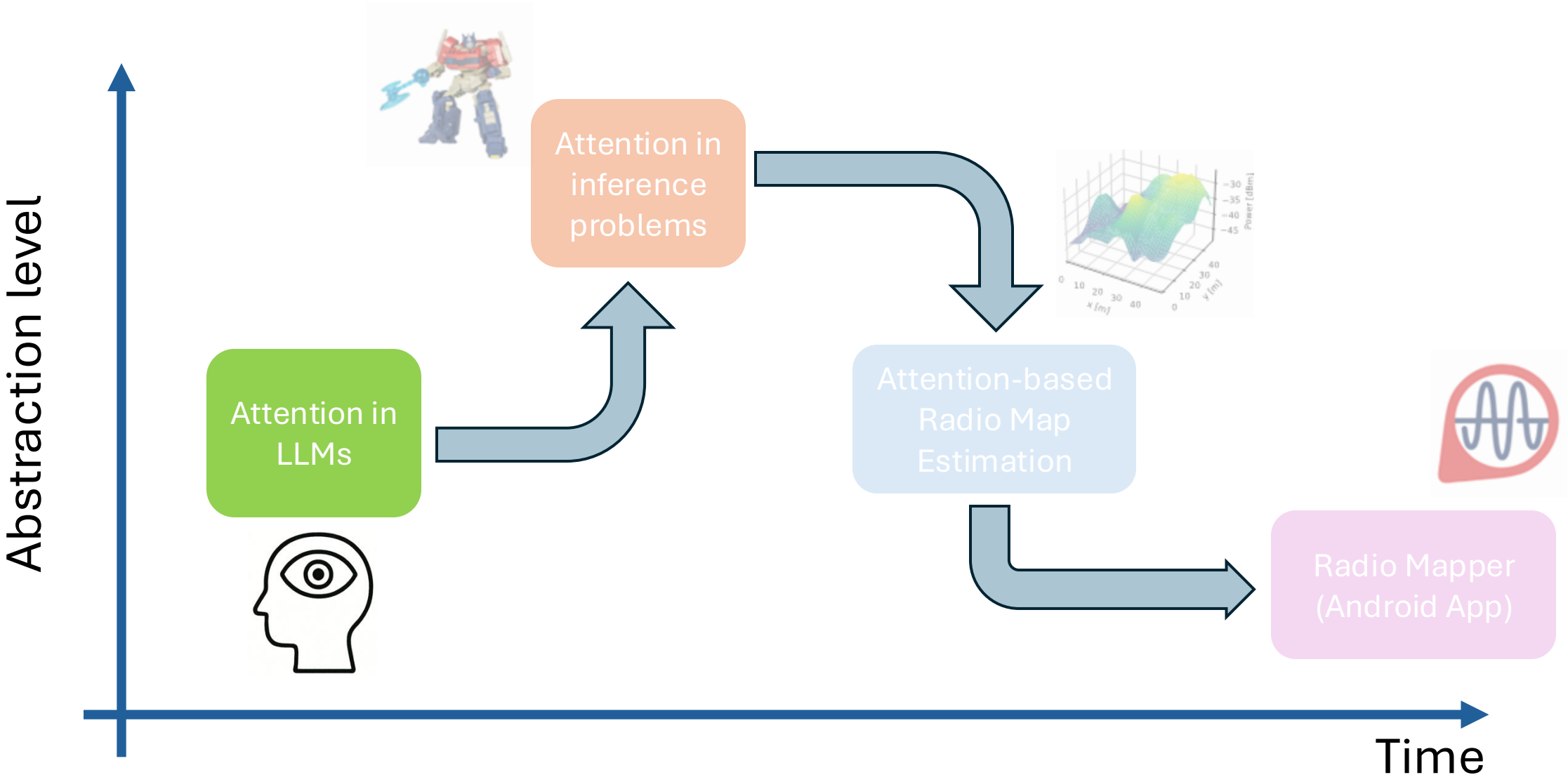
Attention-based Radio Map Estimation

Radio Mapper (Android App)

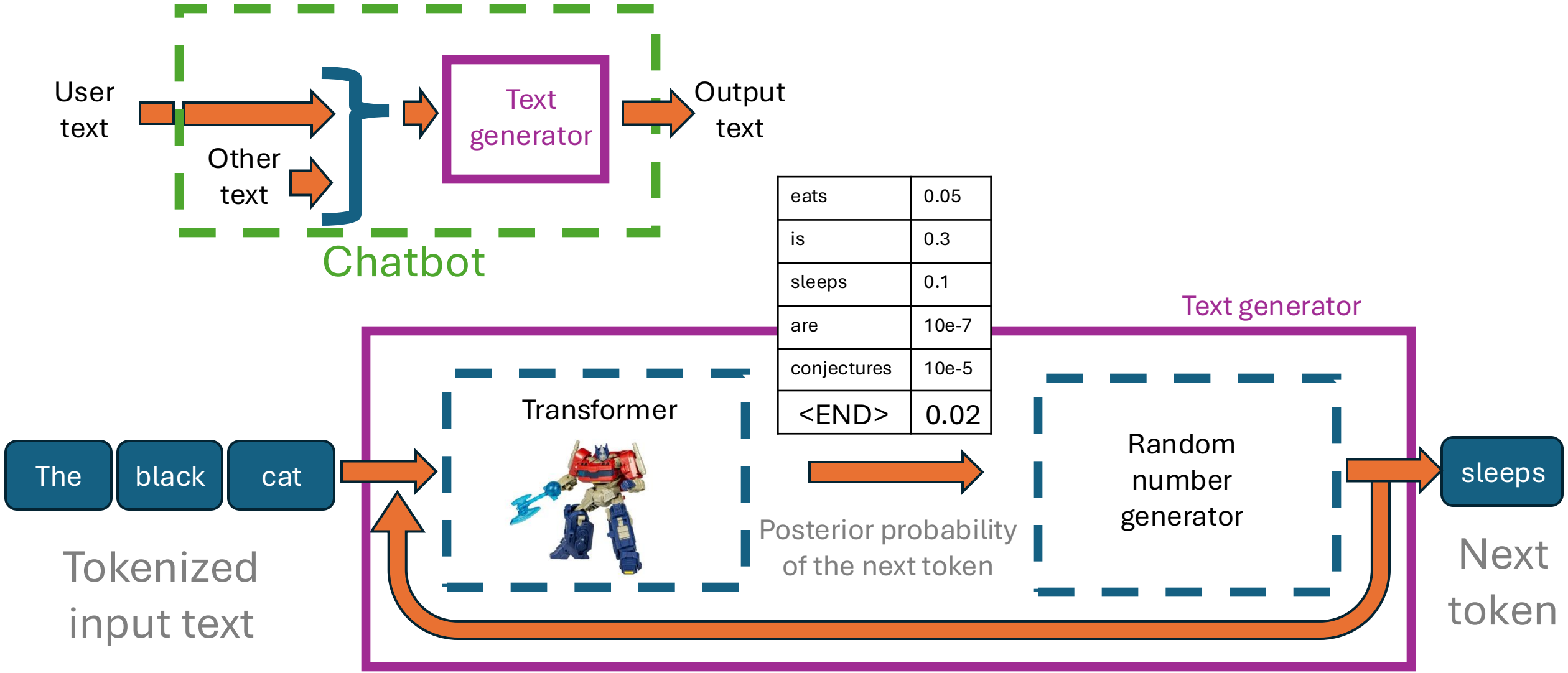
Abstraction level

Time

Outline of this talk



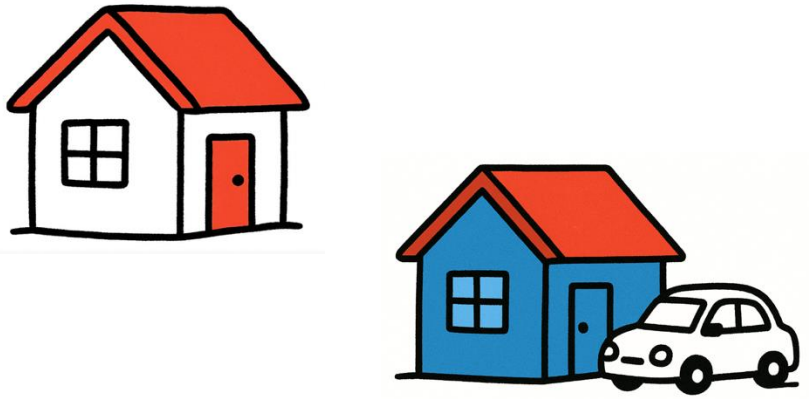
Autoregressive Text Generation



- **Append** the generated token to the input and repeat until <END> is generated

Attention in NLP

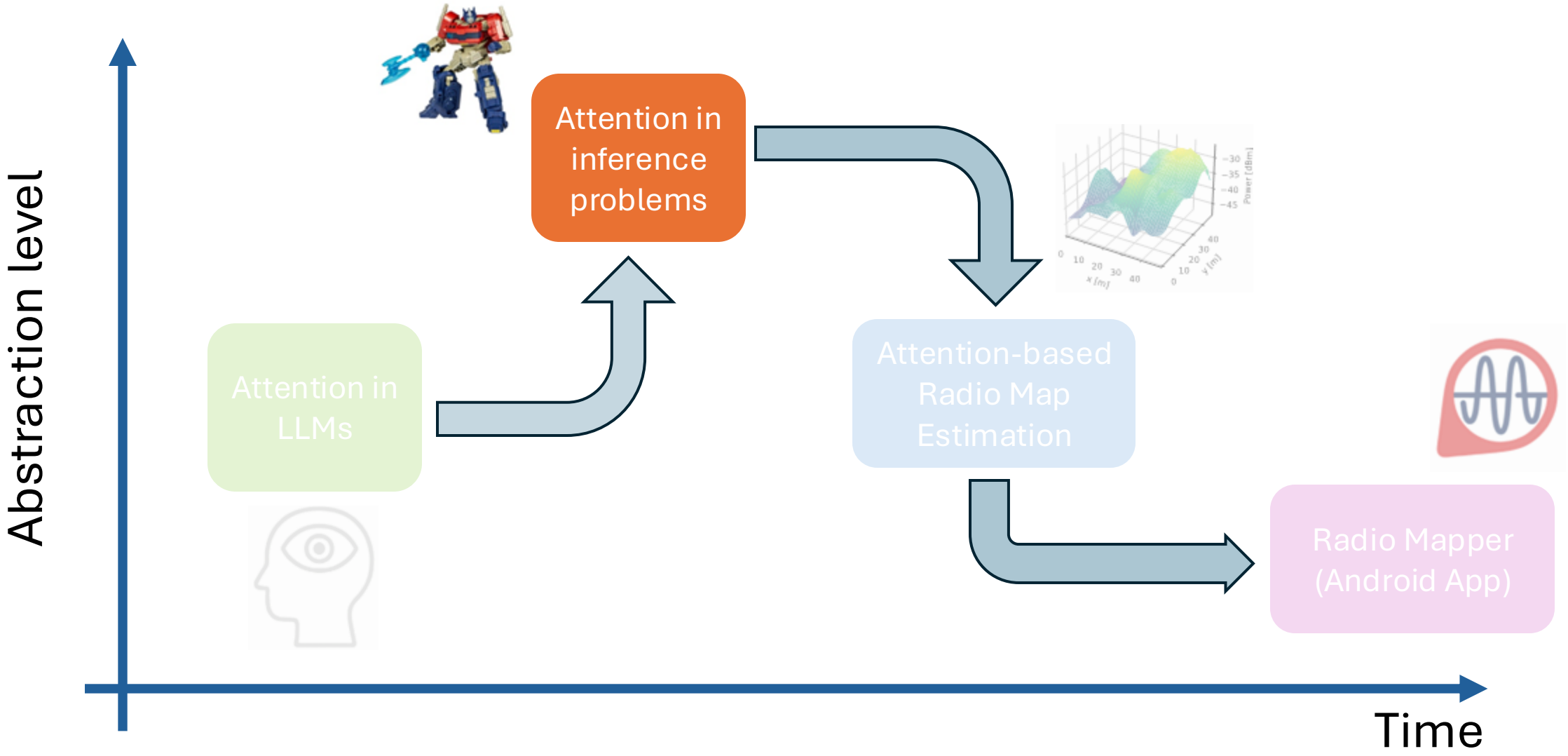
- Combining words results in richer concepts
  - Word relations based on **syntax**
    - The house is white.
    - The car near the house is white.



