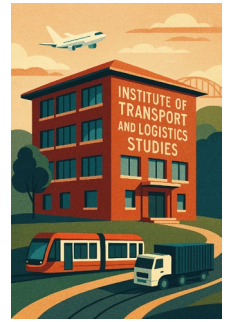
Personalised Travel Behaviour Prediction Using Multi-task Transformers with Mixture-of-ExpertsTransport

David A. Hensher AM, PhD, FASSA, FAITPM, FCILTA, Roads Australia John Shaw Medal
Professor and Founding Director,
Institute of Transport and Logistics Studies,
The University of Sydney Business
https://www.sydney.edu.au/business/about/our-people/academic-staff/david-hensher.html



I acknowledge my colleagues Dr Haoning Xi and Dr Jessica Shao who have been instrumental in the research

Problem Statement (Claims)

Limitations of Traditional Models

'Claim' that existing predictive models generally fail to capture personalised travel behaviour and lack multitask prediction capabilities.

Challenges in Multimodal Travel

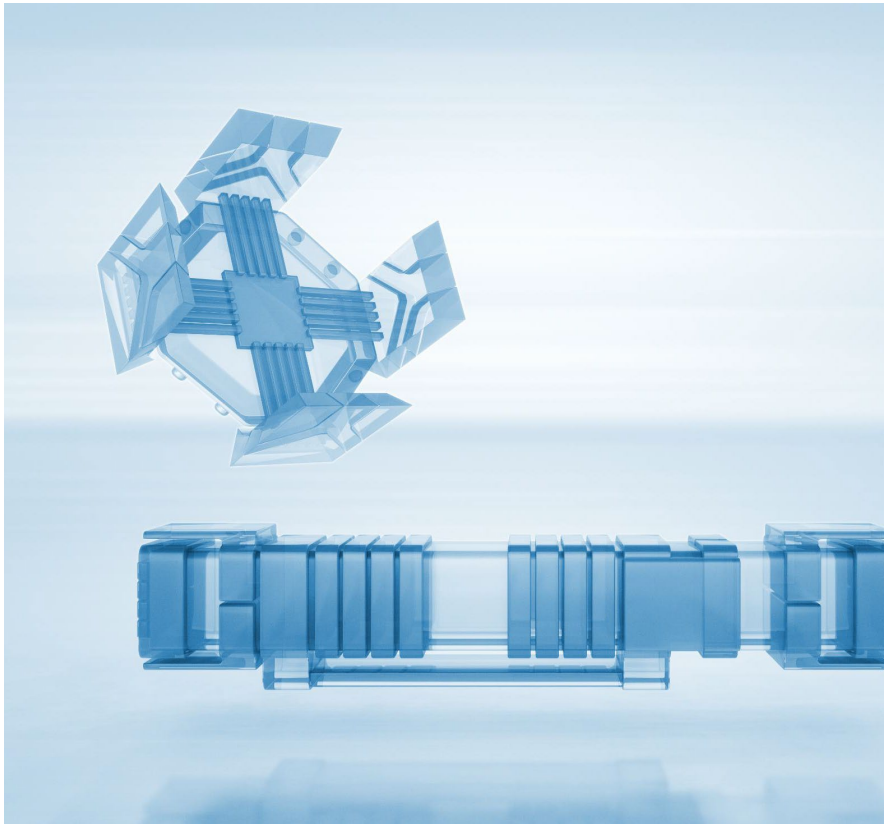
Traditional models do not adapt 'as well' to diverse and dynamic multimodal travel patterns. (But they could?)

Need for User-Centric Solutions

Planners require sophisticated tools that provide granular, personalised predictions across transport modes. Serious Heterogeneity Identification.

Xi, H., Shao, Z., Hensher, D.A., Nelson, J.D., Chen, H. and Wijayaratnae, K. (2025) A multi-task transformer with mixture-of-experts for personalized periodic predictions of individual travel behavior in multimodal public transport *Transportation Research Part C*, 179, 105287,online 5 August 2025,

Proposed Solution



Multi-task Transformer Architecture

The solution uses a multi-task Transformer to simultaneously predict travel mode, purpose, and (over) time effectively.

Mixture-of-Experts (MoE)Mechanism

MoE dynamically selects specialised subnetworks based on input features for personalised travel predictions.

Adaptability and Personalisation

The combined model adapts to individual travel patterns, enabling highly personalised predictions.

Scalable Multimodal Transport Analysis (Controversial)

The framework provides a scalable solution overcoming traditional model limitations for transport analysis.

Individual travel behaviour predictions (Some literature)

Existing literature on predicting users' travel preferences can be divided into two main streams: **statistical and deep learning-based models**. Statistical models, including Markov chains, decision trees and choice modelling (Hensher et al., 2015), have been widely employed in individual travel behaviour predictions.

Recent developments in deep learning have significantly improved individual travel behaviour prediction models **(not necessarily the understanding of behaviour!)**. Feng et al. (2018) introduced DeepMove, a sequence-to-sequence architecture with an attention mechanism improving next-location predictions (temporal).

Rossi et al. (2019) leveraged Long Short-Term Memory (LSTM) networks combined with attention modules for 'accurate' individual taxi destination prediction.

Qin et al. (2024) proposed a Deep Activity-Geographic-Sequential (DeepAGS) model to address variations in individual activity-location combinations.

Liang et al. (2021) developed a Multi-module Deep Learning Network (MDLN), integrating partial trajectories with individual preferences for destination prediction, while Liu et al. (2021) utilized graph embeddings.

Baidu Maps data for personalised multimodal recommendations.

Systematic comparisons between machine learning (ML) and traditional discrete choice models highlight their respective strengths. Mart´ın-Baos et al. (2023) showed models like random forests, extreme gradient boosting (XGBoost), and deep neural networks (DNNs) achieve 'high predictive accuracy', **though discrete choice models yield more stable behavioural estimates such as WTP.**

First paper to compare NNs and DC: Hensher, D.A. and Ton, T. (2000) A predictive assessment of neural networks and discrete choice methods. (Presented at 8th WCTR, Antwerp, July 1998) *Transportation Research Part E*, 36 (3), September, 155-172.

Thus, the Critical Elements

- We propose a **PTBformer-MMoE architecture***, a Transformer-based model within a multi-task i.e., Multi-gate Mixture-of-Experts (MMoE) framework, tailored for individual travel behaviour predictions.
- The proposed PTBformer-MMoE simultaneously handles multiple prediction tasks for each user periodically (e.g., monthly)
  - to **predict each user's monthly mode-specific travel frequency class** (bus and rail: low, medium, high; ferry and tram: low, high)
  - and a regression task for **predicting each user's monthly expected travel fare** (i.e., monthly travel expenditure across all modes serving as a personalised baseline fee) for the upcoming month.

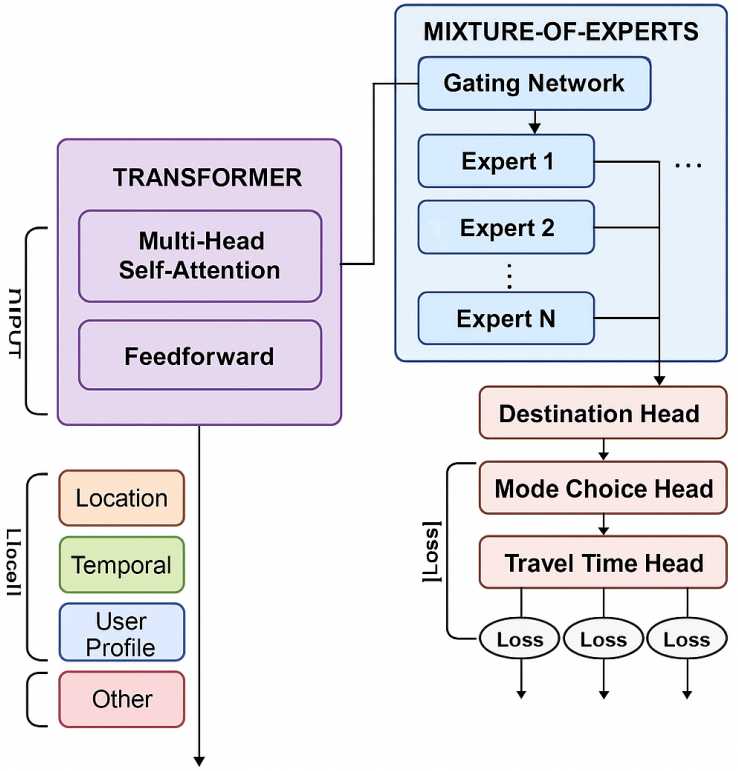
*** PTBformer** (**Public Transport Behaviour Transformer**)–a variant of Transformer-based models tailored for spatio-temporal or behavioural data (e.g., pedestrian trajectory prediction, traffic flow, or mobility patterns).

* **MMoE (Multi-gate Mixture-of-Experts)** – a neural network architecture designed to handle multi-task learning by sharing representations across tasks while allowing task-specific expert selection via gating mechanisms.

Instead of using one large model to handle all tasks, the system has **multiple specialised models** (called "experts"). A **gating mechanism** decides which expert(s) should handle a given input based on the task or data characteristics. **How It Works:** 1. **Experts**: These are sub-models trained to specialise in different aspects or types of data. 2.**Gating Mechanism**: A separate model (often a neural network) that: Takes the input, Predicts which expert(s) should be activated, Routes the input accordingly. **Benefits: Efficiency**: Only a subset of experts is used per input, reducing computation. **Specialisation**: Experts can be fine-tuned for specific tasks or domains. **Scalability**: Easier to scale by adding more experts without retraining the whole system.