  - Word relations based on **semantics**
    - The cat ate the food because it was hungry.
    - The cat ate the food because it was warm.

- Each word
  - "pays more attention" to some words than to others.
  - "enriches the meaning" of related words.

Outline of this talk



Attention in Inference Problems

- **Autoregressive text generation**

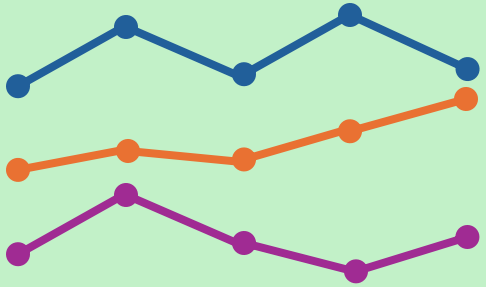
Given $[u[1], \ldots, u[T]]$, obtain the posterior prob. $\boldsymbol{p} \in [0,1]^{N_\mathrm{U}}$ of the next token

$\in \mathcal{U} := \{u^{(1)}, \ldots, u^{(N_\mathrm{U})}\}$

- **Time series forecasting**

Given $\boldsymbol{s}[\tau - T], \ldots, \boldsymbol{s}[\tau - 1]$, predict $\boldsymbol{s}[\tau]$

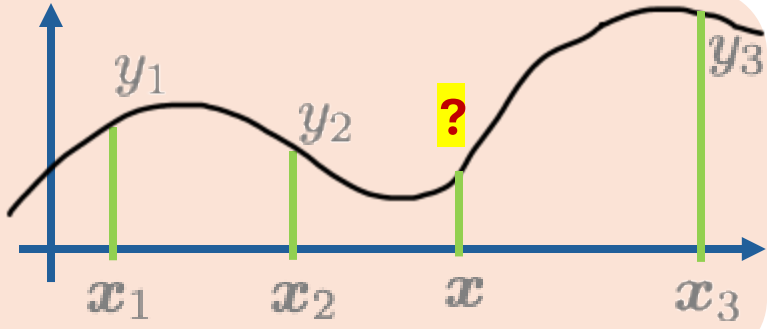
$\in \mathbb{R}^{D_{\boldsymbol{x}}}$



- **Function estimation** $f : \mathbb{R}^{D_{\boldsymbol{x}}} \to \mathbb{R}$

Given $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ and $\boldsymbol{x}$, estimate $f(\boldsymbol{x})$

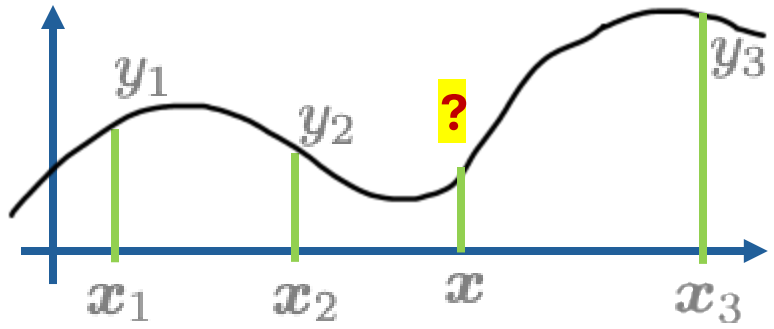
$\boldsymbol{x}_n \in \mathbb{R}^{D_{\boldsymbol{x}}}$ $\quad y_n := f(\boldsymbol{x}_n) + \mathrm{noise}_n$



Note that the number of inputs (N, T) is arbitrary!

Weighted Averages for Function Estimation

Given $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ and $\boldsymbol{x}$, estimate $y := f(\boldsymbol{x})$



- **Estimator** (general form):

$$\hat{y} = \frac{\sum_{n=1}^N \alpha_n y_n}{\sum_{n=1}^N \alpha_n}$$

- Prominent special cases

➢ K-NN	$\alpha_n = \begin{cases} 1 & \text{if } \boldsymbol{x}_n \text{ a NN of } \boldsymbol{x}, \\ 0 & \text{otherwise.} \end{cases}$
➢ Radius-NN	$\alpha_n = \begin{cases} 1 & \text{if } \|\boldsymbol{x} - \boldsymbol{x}_n\|_2 \le R, \\ 0 & \text{otherwise.} \end{cases} := \alpha(\boldsymbol{x}_n, \boldsymbol{x})$
➢ Shepard's method	$\alpha_n = \frac{1}{\|\boldsymbol{x} - \boldsymbol{x}_n\|_2^\gamma} := \alpha(\boldsymbol{x}_n, \boldsymbol{x})$
➢ Nadaraya-Watson estimator	$\alpha_n = \alpha(\boldsymbol{x}_n, \boldsymbol{x})$ a kernel, e.g. Gaussian

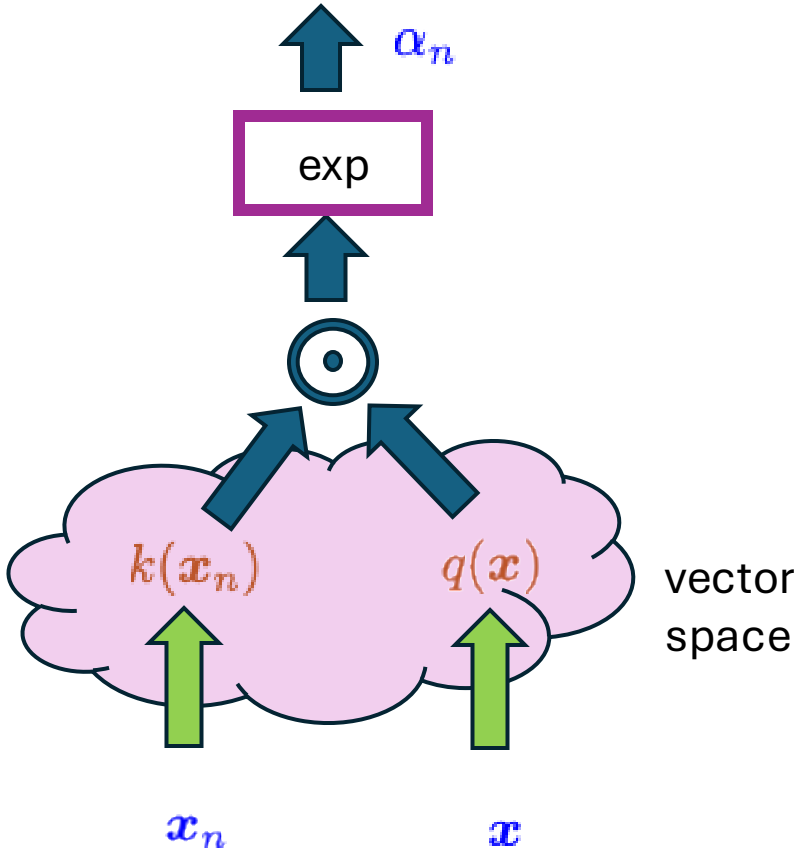
α_n → Quantify how relevant the n-th observation is → **Attention weights**

Inner-Product Attention

- Given by $\alpha_n = \alpha(\boldsymbol{x}_n, \boldsymbol{x}) = \exp\left(k^\top(\boldsymbol{x}_n) \cdot q(\boldsymbol{x})\right)$ quantifies similarity

Learnable

- $k : \mathbb{R}^{D_x} \to \mathbb{R}^{D_k}$ produces *"keys"* $\{k(\boldsymbol{x}_n)\}_n$
  - $k(\boldsymbol{x}_n)$: information that $\boldsymbol{x}_n$ *"offers"*
- $q : \mathbb{R}^{D_x} \to \mathbb{R}^{D_k}$ produces the *"query"* $q(\boldsymbol{x})$
  - $q(\boldsymbol{x})$: information that $\boldsymbol{x}$ is *"interested in"*



- As a soft-max: $\hat{y} = \dfrac{\sum_{n=1}^N \alpha_n y_n}{\sum_{n=1}^N \alpha_n} = \sum_{n=1}^N \bar{\alpha}_n y_n$

$$\bar{\alpha}(\boldsymbol{x}_n, \boldsymbol{x}) = \frac{\exp\left(k^\top(\boldsymbol{x}_n) q(\boldsymbol{x})\right)}{\sum_{n'=1}^N \exp\left(k^\top(\boldsymbol{x}_{n'}) q(\boldsymbol{x})\right)} \approx \delta[n - n_{\max}]$$

if the $n_{\max}$ -th term dominates

Vaswani et al. "Attention is all you need." *Advances in neural information processing systems* (2017).

The Attention Operator

- general
- allows composition

- **DEF:**

Learnable $v : \mathbb{R}^{D_z} \to \mathbb{R}^{D_v}$ produces "values"

$v(z_n)$ information "carried" by z_n

Reference vectors

$\mathcal{Z} := \{z_1, \ldots, z_N\} \subset \mathbb{R}^{D_z}$

Input vector

x

$$A(\mathcal{Z}, x) = \frac{\sum_{n=1}^{N} \alpha(z_n, x) v(z_n)}{\sum_{n=1}^{N} \alpha(z_n, x)} \in \mathbb{R}^{D_v}$$

$$\alpha(x_n, x) = \exp\left(\frac{1}{\sqrt{D_k}} k^\top(z_n) q(x)\right)$$

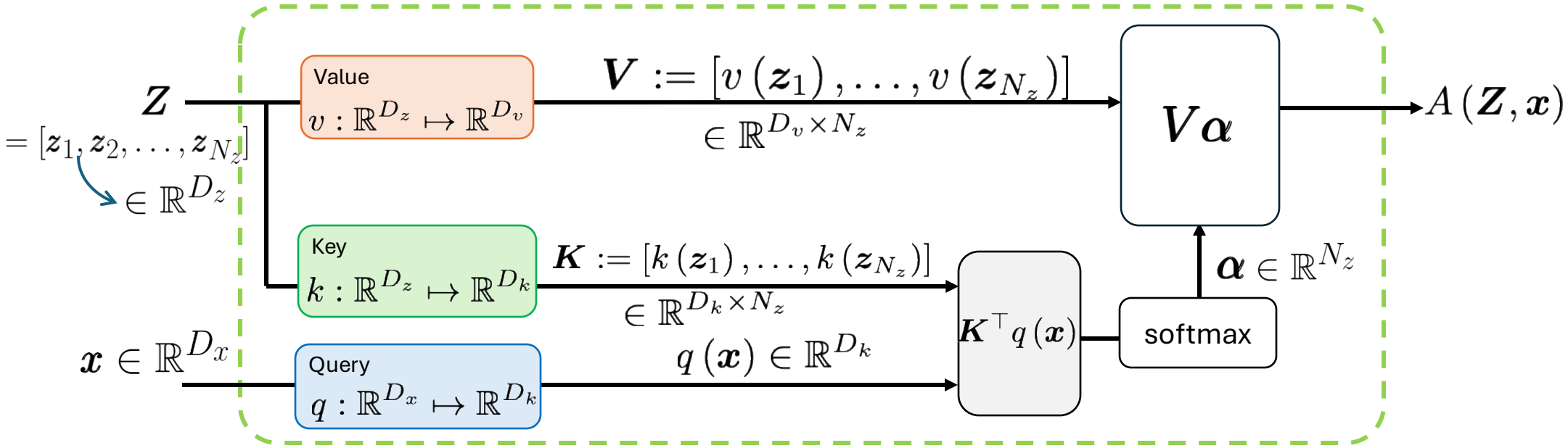
Intuition: This operator enriches the information in x with the information in the reference vectors

- These functions are typically **linear**: $k(z_n) = W_k z_n, \quad q(x) = W_q x, \quad v(z_n) = W_v z_n$

Trainable parameters: $\Theta := (W_k, W_q, W_v)$

The Attention Operator

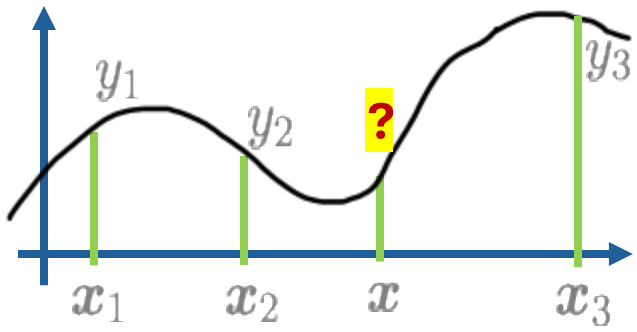
$$A(\boldsymbol{Z}, \boldsymbol{x}) = \boldsymbol{V}\mathrm{softmax}\left(\boldsymbol{K}^\top q(\boldsymbol{x})\right)$$



Attention Operator in Function Estimation

- The attention operator generalizes the function estimator seen earlier:

$$A(\mathcal{Z}, \boldsymbol{x}) = \frac{\sum_{n=1}^{N} \alpha(\boldsymbol{z}_n, \boldsymbol{x}) v(\boldsymbol{z}_n)}{\sum_{n=1}^{N} \alpha(\boldsymbol{z}_n, \boldsymbol{x})} \quad = \quad \hat{y} = \frac{\sum_{n=1}^{N} \alpha_n y_n}{\sum_{n=1}^{N} \alpha_n}$$



$$\boldsymbol{z}_n = \begin{bmatrix} y_n \\ \boldsymbol{x}_n \end{bmatrix} \quad k(\boldsymbol{z}_n) = [\boldsymbol{0}, \bar{\boldsymbol{W}}_k]\boldsymbol{z}_n \quad v(\boldsymbol{z}_n) = [1, \boldsymbol{0}^\top]\boldsymbol{z}_n = y_n$$

- In practice, if the problem is invariant to translations

→ $\boldsymbol{z}_n = \begin{bmatrix} y_n \\ \boldsymbol{x}_n - \boldsymbol{x} \end{bmatrix} \qquad \hat{y} = A(\mathcal{Z}, \boldsymbol{z}_N)$

(the estimator above is not invariant)

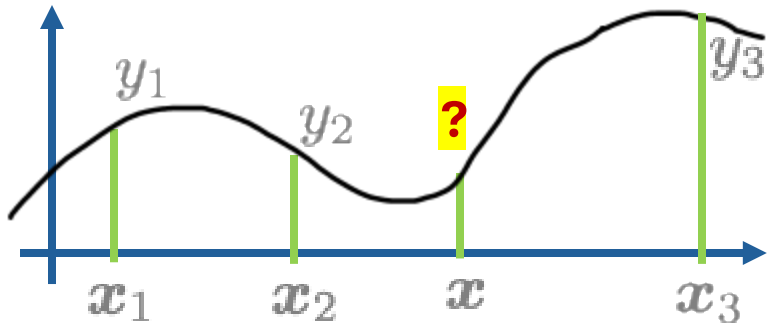
- If the problem is invariant to rotations → see later

Training (Toy) Example

- Consider a **dataset** for the function estimation problem

$$\{(\boldsymbol{x}_n^{(j)}, y_n^{(j)}) : \quad n = 0, \ldots, N, j = 1, \ldots, J\}$$

number of realizations



- Simple **MSE loss** (1 target/realization)

$$\mathcal{L}(\Theta) = \frac{1}{J} \sum_{j=1}^{J} \left(y_0^{(j)} - A(\mathcal{Z}^{(j)}, \boldsymbol{z}_0^{(j)}) \right)^2$$

$$\boldsymbol{z}_n^{(j)} := \begin{bmatrix} y_n^{(j)} \\ \boldsymbol{x}_n^{(j)} \end{bmatrix}$$

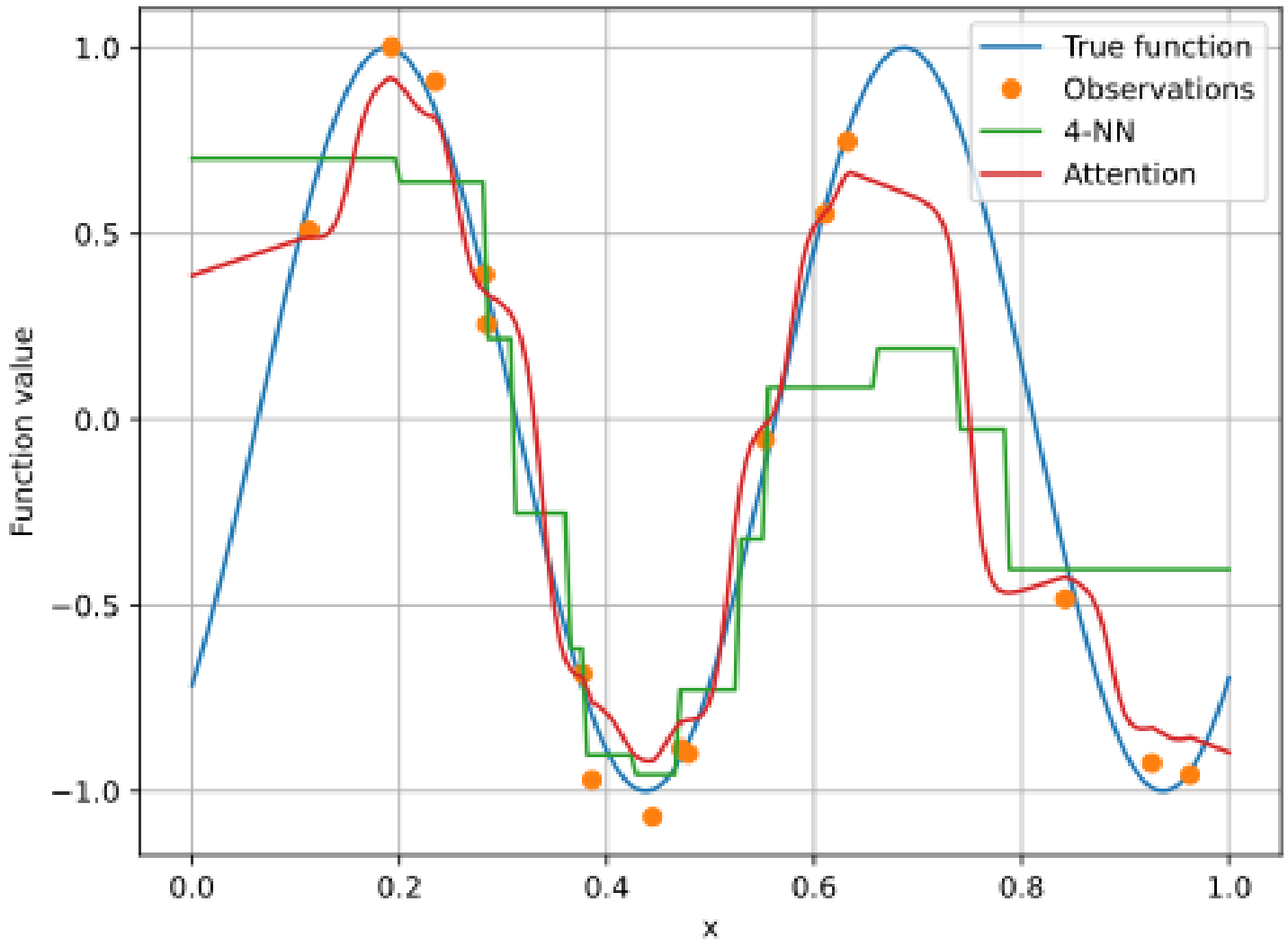
$$\mathcal{Z}^{(j)} := \left\{ \boldsymbol{z}_1^{(j)}, \ldots, \boldsymbol{z}_N^{(j)} \right\}$$

- Multiple targets per realization also possible

- **Training** means finding a local optimum of

$$\underset{\Theta}{\text{minimize}} \quad \mathcal{L}(\Theta)$$

Training Example



Attention Heads

- **Cross-attention** head

$$\mathcal{Z} := \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\} \subset \mathbb{R}^{D_z}$$

order not important

$$\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M]$$

order matters

$$\longmapsto \quad A(\mathcal{Z}, \boldsymbol{X}) := [A(\mathcal{Z}, \boldsymbol{x}_1), \ldots, A(\mathcal{Z}, \boldsymbol{x}_M)] \in \mathbb{R}^{D_v \times M}$$