Various Components in public transport systems: Summary

- **Individual travel behaviour predictions**
- **Multimodal transport bundling**
- **Multi-task learning**
- **Transformer-based models**
- **A key behavioural component: Attention Scores**



Personalized Travel Behaviour Prediction Using Multi-task Transformers with Mixture-of-Experts

Transformer-based models in public transport systems

The Transformer model has gained significant popularity due to its 'exceptional' capability in capturing long-range dependencies, especially evident in large language models (LLMs).

(Q: Can we do this using traditional time series?)

It has shown promise in time series forecasting and classification tasks. Despite their strengths, **traditional Transformers face challenges with multimodal and traffic-related data due to high complexity**.

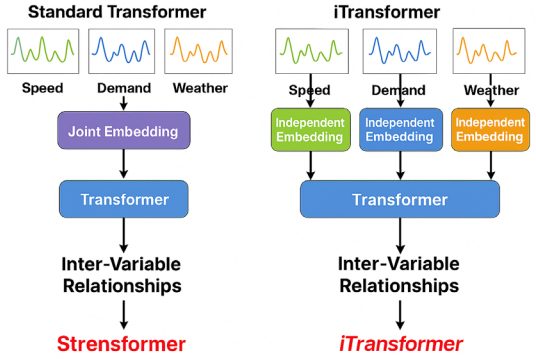
- **To overcome this, Liu et al. (2024) introduced iTransformer, which independently embeds variable-specific time series data to better model inter-variable relationships**. (e.g. between available modes)

iTransformer addresses this by:

- **Independently embedding each variable's time series**: Instead of combining all variables into one embedding, it treats each variable separately.
- This allows the model to **preserve the unique temporal patterns** of each variable.
- After embedding, it can **better learn how variables interact** over time, improving the modelling of **inter-variable relationships**.

Why It Matters

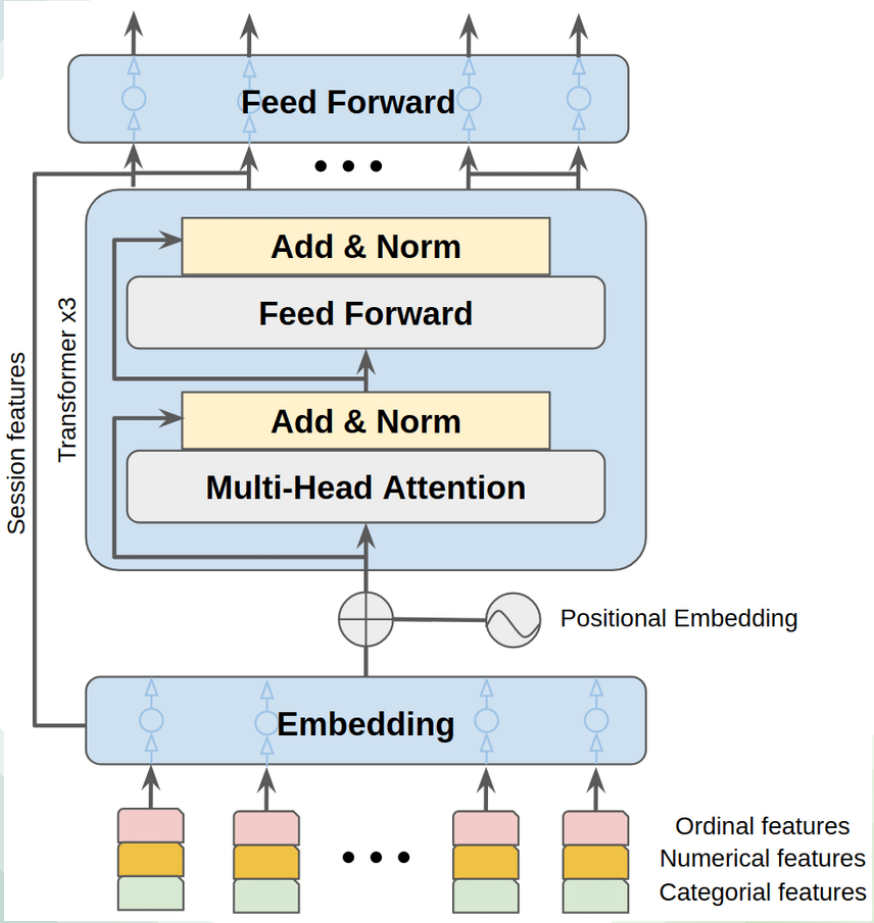
This approach is especially useful in domains like Transport and mobility modelling where different variables (e.g., speed, location, demand, modes, weather) have **distinct behaviours but still influence each other**.



Wang and Osaragi (2024) proposed a Transformer-based model, incorporating a self-attention mechanism and a Generative Pre-trained Transformer (GPT) to accurately generate and predict daily human mobility patterns responsive to contextual factors.

(1) input embedding, (2) transformer blocks and (3) feed forward layer

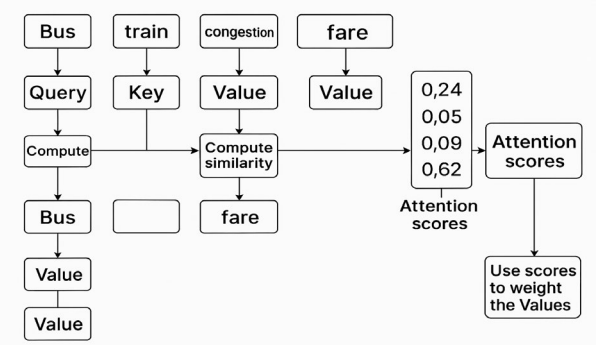
A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence.



key behavioural component Attention Scores Meaning Explained

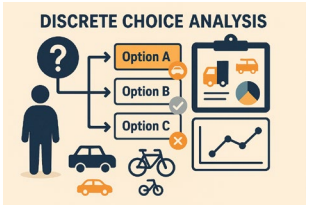
Definition: Attention scores quantify how much focus the model places on different parts of the input when making predictions.
In time series or transport modelling, this could mean:

- **How much past data (e.g., previous traffic flow) influences current predictions.**
- **Which variables (e.g., weather, demand, modal use, delays) are most influential.**



- For each feature (bus trips, rail trips, tram time, ferry time), the model asks: “Whom should I pay attention to when refining this feature?”
- It scores how related the current feature is to every other feature (including itself). Higher score ⇒ stronger relevance/correlation.
- These scores are turned into weights that sum to 1 (via softmax function)
  - The softmax function is a mathematical function often used in machine learning, especially in classification tasks involving neural networks. It converts a vector of raw scores (called logits – Hensher et al. 2015) into probabilities, making it easier to interpret the output of a model.
  - Here's the formula: Given a vector of scores z=[z1,z2,...,zn]z=[z1,z2,...,zn], the softmax function outputs:

$$Soft\max(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^{n} \exp(z_i)}$$



  - Outputs are probabilities: Each value is between 0 and 1, and the sum of all outputs is 1.
  - Emphasises larger values: Higher input scores get disproportionately higher probabilities.
  - Used in classification: Commonly used in the final layer of a neural network for multi-class classification.

- The feature is then updated as a weighted blend of all features’ information, where the weights reflect those relevance scores.
- In example, bus assigns most weight to rail and some to itself, and very little to tram/ferry, so the bus representation is mainly shaped by rail (capturing their coupling).
- Sparse modes (tram/ferry) still get non-zero weights, so the model can retain useful cross-mode signals even when data are thin.
- **That’s self-attention: each feature learns which other features matter right now and mixes them accordingly.**

For each token (e.g., “Bus”), we calculate how similar its **Query vector** is to the **Key vectors** of all tokens (Bus, train, congestion, fare).
This is done using a **dot product**:

$$\text{score}_{ij} = Q_i \cdot K_j$$

So if “Bus” is the current token, we compute:

$$\text{score}_{Bus,Bus}, score_{Bus,train}, score_{Bus,congestion}, sco$$

These raw scores could look like:

$$[2.3, 0.5, 0.9, 5.8]$$

How to calculate:

Each token (e.g., “Bus”) is represented by a **vector of numbers** (embedding).The model learns **weight matrices** for Queries (Q), Keys (K), and Values (V). For a token: $Q = Embedding \times W_Q, K = Embedding \times W_K$ where W_Q and W_K are learned during training. Then the raw score is $score_{ij} = Q_i \cdot$(dot product between the Query of token i and Key of token j). **2.3** comes from the dot product of two learned vectors after many training steps on large datasets

- Normalise similarities into **attention scores** (e.g., 0.24, 0.05, 0.09, 0.62). (Logit model)
- Use these scores to weight the Values and build the final representation.

In this example, the model might pay **most attention to “fare” (0.62)** when processing “Bus,” meaning fare strongly influences bus-related decisions

Who (What) are Expert 1,2,...(TBFormer)?

Instead of using one large model to handle all tasks, the system has **multiple specialised models** (called "experts"). A **gating mechanism** decides which expert(s) should handle a given input based on the task or data characteristics. **How It Works:** 1. **Experts**: These are sub-models trained to specialise in different aspects or types of data. 2.**Gating Mechanism**: A separate model (often a neural network) that: Takes the input, Predicts which expert(s) should be activated, Routes the input accordingly. **Benefits: Efficiency**: Only a subset of experts is used per input, reducing computation. **Specialisation**: Experts can be fine-tuned for specific tasks or domains. **Scalability**: Easier to scale by adding more experts without retraining the whole system.