- **Self-attention** head

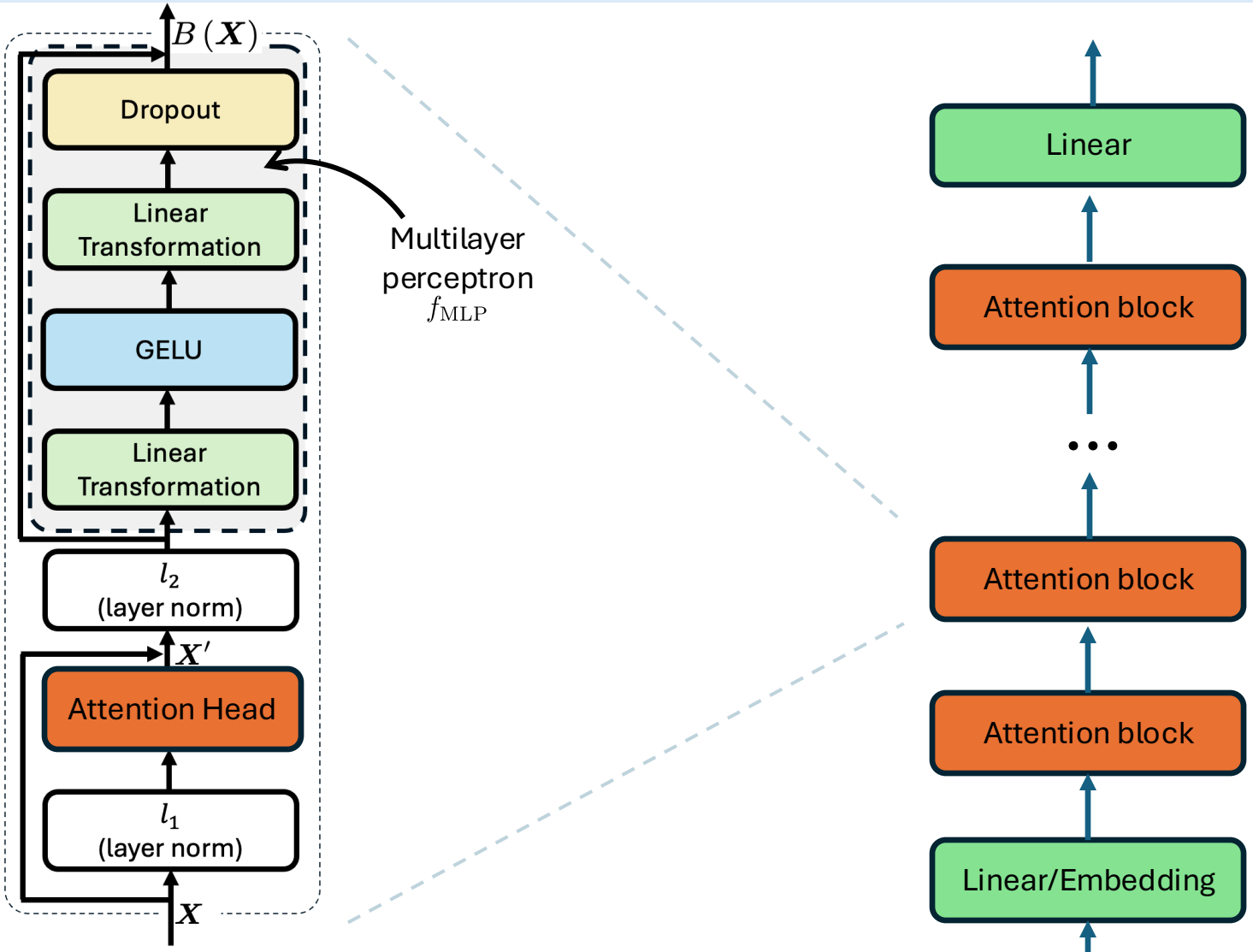
$$\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M] \quad \longmapsto$$

$$A(\boldsymbol{X}) := A([\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M]) := A(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}, [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M])$$
$$= \left[A(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}, \boldsymbol{x}_1), A(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}, \boldsymbol{x}_2), \ldots, A(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}, \boldsymbol{x}_M)\right]$$

- **Causal** self-attention head

$$\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M] \quad \longmapsto \quad A_c(\boldsymbol{X}) := \left[A(\{\boldsymbol{x}_1\}, \boldsymbol{x}_1), A(\{\boldsymbol{x}_1, \boldsymbol{x}_2\}, \boldsymbol{x}_2), \ldots, A(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}, \boldsymbol{x}_M)\right]$$

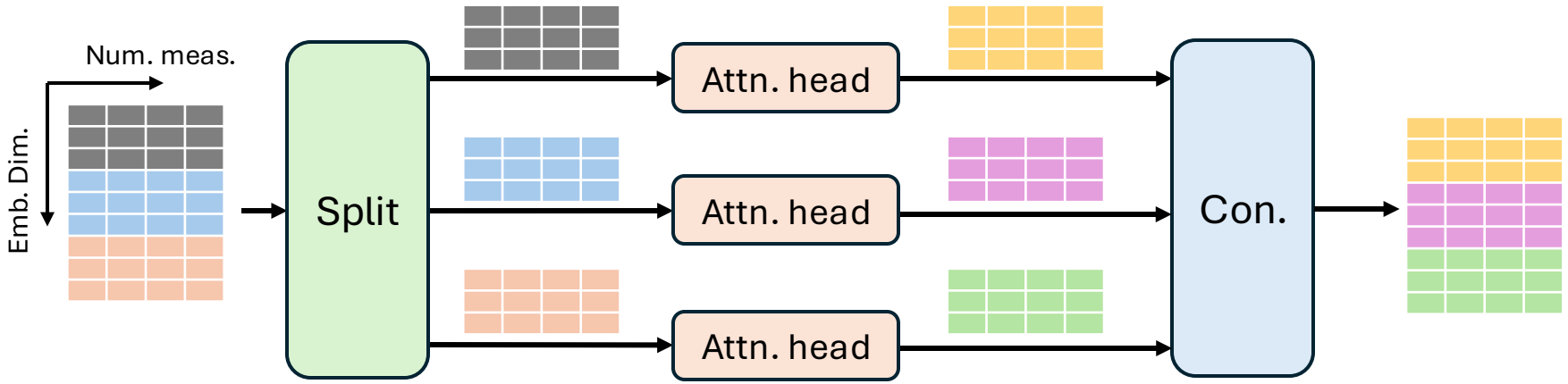
Transformers



Attention block

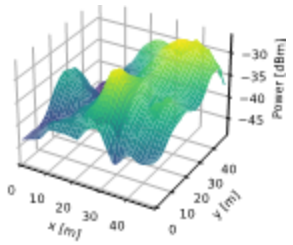
Transformer

Multi-head Attention



Outline of this talk

Attention in inference problems



Attention in LLMs

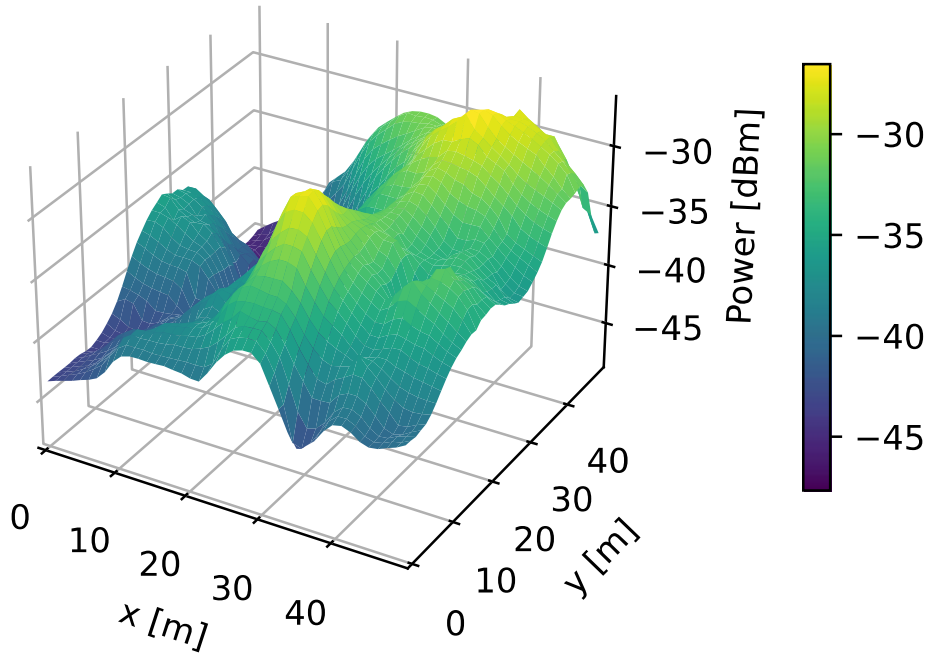
Attention-based Radio Map Estimation

Radio Mapper (Android App)

Abstraction level

Time

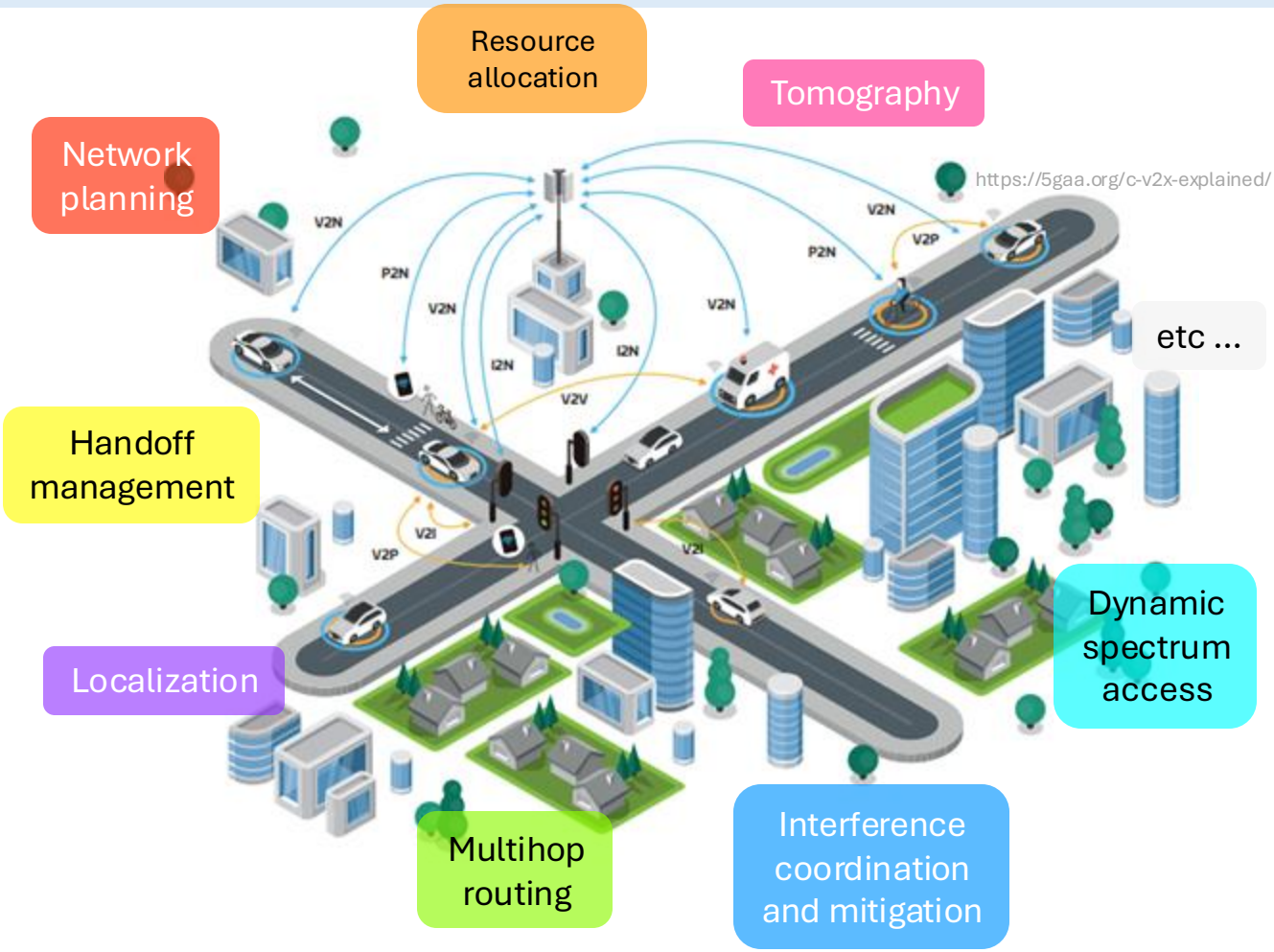
Radio Maps



A **radio map** provides a **metric of interest** in a radio communication environment.

Examples:

- Received signal power
- Interference power
- Power spectral density (PSD)
- Electromagnetic absorption
- Channel gain



D. Romero and S.-J. Kim, "Radio map estimation: A data-driven approach to spectrum cartography," IEEE Signal Process. Mag., 2022.

Overview

Radio map estimation (RME)

Motivation

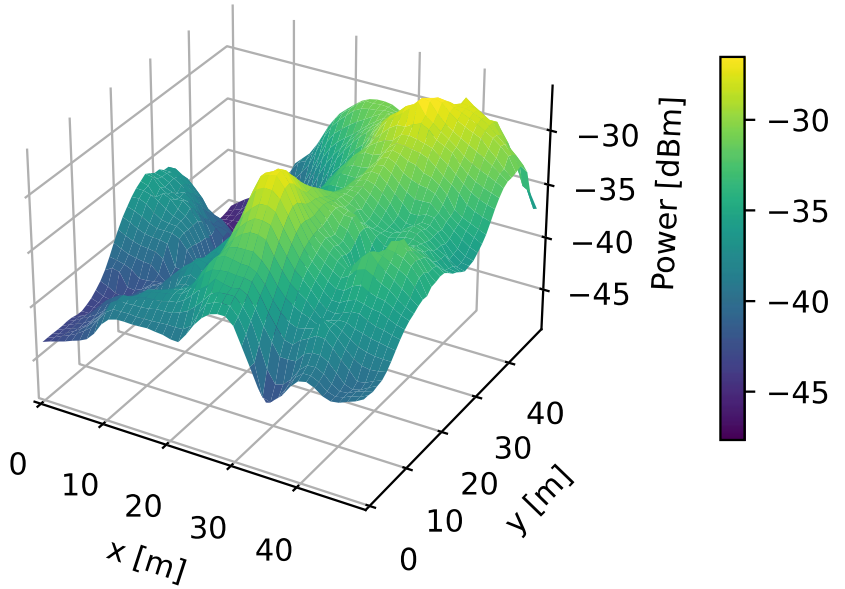
Determine radio map values where no measurements have been collected

Goal

Interpolative inference based on spatially distributed measurements

Contribution

Transformer-based Estimator (STORM)



Active sensing for RME

Motivation

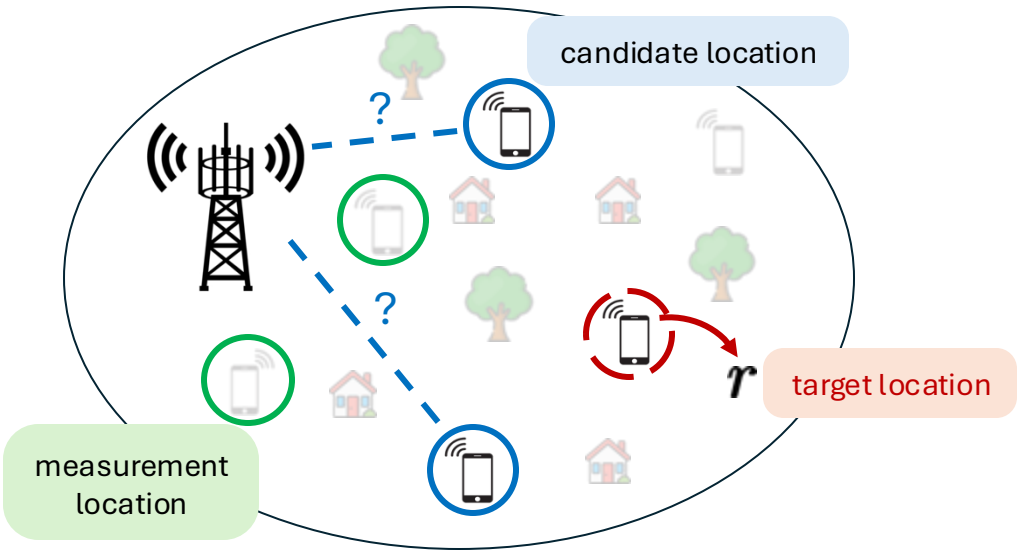
Minimization of drive tests (MDT)

Goal

Choose where to measure next

Contribution

Extended transformer-based network



Viet & Romero, “Spatial Transformers for Radio Map Estimation”, IEEE ICC 2025.

Radio Map Estimation: Problem Formulation

❖Given

- Measurements $\{(\boldsymbol{r}_n, \tilde{\gamma}_n)\}_{n=1}^{N}$

measurement location

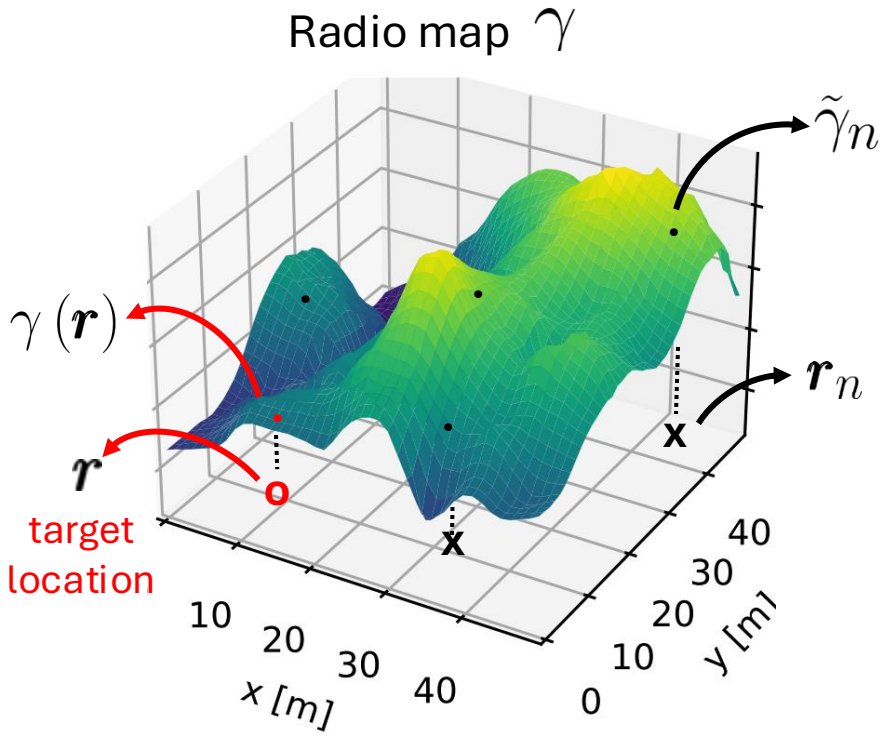
measurement value

$$\tilde{\gamma}_n = \gamma(\boldsymbol{r}_n) + \zeta_n$$

true signal power

measurement error

- Target location $\boldsymbol{r} \in \mathbb{R}^3$



❖Estimate $\gamma(\boldsymbol{r})$

Contribution in RME

Related work

- **Non-DNN**
  - KNN
  - Kriging
  - Sparsity-based inference
  - Dictionary learning
  - Kernel-based learning
  - Matrix completion
  - Graphical models

[Alaya-Feki et al. 2008]
[Bazerque et al. 2010]
[Kim et al. 2013]
[Romero et al. 2017]
[Schaufele et al. 2019]
[Ha et al. 2024]

- Oversimplify propagation characteristics
- Cannot learn complex patterns from data