- **TBFormer** stands for **Public Transport Behaviour Transformer**. It's a specialised Transformer-based architecture designed to handle **periodic predictions of individual travel behaviour** using large-scale public transport data. It was developed as part of a framework called **PTBFormer-MMoE** (Multi-gate Mixture-of-Experts), which combines:
- **Multi-mode Transformer**: Uses multi-feature self-attention to model continuous time-series travel data.
- **OD Transformer**: Captures origin-destination-specific travel features using multi-OD self-attention.
- This model was trained on a massive dataset of **0.96 billion travel records** from **1.58 million users in Queensland, Australia**, covering buses, trains, ferries, and trams between January 2021 and January 2023.

Expert 1: Multi-mode Transformer

- **Focus**: Captures **temporal patterns** in travel behaviour.
- **Why it's needed**: Travel behaviour varies over time (e.g., weekdays vs weekends, peak vs off-peak).
- **How it works**: Uses **multi-feature self-attention** to model time-series data like trip frequency, duration, and mode usage.

Expert 2: OD Transformer

- **Focus**: Captures **origin-destination (OD) specific features**.
- **Why it's needed**: Travel decisions are influenced by where people start and end their trips.
- **How it works**: Uses **multi-OD self-attention** to learn patterns based on OD pairs (e.g., home to work, school to gym).

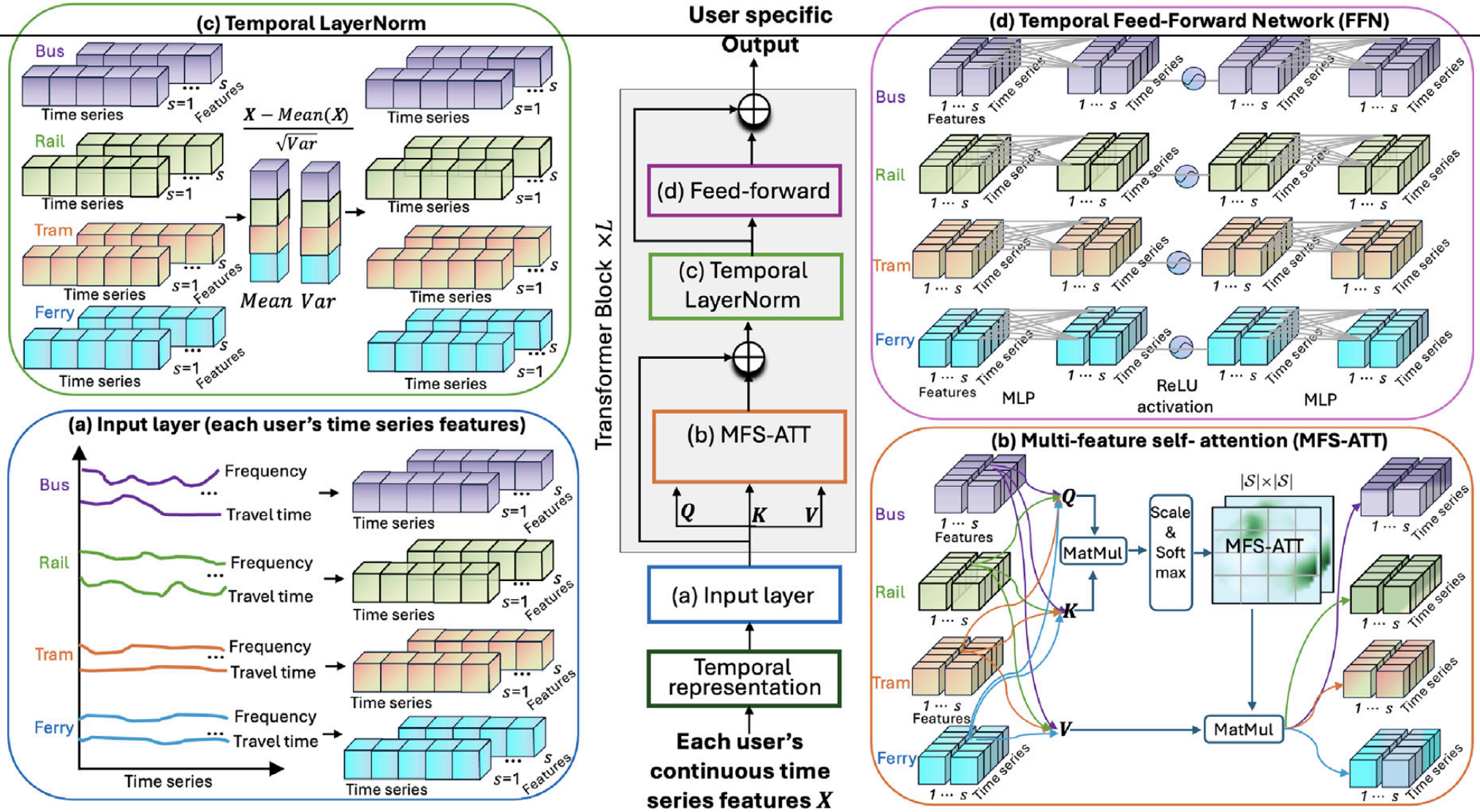
Expert 3: Behaviour Fusion Expert

- **Focus**: Combines insights from Experts 1 and 2.
- **Why it's needed**: Travel behaviour is a **mix of temporal and spatial patterns.**
- **How it works**: Fuses outputs from the other experts to make **final predictions** about individual travel behaviour.

The structure of Multi-mode Transformer for our PT application (The Data and linkages)

The Multi-mode Transformer is designed to process each user's continuous time-series travel features:the Multi-mode Transformer consists of a stack of $\mathcal{L}$ blocks, each including an input layer, multi-feature self-attention (*MFS-ATT*), and temporal layer normalisation, feed-forward networks, etc. The core of Multi-mode Transformer is multi-feature self attention, shorted as MFS-ATT.

Temporal Layer Normalisation =*family of normalisation techniques where normalisation statistics or parameters are computed using temporal information—either from past timesteps, time-dependent learnable parameters, or temporal context networks—allowing better modelling of sequential and time-series data.*



Example of a Multi-head multi-feature self-attention (*MFS-ATT*) score matrix.

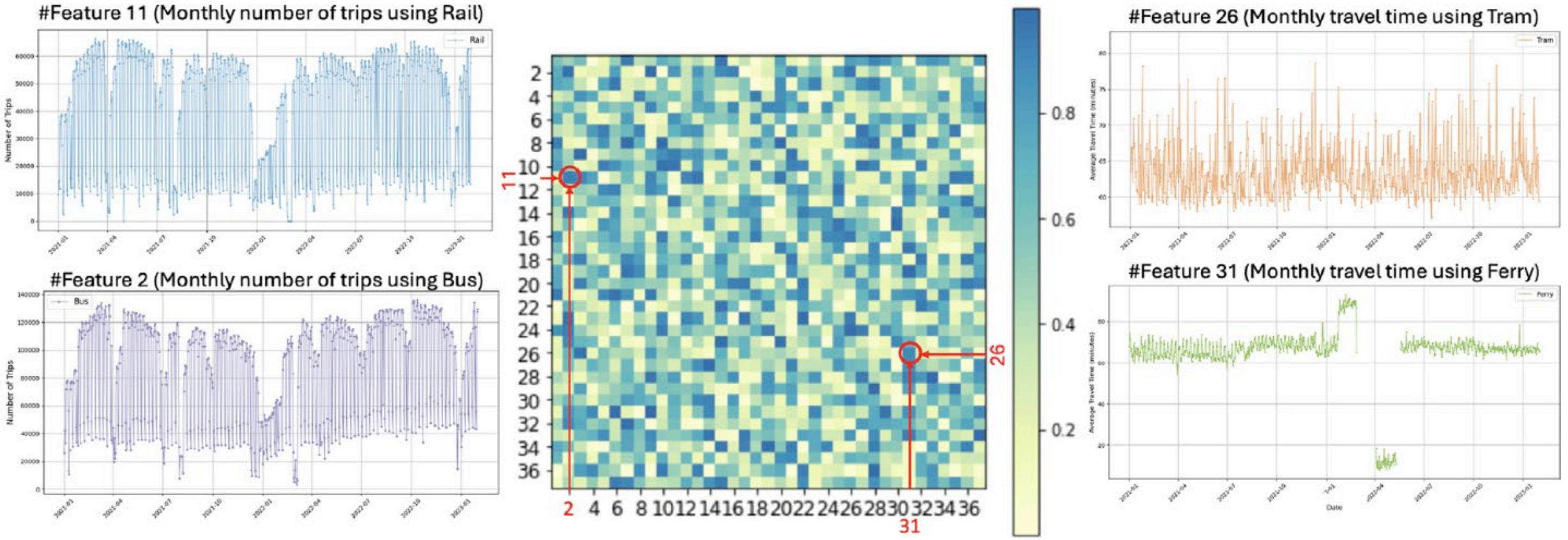


Fig. Visualisation of an attention matrix of MFS-ATT for a user sample **(lighter shade=less interdependency).**

This figure visualises an attention matrix of MFS-ATT from a Multi-mode Transformer for a user sample, showing **interdependencies between mode-specific and cross-mode travel features**. The intensity of each cell indicates how strongly one feature attends to another; for example, Feature 2 (monthly bus trip count) and Feature 11 (monthly rail trip count) exhibit particularly high mutual attention. **This strong coupling between bus and rail usage suggests integrated or complementary multimodal travel behaviour, implying that the user tends to utilise bus and rail in tandem**. In contrast, features representing rarely used modes (Feature 26: monthly tram travel time, and Feature 31: monthly ferry travel time) still receive meaningful attention from the model despite the data sparsity. The presence of non-negligible attention weights for these sparse features indicates that the model captures their latent relationships with other modes rather than ignoring them.

The attention matrix highlights the Multi-mode Transformer's capacity to learn complex cross-modal dependencies among travel modes in a multimodal transit system, underscoring the MFS-ATT model's effectiveness in modeling a user's multimodal travel behaviour. You could say there are some significant cross-elasticity effects. (Note: This can be handled in DCA with advanced choice models).

OD Transformer (MOD-ATT): Giving spatial OD identity to the evidence

OD Transformer consists of a stack of N blocks, each including an input layer based on Bipartite Graph embedding, multi-OD self-attention, FFN, layer normalisation, etc.

An **FFN (Feed-Forward Neural Network)**—also called a **fully connected network** or **MLP (multi-layer perceptron)**—is one of the foundational neural network architectures used in deep learning.

It is called *feed-forward* because information flows **one way: input → hidden layers → output**, with no cycles or recurrence.

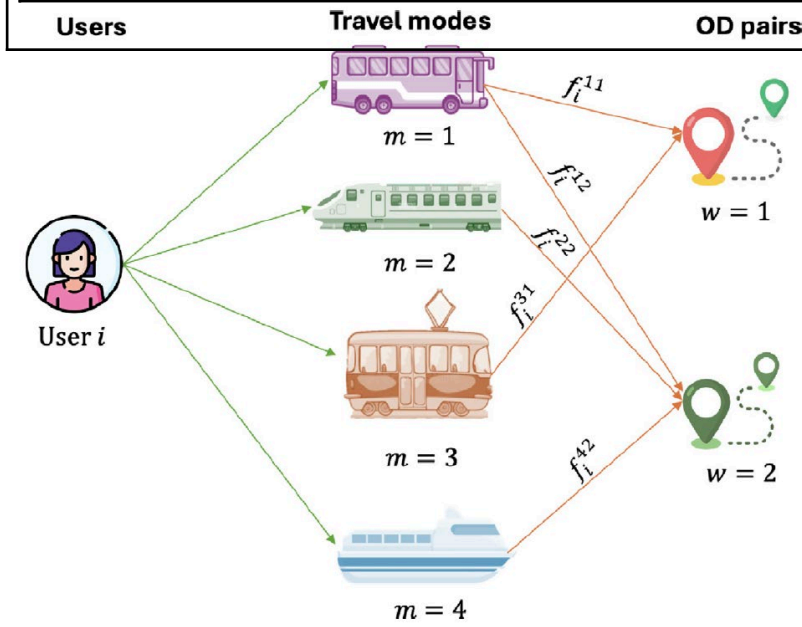


Fig. Bipartite graph representation

OD feature extraction is based on Bipartite Graph embedding since users frequently make consistent trips between specific origin–destination (OD) pairs via specific travel modes.

This is a key component in the OD Transformer module, which is designed to model complex spatial (OD-pair) and multi-mode relationships using attention scores. *MOD-ATT* allows OD Transformer to dynamically learn and prioritise dependencies across different modes and different types of OD pairs.

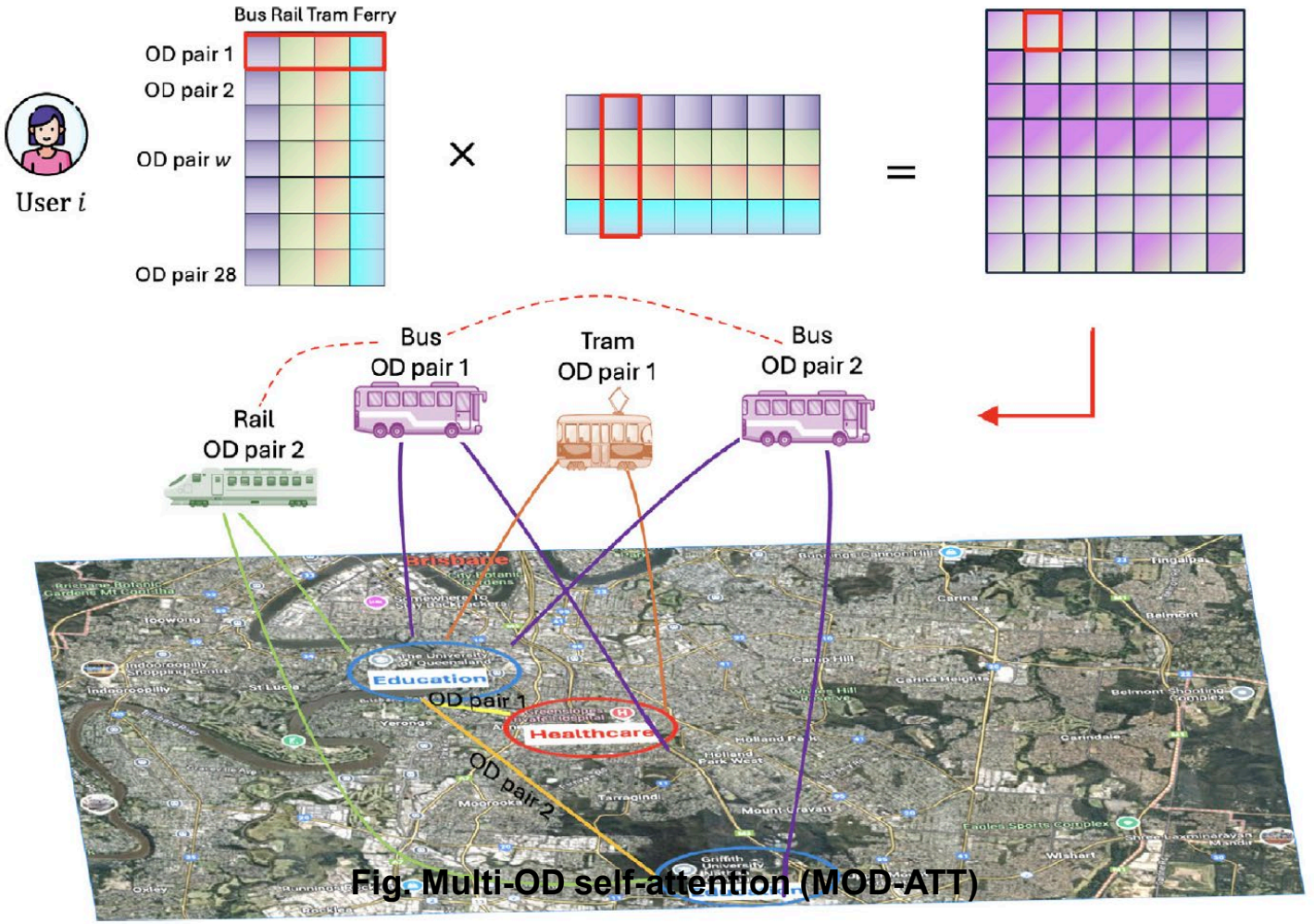
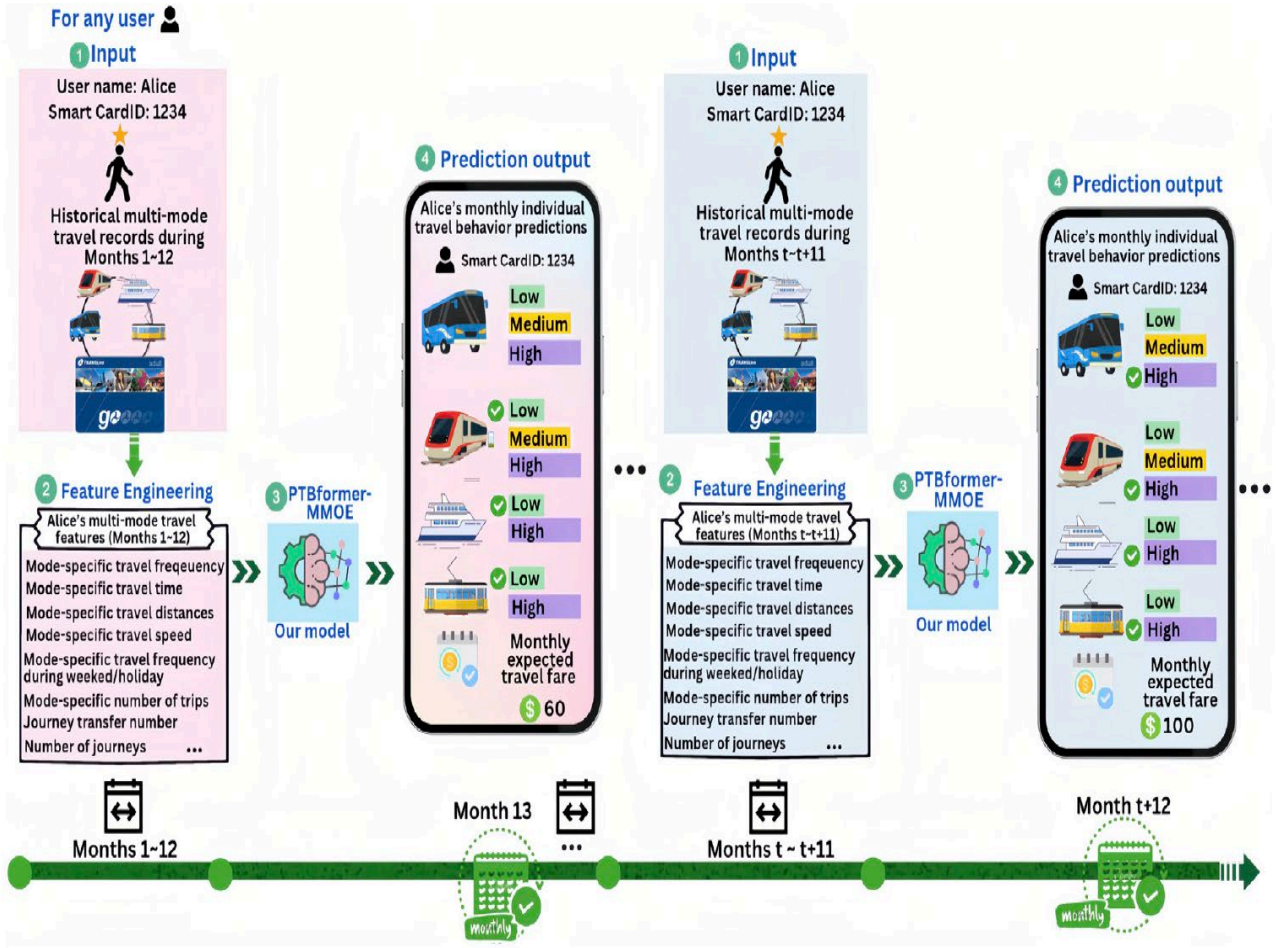