- **DNN**
  - U-Net
  - Deep Completion Autoencoders
  - RadioUNet
  - DRUE

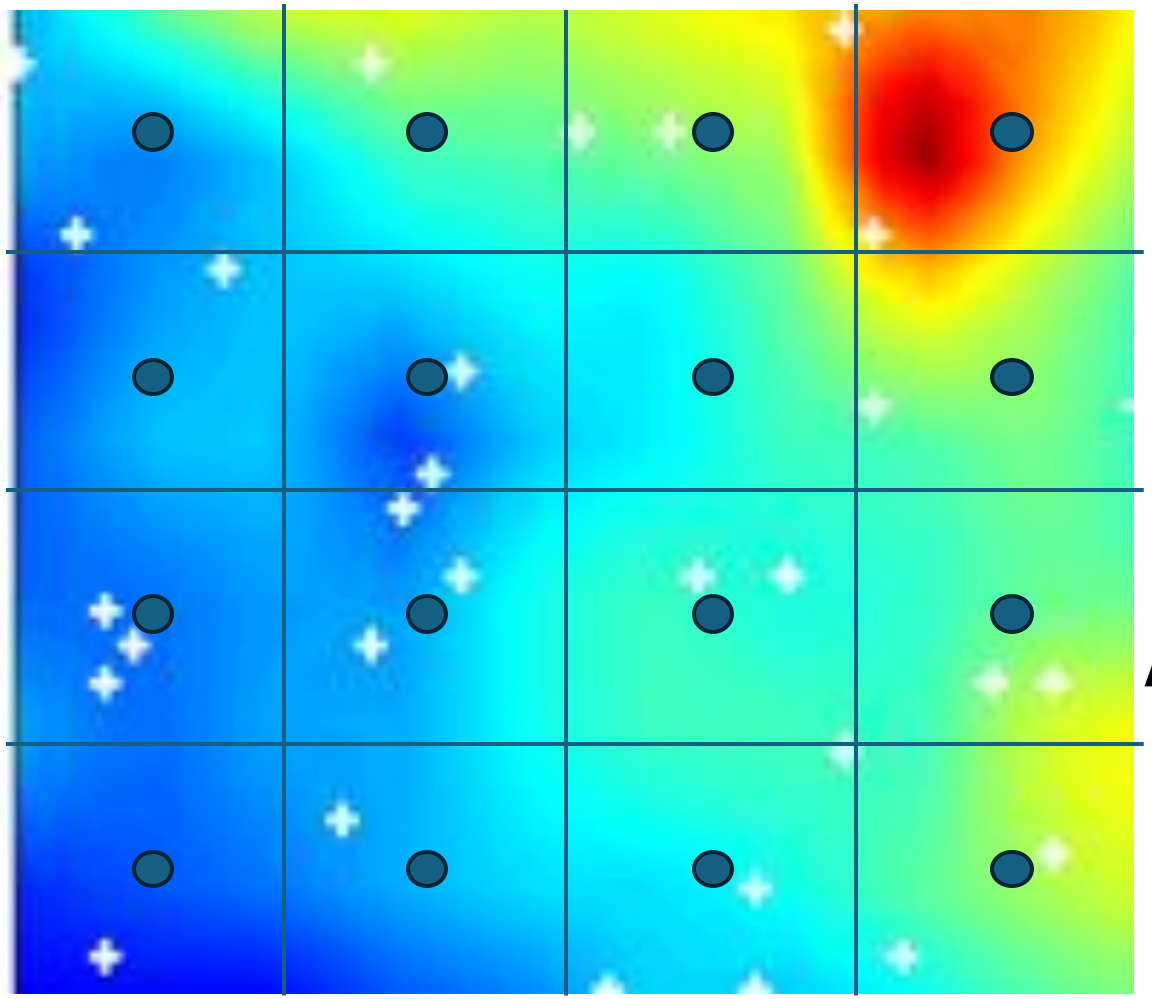
[Krijestorac et al. 2021]
[Teganya et al. 2021]
[Levie et al. 2021]
[Shrestha et al. 2022]

- Grid discretization
  - No translation equivariance
  - No rotation equivariance
  - Limited spatial resolution
- Large number of parameters

Contributions: **S**patial **T**ransf**O**rmer for **R**adio **M**ap estimation (STORM)

1. Full spatial resolution (gridless) -> **translation equivariance** + **rotation equivariance**
2. Estimate only at desired locations
3. Accommodate measurements outside the region of interest
4. Low computational complexity
5. Estimation performance sets the state of the art

Gridless vs. grid-based RME



Existing deep learning estimators rely on a grid

Disadvantages of grid discretization

- Limited spatial resolution
- Polynomial growth of the grid size
- Lack of translation equivariance
  (even discontinuous!)
- Lack of rotation equivariance

Additional limitations of grid-based estimators

- They estimate the map at **all** grid points
- Cannot accommodate measurements **out** of the estimation region

Function estimation formulation + transformer → Gridless estimator

Attention-Based RME

Spatial **T**ransf**O**rmer for **R**adio **M**ap estimation (**STORM**)

Desiderata

- Arbitrary resolution
- Invariant to permutations, translation, rotation
- Accommodates an arbitrary number of measurements

❖ Key idea:

transformer for function estimation

$$\theta^* := \measuredangle \sum_{n=1}^{N} \exp\left(\tilde{\gamma}_n\right)\left(\boldsymbol{r}_n - \boldsymbol{r}\right)$$

❖ Feature design

$$\{(\boldsymbol{r}_n, \tilde{\gamma}_n)\}_{n=1}^{N}$$

$$\boldsymbol{r}$$

Num. realizations

$$\left\{\left(\boldsymbol{X}^{(j)}, \boldsymbol{r}^{(j)}, \tilde{\gamma}^{(j)}\right)\right\}_{j=1}^{J}$$

Translation (ensure invariance)

Rotation (ensures invariance)

$$\begin{bmatrix} \tilde{\gamma} \\ \boldsymbol{U}\left(\theta^*\right)\left(\boldsymbol{r}_n - \boldsymbol{r}\right) \\ \theta_n \\ \cos\left(\theta_n\right) \\ \sin\left(\theta_n\right) \\ \left\|\boldsymbol{r}_n - \boldsymbol{r}\right\| \end{bmatrix}$$

n-th feature vector

$$F\left(\boldsymbol{X}\right) = \left[F_1\left(\boldsymbol{X}\right), \ldots, F_N\left(\boldsymbol{X}\right)\right]$$

F^{D}

F^{E}

L

$\mathrm{B_c}$

$\mathrm{B_c}$

$\mathrm{B_c}$

$\mathrm{B_c}$

L

Causal attention block

Linear transformation

$\boldsymbol{X}$

❖ Training

- Dataset: $\left\{\left(\boldsymbol{X}^{(j)}, \boldsymbol{r}^{(j)}, \tilde{\gamma}^{(j)}\right)\right\}_{j=1}^{J}$
- Loss function $\frac{1}{JN}\sum_{j=1}^{J}\sum_{n=1}^{N}\left(\tilde{\gamma}^{(j)} - F_n\left(\boldsymbol{X}^{(j)}\right)\right)^2$

Estimate of $\gamma\left(\boldsymbol{r}^{(j)}\right)$ given $\boldsymbol{X}_{1:n}^{(j)}$

Measurements

Features (other than measurements)

Estimates

Experimental Setup

❖ STORM:

Parameter	Value
Num. heads	2
Embedding dimension	48
Num. layers	4
	~100 k paramters

❖ Datasets:

Name	
Ray-tracing	[Teganya and Romero, 2020]
USRP	[Shrestha et al. 2023]
4G Gradiant	[Shrestha et al. 2024]

❖ Benchmarks*:

Name		Num. Param	
KNN			
Kriging			[Alaya-Feki et al. 2008]
KRR	Kernel ridge regression		[Romero et al. 2017]
DNN-1	Deep Completion Autoencoders	140 k	[Teganya et al. 2020]
DNN-2	RadioUNet	9 M	[Levie et al. 2022]
DNN-3	U-Net	9 M	[Krijestorac et al. 2021]
DNN-4	Autoencoders	60 M	[Shrestha et al. 2022]

* The benchmarks are trained as described in [Shrestha et al. 2024]

❖ Performance metric:

$$\mathrm{RMSE} \triangleq \sqrt{\mathbb{E}\left[\frac{1}{|\mathcal{N}_{\mathrm{nobs}}|}\sum_{n\in\mathcal{N}_{\mathrm{nobs}}}\left|\tilde{\gamma}_n-\hat{\gamma}\left(\boldsymbol{r}_n\right)\right|^2\right]}$$

Set of non-observed locations

Experimental Results

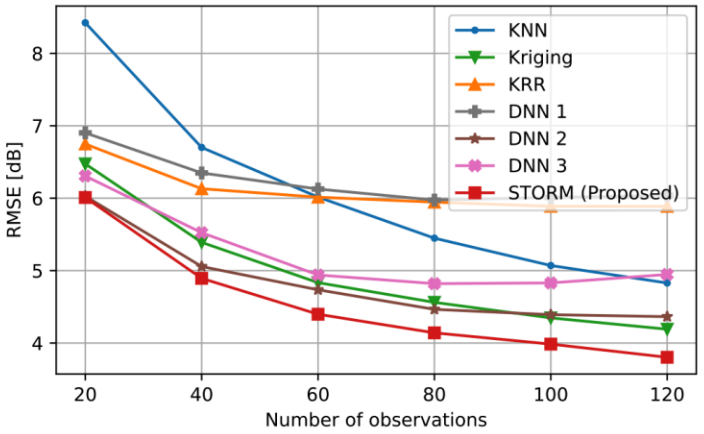


Fig. 3: RMSE for ray-tracing data vs. N when $L = 64$ m, $\Delta = 4$ m, and the estimators are trained with $N \in [20, 100]$.

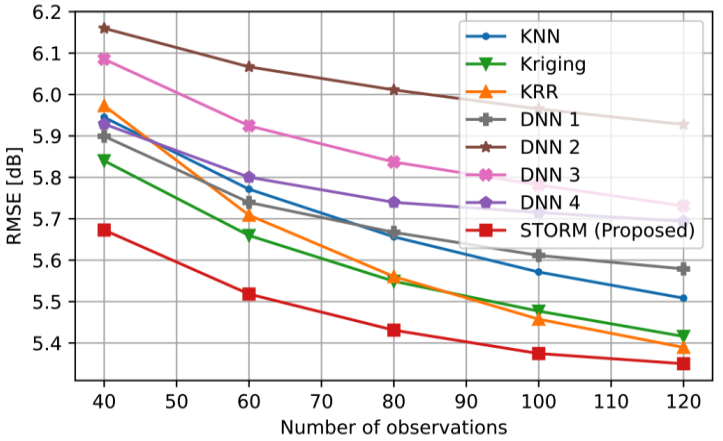


Fig. 4: RMSE for USRP data vs. the N when $L = 38.4$ m, $\Delta = 1.2$ m, and the estimators are trained with $N \in [40, 100]$.

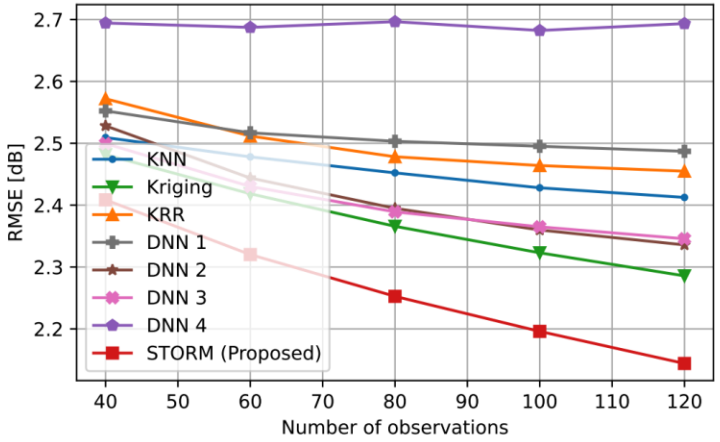


Fig. 5: RMSE for 4G data vs. N when $L = 64$ m, $\Delta = 4$ m, and the estimators are trained with $N \in [20, 100]$.