Fig. Multi-OD self-attention (MOD-ATT)

Example Application

Personalised periodic predictions of individual travel behaviour in multimodal public transport

- Figure illustrates the personalised prediction process with Alice as an example.
- Initially, Alice's historical multimodal travel data (bus, rail, ferry, tram) from smart card records are collected **(Step 1)**, and processed into monthly travel features such as travel frequency, time, distance, speed, weekend/holiday usage, and trip count **(Step 2).**
- The input features are then used to train the proposed PTBformer-MMoE model **(Step 3).** Utilising a rolling 12-month historical window, the model predicts each user's upcoming monthly mode-specific travel frequency class and monthly expected travel fare **(Step 4).**
- For example, Alice's predicted bus usage frequency class remains "high" in Month 13 and Month *t*+12, while her predicted monthly expected fare increases from $60 to $100 due to usage patterns on other modes (e.g., rail, ferry, tram), travel distances, or transfer numbers, etc. **(Far right box)**
- This example indicates the practical importance of jointly predicting monthly mode-specific travel frequency classes and monthly expected travel fare, as each output is essential to **designing personalised multimodal subscription plans that dynamically adapt to evolving user travel needs.**
  - **I query whether an advanced DCM is just as good?**



Experiments and evaluation

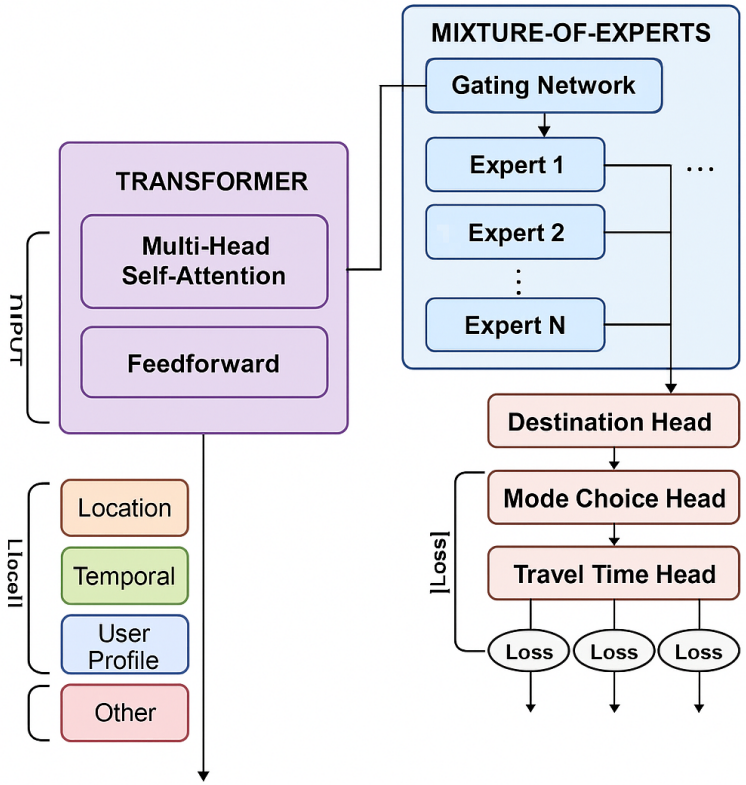
An **ablation study** is a method used in machine learning and deep learning research to understand the contribution of different components of a model or system. We **remove one component at a time** (e.g., remove the spatial encoder), r**etrain the model** or run it on the same dataset, and **compare performance metrics** (e.g., accuracy, RMSE, MAE) with the full model. DWA = dynamic weighting average (DWA) mechanism to adjust the weights of different prediction tasks during the learning process (To address the inherent data imbalance in the dataset, where bus and rail provide abundant data while ferry and tram generate sparse data)

Ablation study: Performance comparison of different modules in PTBformer-MMoE.

Metrics	Tasks	Without Multi-mode Transformer	Without OD Transformer	Without MMOE (Single task)	Without DWA	**PTBformer-MMoE**
Precision (%)	Bus	69.73%	70.67%	72.08%	74.1586	**74.82%**
	Ferry	94.78%	97.27%	97.43%	97.33%	**97.44%**
	Tram	99.75%	99.76%	99.83%	99.79%	**99.89%**
	Rail	70.28%	71.18%	75.11%	74.74%	**75.57%**
	All modes	83.64%	84.72%	86.16%	86.51%	**86.91%**
Recall (%)	Bus	71.74%	72.54%	72.84%	74.97%	**75.48%**
	Ferry	97.23%	97.35%	97.77%	97.72%	**98.17%**
	Tram	99.75%	99.76%	99.71%	**99.79%**	99.78%
	Rail	73.91%	74.43%	76.58%	76.21%	**76.80%**
	All modes	85.66%	86.02%	86.85%	87.17%	**87.47%**
F1 score	Bus	0.6881	0.6964	0.7145	0.7421	**0.7494**
	Ferry	0.9563	0.9589	0.9632	0.9722	**0.9723**
	Tram	0.9632	0.9810	0.9732	0.9858	**0.9982**
	Rail	0.6810	0.7085	0.7453	0.7578	**0.7591**
	All modes	0.8413	0.8384	0.8564	0.8655	**0.8696**
MSE	Expected travel fare	1.8981	1.8120	1.6832	1.6322	**1.2287**
MAE		0.6226	0.5856	0.5320	0.4965	**0.4812**
MAPE (%)		25.32%	23.82%	20.25%	18.57%	**17.65%**

This Table presents an ablation study designed to evaluate the contributions of each module in the proposed PTBformer-MMoE framework, e.g., the Multi-mode Transformer, OD Transformer, MMoE, and DWA, by comparing each variant against the proposed PTBformer-MMoE. **F1 score is the harmonic Mean of two metrics (precision vs recall). The results demonstrate the "superior" performance of the PTBformer-MMoE architecture across all evaluation metrics**

Final commentary:Claims



Improved Prediction Accuracy

The model 'significantly' improves accuracy in predicting mode choice, trip purpose, and travel time compared to traditional models.

Enhanced Interpretability

Better interpretability of results supports clearer insights for transport planners and stakeholders.

Personalised Transport Planning

Personalised predictions enable more targeted and efficient transport planning tailored to user needs.

Multi-Task Efficiency

The model handles multiple tasks simultaneously, reducing the need for separate models and streamlining analysis.

Caveats re Hybrid Modelling: Combining Theory and Flexibility – are we overdoing it and not giving enough credit to traditional DCA?

One claimed promising direction is the development of **hybrid models** that retain the interpretability and theoretical grounding of DCM while leveraging the flexibility of ML.

For instance:
- ML can be used to estimate **latent variables** (e.g., attitudes, perceptions) that feed into a structural choice model.
- ML can model **error components** or **heterogeneity** in preferences, which are then incorporated into a mixed logit framework.
- Bayesian approaches can integrate ML priors into probabilistic choice models.

This synergy allows for richer behavioural insights while maintaining the ability to conduct welfare analysis, policy simulation, and elasticity estimation.

Seems to me that the Machine Learning (ML) folks typically assume we just have an MNL choice model with linear in parameters and attributes!

Big mistake, and we have captured a lot of heterogeneity in our advanced behaviourally rich models with process heuristics etc. (Hensher et al. 2025 and Third Edition 2027 in progress of ACA (CUP).

We do most of these hybrid claims within a DCA setting already!

A synthesis perspective on Discrete Choice Models (DCMs) versus Machine Learning (ML) approaches, especially in the context of behavioral modelling and policy analysis:

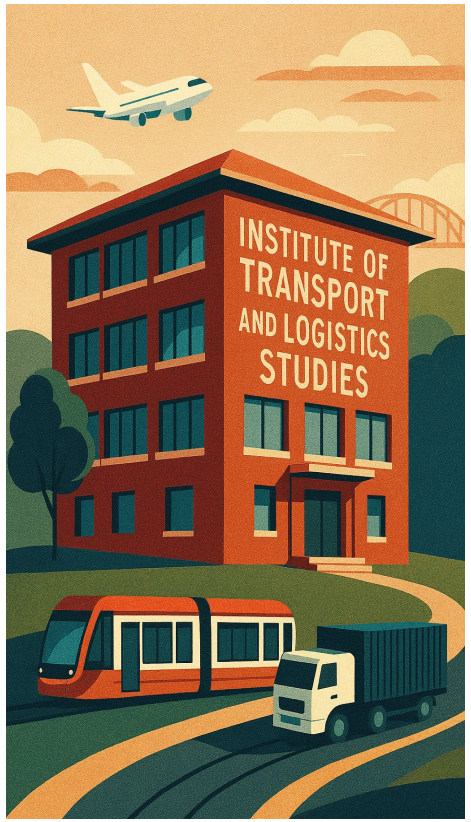
Aspect	Discrete Choice Models (DCMs)	Machine Learning (ML)
Theoretical Foundation	Grounded in **random utility theory** and economic rationality	**Data-driven**, with minimal or no theoretical assumptions
Interpretability	High – parameters have **clear economic meaning** (e.g., willingness to pay)	Often low – models like neural nets are **black boxes**
Predictive Accuracy	Good, especially when theory aligns with behavior	Often superior in **pure prediction tasks**, especially with large data
Policy Simulation	Strong – can simulate **counterfactuals** and **policy impacts**	Weak – lacks structural basis for reliable counterfactuals
Data Requirements	Performs well with **smaller, structured datasets**	Excels with **large, high-dimensional datasets**
Behavioral Insight	Designed to uncover **preferences and decision processes**	Focuses on **patterns**, not necessarily behaviorally meaningful
Transparency	Transparent and **auditable**	Often opaque and **hard to validate**
Flexibility	Limited by **functional form assumptions**	Highly flexible – can model **nonlinearities and interactions**
Use in Economics	Widely used in **transport, health, and environmental economics**	Increasingly used in **complementary roles** (e.g., feature extraction)
McFadden's View	Essential for **scientific modelling** and understanding behavior	Useful for **prediction**, but must be used with caution in economics

Greene, W. H.. and Hensher, D.A. (2026) Hybrid choice modelling and transportation research, *Foundations and Trends in Econometrics,* Now Publishers Inc., Boston (part of Emerald Publishing) 174 pp.

Personalised Travel Behaviour Prediction Using Multi-task Transformers with Mixture-of-ExpertsTransport

David A. Hensher AM, PhD, FASSA, FAITPM, FCILTA, Roads Australia John Shaw Medal
Professor and Founding Director,
Institute of Transport and Logistics Studies,
The University of Sydney Business

https://www.sydney.edu.au/business/about/our-people/academic-staff/david-hensher.html



ITLS

I acknowledge my colleague Dr Haoning Xi who has been instrumental in the research

Multi-task learning in public transport systems

Multi-task learning (MTL) is a machine learning (ML) paradigm that simultaneously trains related tasks using shared representations, 'improving' predictive accuracy and generalisation compared to single-task learning.

Recent developments demonstrate the substantial potential of MTL to improve accuracy and efficiency in the prediction tasks of PT systems – **our key focus.**

For example, Bei et al.(2023) proposed a Multi-task Learning Deep Neural Network (MTLDNN), which maintains shared features while addressing task-specific variations, **to jointly predict travel mode and purpose, significantly outperforming single-task learning.**

Notably, Mixture-of-Experts (MoE) frameworks dynamically select task-specific experts via gating mechanisms, reducing computational complexity.

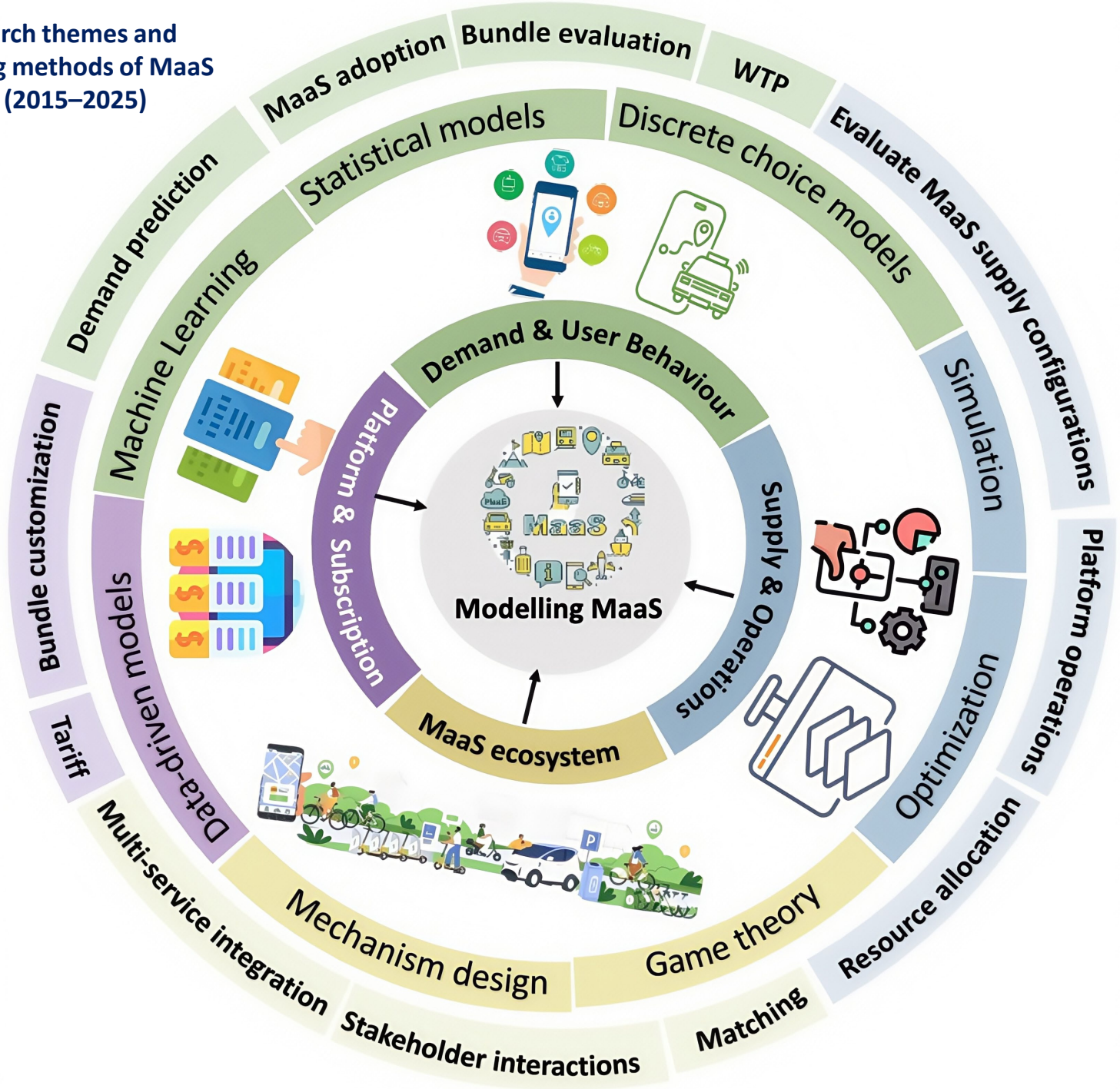
The Multi-gate Mixture-of-Experts (MMoE) framework extends this concept by optimising experts for subtasks, further improving predictive capability. Transformer-based architectures, such as the Transformer-Encoder-based Neural Process (TENP) effectively model complex ***temporal dynamics*** **(the over time evidence and influence).**