Contribution in Active Sensing

Related work

- **Non-DNN**
  - Uncertainty metric + shortest-path algorithm [Romero et al. 2020]
  - Random-active sampling selection [Liu et al. 2024]
  - Uncertainty-aware dynamic programming [Chen et al. 2025]
  - K-means clustering + maximum a posteriori [Chen et al. 2025]

- Grid discretization
  - No translation equivariance
  - No rotation equivariance
  - Limited spatial resolution
- Ray-tracing simulation dependence, limited applicability

- **DNN**
  - DRUE [Shrestha et al. 2022]
  - Kriging + Deep Q learning [Krijestorac et al. 2023]
  - Convolutional autoencoder + A* [Chen et al. 2025]
  - Uncertainty-Aware U-Net [Shi et al. 2025]

- Grid discretization
  - No translation equivariance
  - No rotation equivariance
  - Limited spatial resolution
- Large number of parameters

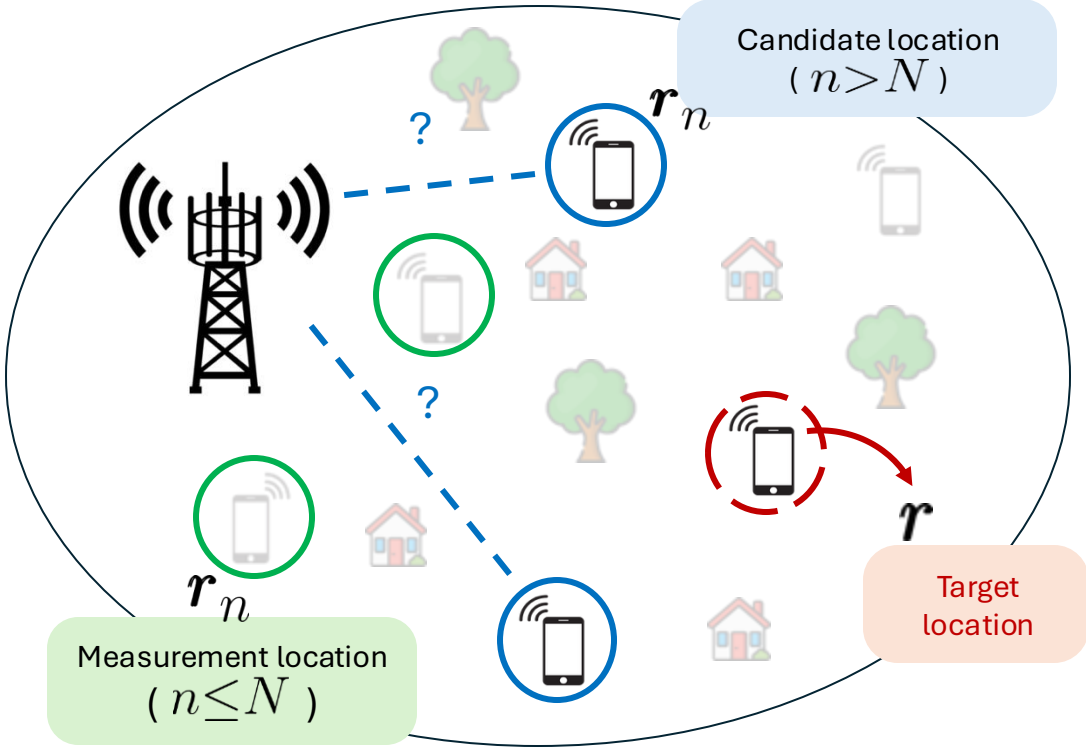
Contributions: Extended transformer-based network for *active sensing*

- **Attention-based selection** of the next measurement location
- Applicable to *minimization of drive tests* (MDT)
- **Low** computational complexity

Active Sensing: Problem Formulation

❖ **Given**

- Measurements $\{(\boldsymbol{r}_n, \tilde{\gamma}_n)\}_{n=1}^{N}$
- Candidate locations $\{\boldsymbol{r}_n\}_{n=N+1}^{N+M}$
- Target location $\boldsymbol{r}$



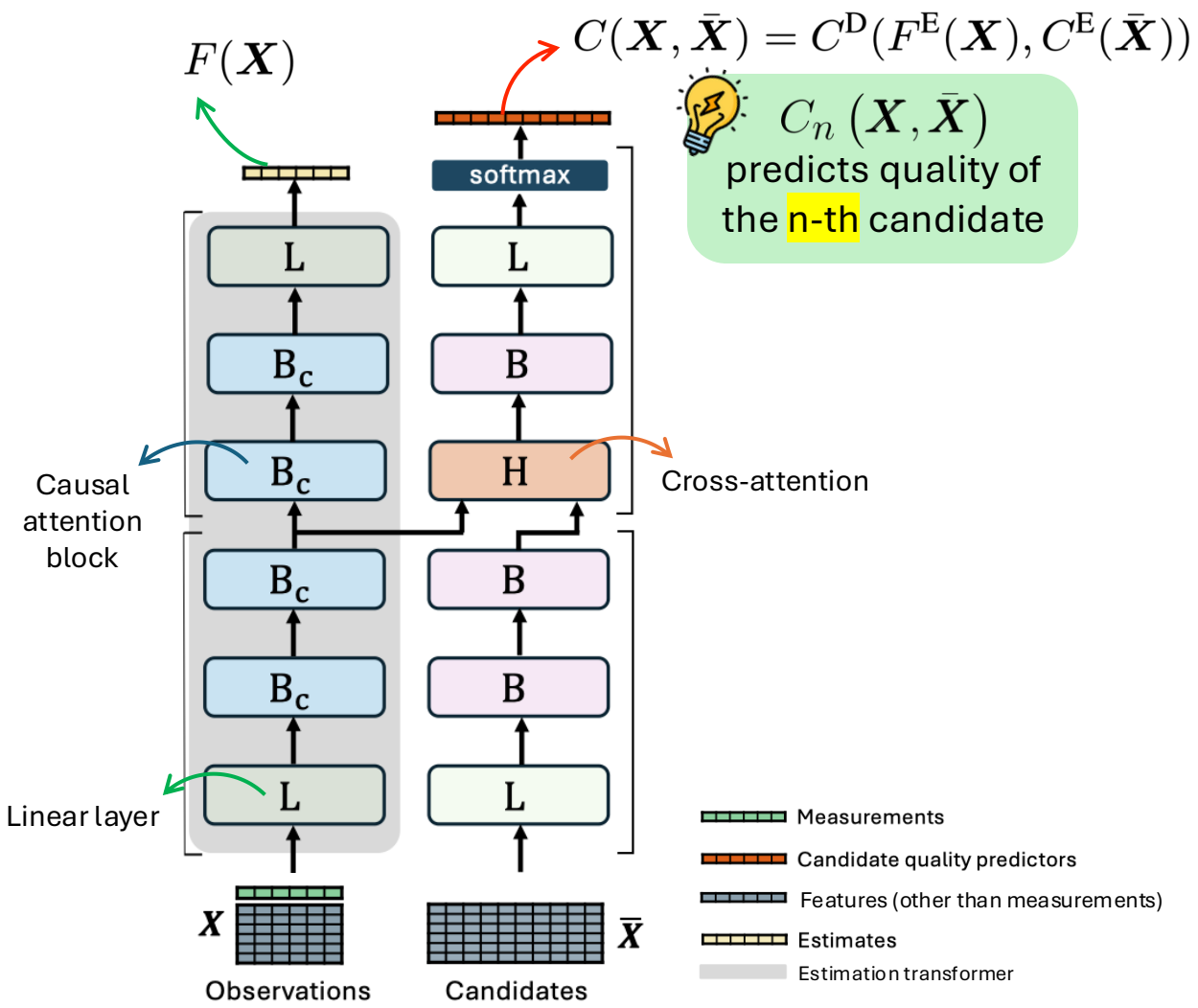
❖ **Find** $\underset{m}{\operatorname{argmin}} \left(\gamma(\boldsymbol{r}) - \hat{\gamma}\left(\boldsymbol{r}|\mathcal{M}(m)\right)\right)^2$

$m \in \{N+1, \ldots, N+M\}$

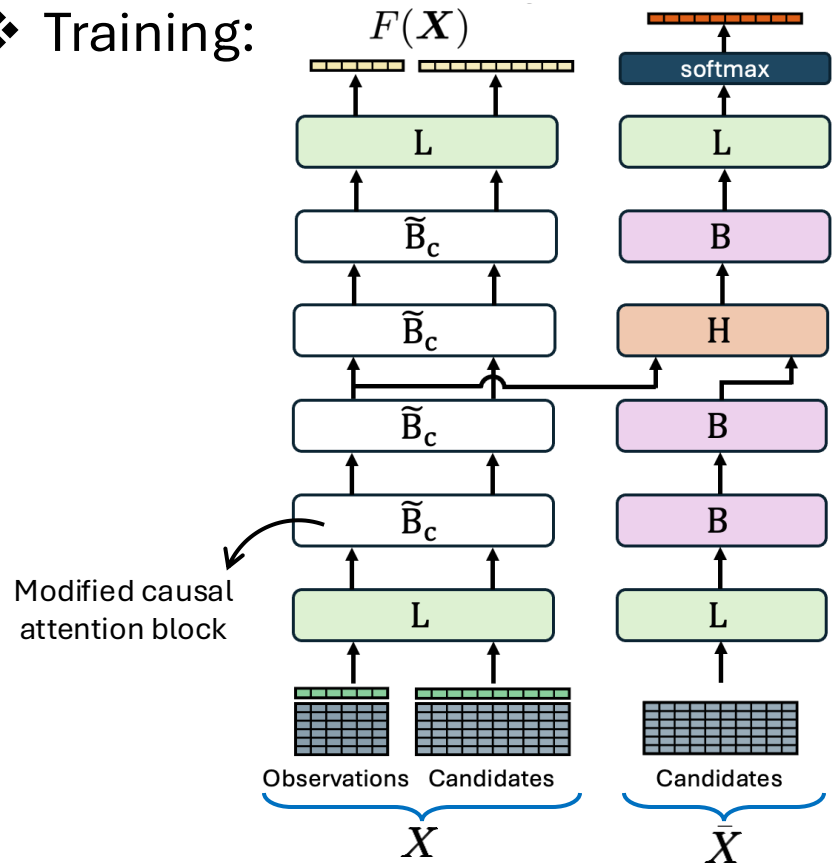
$\mathcal{M}(m) := \{(\boldsymbol{r}_n, \tilde{\gamma}_n)\}_{n=1}^{N} \cup \{(\boldsymbol{r}_m, \tilde{\gamma}_m)\}$

Attention-Based Active Sensing

❖ Inference mode



❖ Training:



$$\frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{2N}\sum_{n=1}^{N}\left(\tilde{\gamma}^{(j)} - F_n\left(\boldsymbol{X}^{(j)}\right)\right)^2 + \frac{1}{2}\left(\tilde{\gamma}^{(j)} - \sum_{m=1}^{M} C_m\left(\boldsymbol{X}_{1:N}^{(j)}, \bar{\boldsymbol{X}}^{(j)}\right) F_{N+m}\left(\boldsymbol{X}^{(j)}\right)\right)\right]$$

Estimate of $\tilde{\gamma}^{(j)}$ given $\boldsymbol{X}_{1:N}^{(j)}$ and $\boldsymbol{X}_m^{(j)}$

m-th entry

(N+m)-th entry

Experimental Results

❖ **STORM:**

Parameter	Value
Num. heads	2
Embedding dimension	48
Num. layers	4
	~100 k paramters