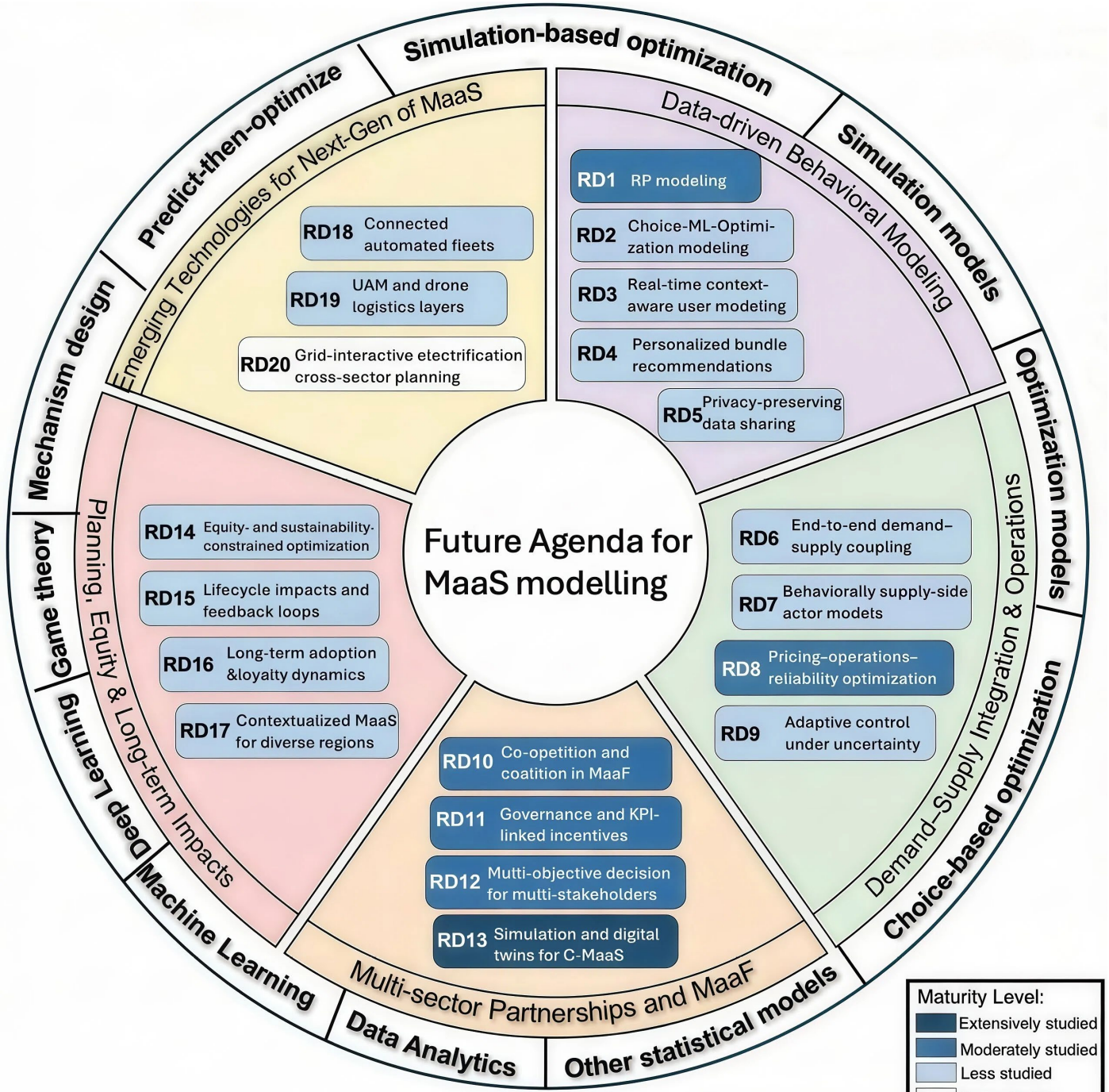
Additional models like the dual information Transformer and the adaptive MultiMode-former (M2-former) have shown significant improvements in multimodal transport predictions.

Shao et al. (2025) proposed a novel spatial-temporal dynamic attention-based state-space model (STDAtt-Mamba) tailored for multi-type passenger demand prediction in multimodal PT systems.

Shao, Z., Xi, H. Hensher, D.A., Wang, Z. and Gao, J. (2025) A spatial-temporal dynamic attention-based Mamba model for multi-type passenger demand prediction in multimodal public transit systems, *Transportation Research Part E,* 202, 104282

Key research themes and modelling methods of MaaS literature (2015–2025)

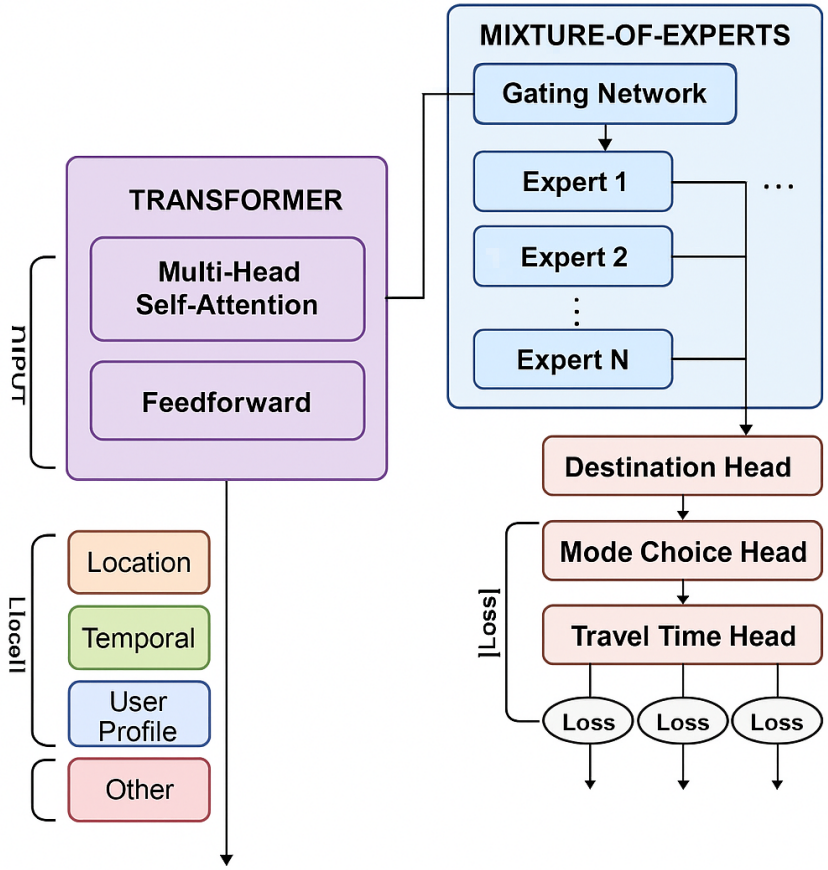



Future Agenda for MaaS modelling
Emerging Technologies for Next-Gen of MaaS
RD18 Connected automated fleets
RD19 UAM and drone logistics layers
RD20 Grid-interactive etectrification cross-sector planning
Data-driven Behavioral Modeling
RD1 RP modeling
RD2 Choice-ML-Optimi-zation modeling
RD3 Real-time context-aware user modeting
RD4 Personalized bundle recommendations
RD5 Privacy-preserving data sharing
Demand–Supply Integration & Operations
RD6 End-to-end demand–supply coupling
RD7 Behaviorally supply-side actor modets
RD8 Pricing–operations–reliability optimization
RD9 Adaptive control under uncertainty
Multi-sector Partnerships and MaaF
RD10 Co-opetition and coatition in MaaF
RD11 Governance and KPI-linked incentives
RD12 Mutti-objective decision for mutti-stakeholders
RD13 Simulation and digital twins for C-MaaS
Planning, Equity & Long-term Impacts
RD14 Equity- and sustainability-constrained optimization
RD15 Lifecycle impacts and feedback loops
RD16 Long-term adoption &loyalty dynamics
RD17 Contextualized MaaS for diverse regions
Predict-then-optimize
Simulation-based optimization
Simulation models
Optimization models
Choice-based optimization
Other statistical models
Data Analytics
Machine Learning
Deep Learning
Game theory
Mechanism design
Maturity Level:
Extensively studied
Moderately studied
Less studied
None studied

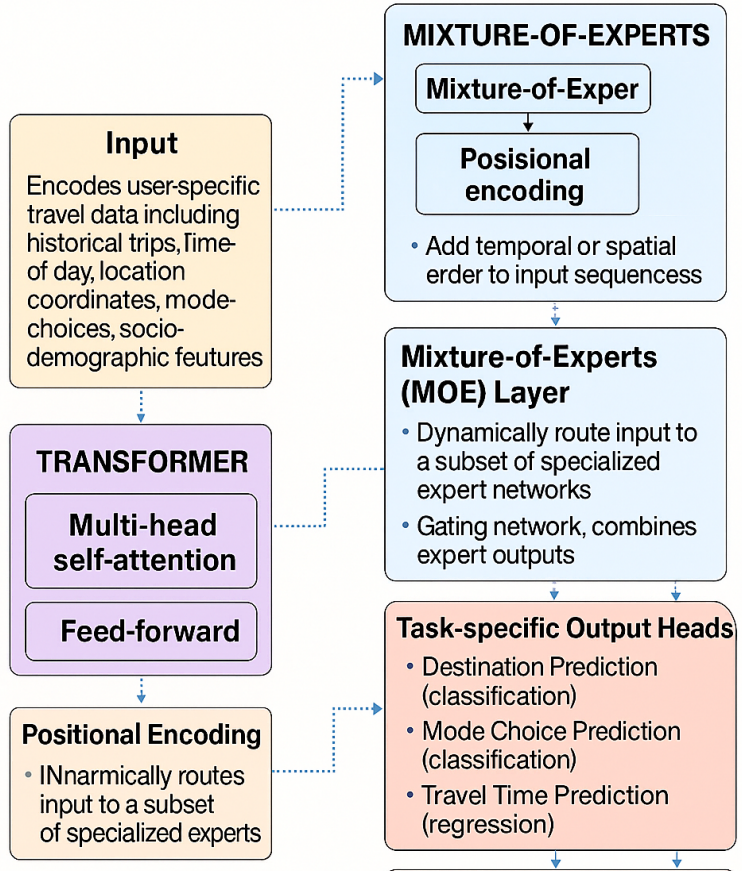
2025 podcast on MaaS and MaaS: https://mobilitaetsfunk.de/multiservice-statt-multimodal-wie-mobilitaetsverhalten-wirklich-veraendert-werden-kann/

2025 see short video on the Car Community Club (CCC) https://mitfahrverband.org/beyond_ride-sharing/