❖ **Datasets:**

Name	
Ray-tracing	
USRP	[Shrestha et al. 2023]

❖ **Performance metric:**

$$\mathrm{RMSE} \triangleq \sqrt{\mathbb{E}\left[\frac{1}{|\mathcal{N}_{\mathrm{nobs}}|}\sum_{n\in\mathcal{N}_{\mathrm{nobs}}}\left|\tilde{\gamma}_n-\hat{\gamma}\left(\boldsymbol{r}_n\right)\right|^2\right]}$$

Set of non-observed locations

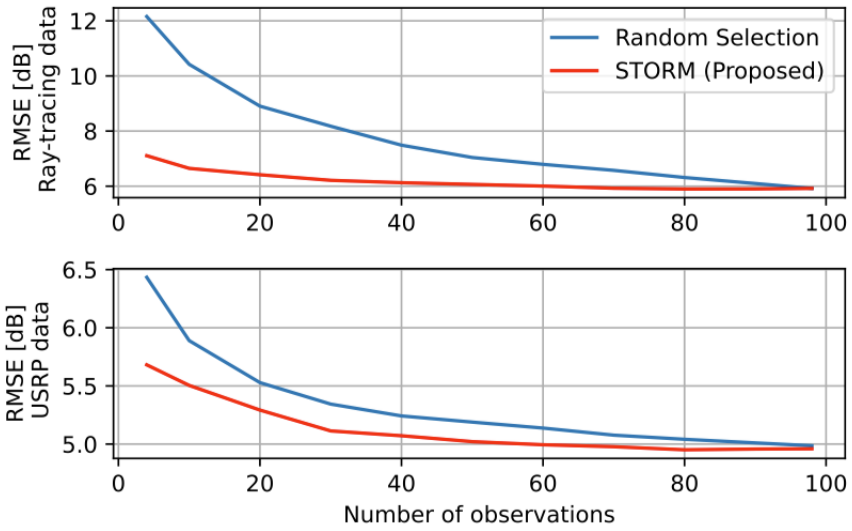
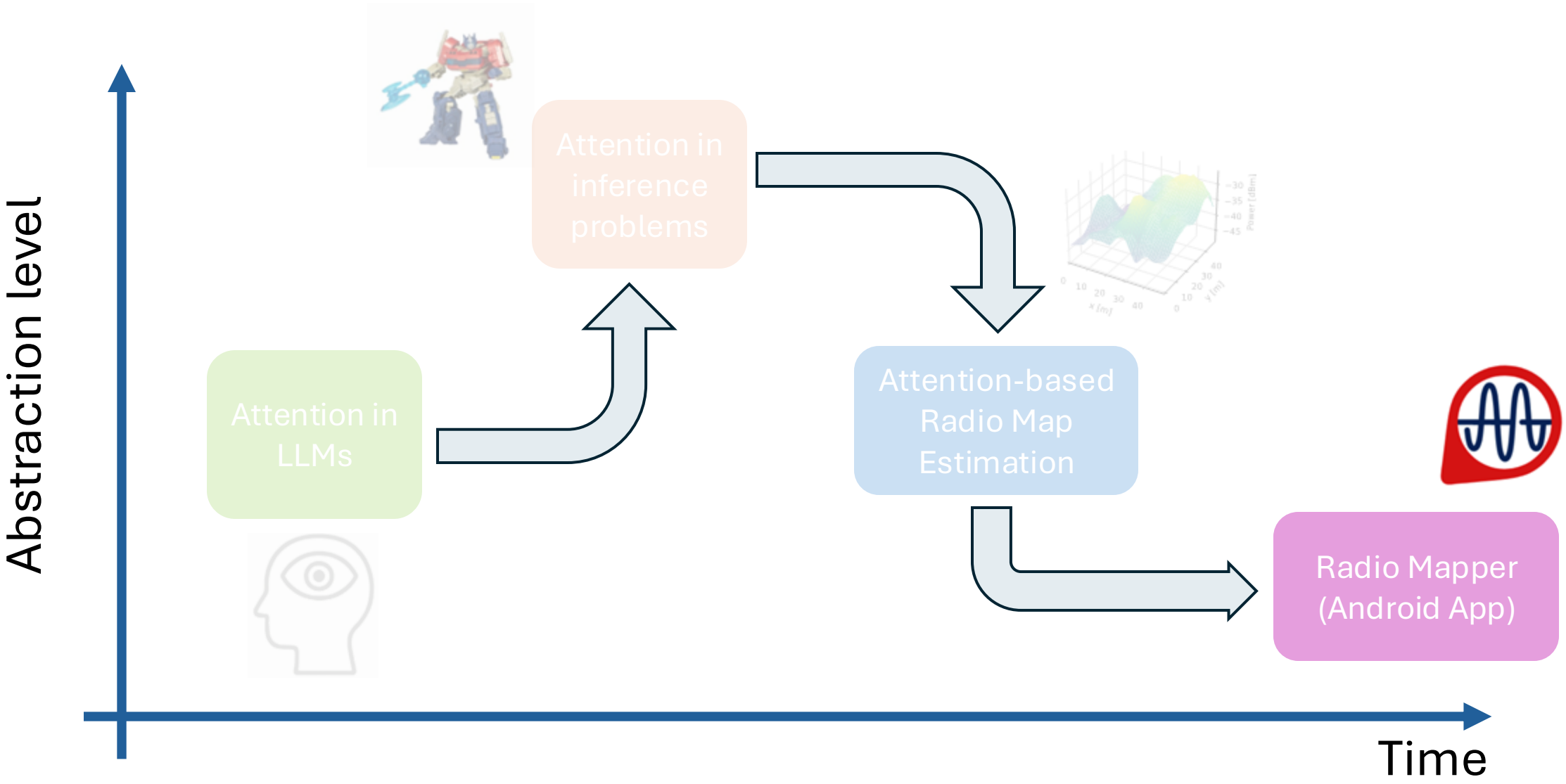


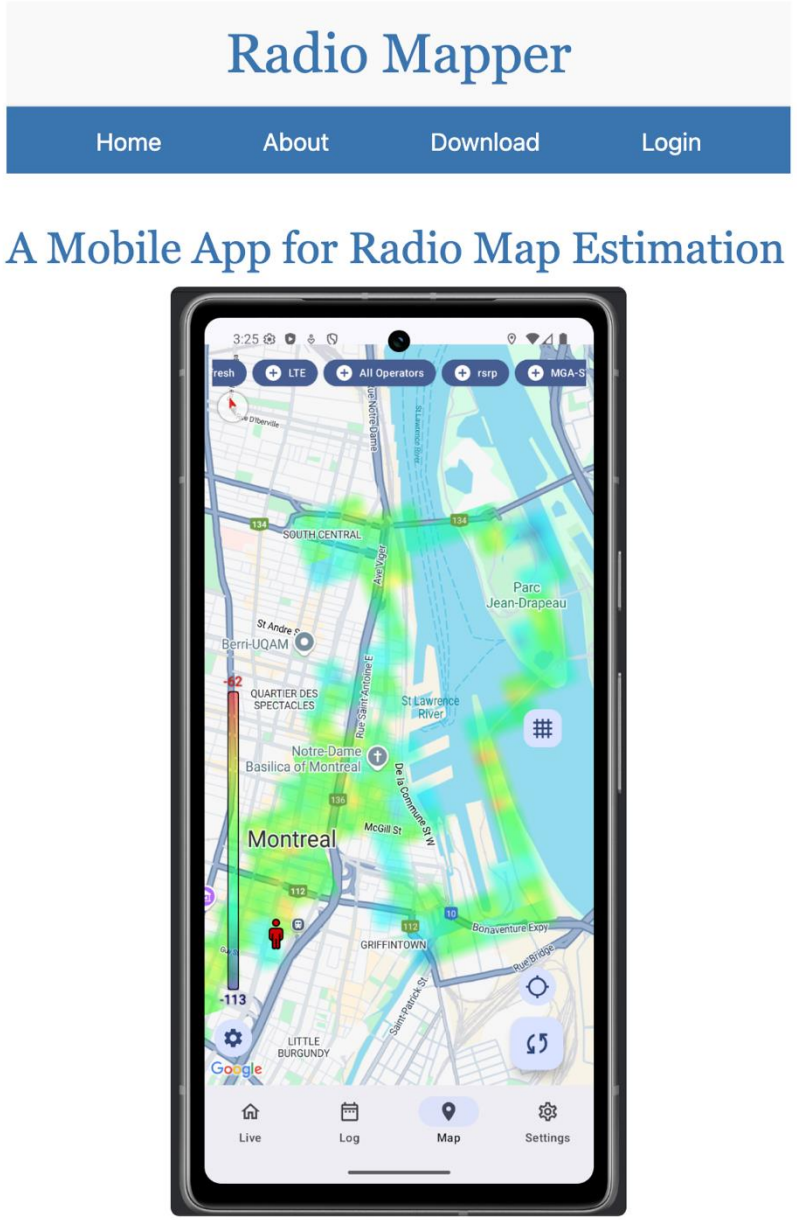
Fig. 6: RMSE vs. N for the active sensing problem with ray-tracing and USRP data. $D = 20$.

Outline of this talk



Radio Mapper

- **App** for RME with smartphones
  - ➢ Measurement collection
  - ➢ Map visualization
- Uses **STORM** (among other estimators)
  - ➢ Measurements sent to the server
  - ➢ Estimation in the server
- Mapped **metrics**
  - ➢ 5G → CSI-RSRP, CSI-RSRQ, CSI-SINR, SS-RSRP, SS-RSRQ, SS-SINR
  - ➢ LTE → RSRP, RSRQ, SINR, CQI
  - ➢ GSM → RSSI



Radio Mapper

- **Platforms**
  - Android: Testing phase → download at rme.uia.no

  - iOS: No → limitations of Apple devices
- **Challenges**
  - Earth is round → Mercator system → Estimation tiles → Zoom regions
  - Computation
    - Fast response → Multi-threading, code optimization…
    - GPU → memory, race conditions…
  - Storage
    - Database access
  - Security, privacy…

Radio Mapper

- **Development team**

UiA University of Agder Norway

Daniel Romero

gyleven development & design

Salatiel Genol

Paola Quintero

Alejandro Vidal

- **Sponsors**

The Research Council of Norway

UiA University of Agder Norway

Conclusions and Future Work

❖ **Conclusions**:

- **Attention is a general inference tool applicable in many different problems**
- **S**patial **T**ransf**O**rmer for **R**adio **M**ap estimation (STORM)
- **Active sensing**
- **Radio Mapper**

❖ **Future work**:

- **RME**
  - ➢ Experiments with indoor radio maps.
  - ➢ Accommodate side information, e.g. building locations.
- **Active sensing**
  - ➢ Metrics that account for estimation error on a set of locations
  - ➢ A more extensive experimental study

Thank you!!!