The Architecture overview Summary



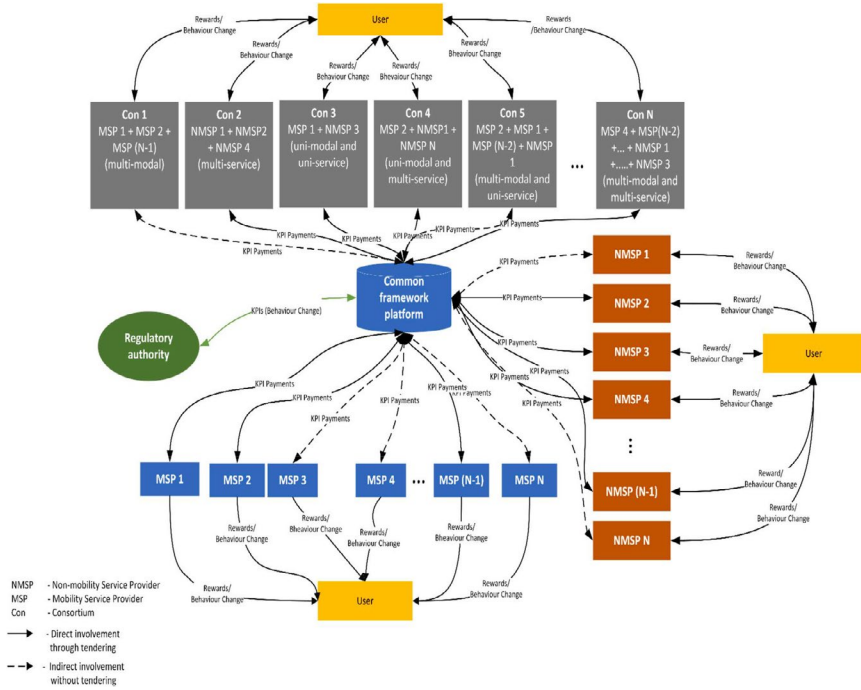
Personalized Travel Behaviour Prediction Using Multi-task Transformers with Mixture-of-Experts

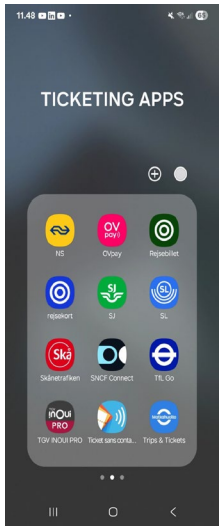


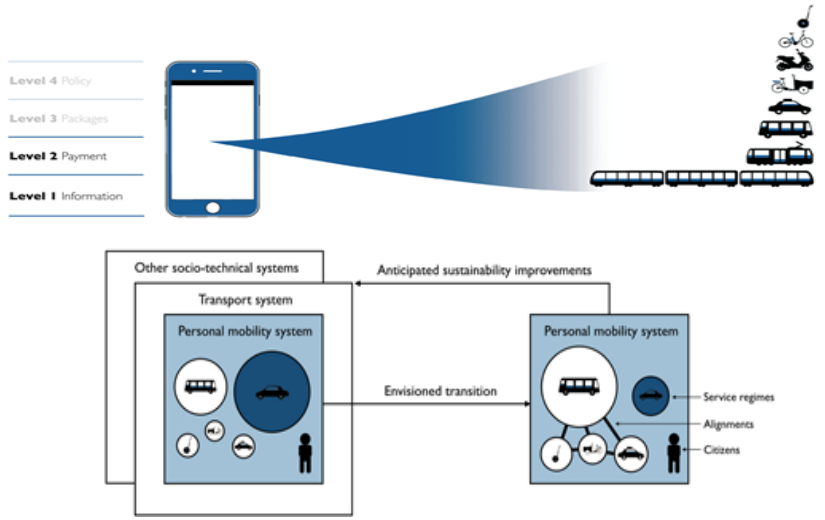
Personalized Travel Behaviour Prediction Using Multi-task Transformers with Mixture-of-Experts

Multimodal transport bundling

- Emerging digital platforms, such as smartphones, smart ticketing, and real-time booking systems enable tailored multimodal transport bundles that integrate various modes into unified service offerings.
- These innovations 'promise' seamless access to diverse transport options, potentially increasing public transit use and reducing private vehicle reliance (Hensher, 2017).
- **There is potential here for MaaS and MaaF.**







Hensher, D.A. and Heitenan, S. (2023) Mobility as a Feature (MaaF): Rethinking the Focus of the second generation of Mobility as a Service (MaaS), *Transport Reviews* DOI: 10.1080/01441647.2022.2159122

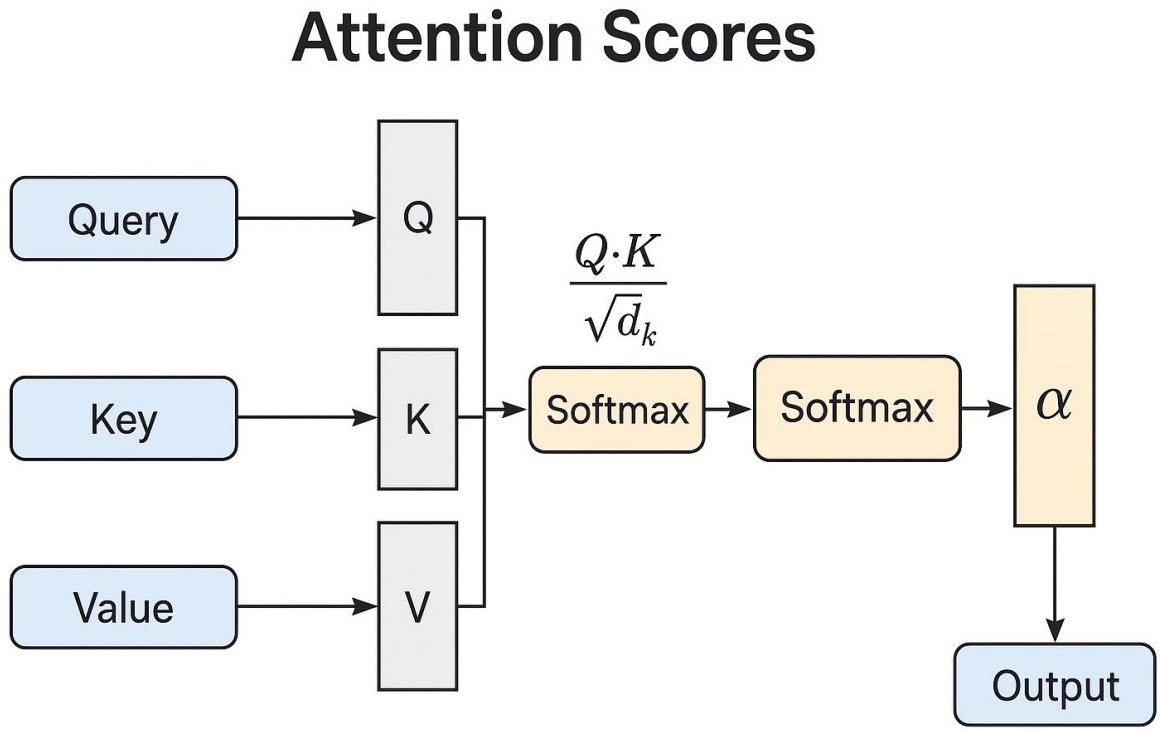
Hensher, D.A., Mulley, C., Ho, C., Nelson, J., Smith, G. and Wong, Y. (2020) *Understanding Mobility as a Service (MaaS) - Past, Present and Future.* Elsevier.

Hensher, D.A., Nelson, J. D., and Mulley, C. (2026) Mobility as a Service: Challenges and Opportunities, for *Handbook on Research Methods in Transport Economics and Policy*, edited by Andrew Smith and Chris Nash, Edward Elgar Publishers.

A key behavioural component: Attention Scores?

- In Transformer models (like BERT, GPT, T5), **attention scores** measure how much focus each token (word or sub-word) should give to other tokens in the input sequence.
- These scores are computed in the **self-attention mechanism**, which allows the model to weigh the importance of different words relative to each other.
- Attention scores tell the model which parts of the input are most relevant.
- They are computed using dot products of query and key vectors, scaled and normalised.
- In multi-task setups, attention can be influenced by task-specific signals to adapt behaviour across tasks.
- **So, what does this mean intuitively in our context?**

https://www.youtube.com/watch?v=bavfa_Rr2f4



They are *learned vector representations* of the input, created through linear transformations.

For an input sequence $X = [x_1, x_2, \ldots, x_n]$,
each token embedding x_i is projected into **three different spaces** using learned weight matrices:
Q= $XW_Q, K = XW_K, V = XW_V$
W_Q, W_K, W_V are trainable matrices. This means **the same input** gets turned into three versions of itself, each with a different purpose.

Another Interpretation:Summary

The attention score (e.g., 2.3) is a pre-softmax logit computed by (scaled) dot-product attention. For token *i* (e.g., Bus) and token j, we project their embeddings using learned matrices to obtain $q_i = x_i W_Q$ and $k_j = x_j W_K$, then compute $e_{ij} = q_i^\top k_j$.

In a Transformer attention block, for a given "query token" *i* (e.g., Bus) and a "key token" *j* (e.g., Bus, Train, Fare), the model first computes:
a Query vector q_i
a Key vector k_j

Then it computes a raw similarity score:

$$e_{ij} = \mathbf{q}_i^\top \mathbf{k}_j$$

This raw similarity e_{ij} is what the slide shows as 2.3 (before softmax).

where d_k is the key/query dimension. Some implementations also add a bias term (e.g., positional bias), but the core source of the number is still $\mathbf{q}_i^\top \mathbf{k}_j$.

Example, suppose the projected vectors are 3-dimensional:

$$\mathbf{q}_{bus} = [1.0,\ 0.5,\ 0.2]$$

$$\mathbf{k}_{bus} = [1.5,\ 1.0,\ 1.5]$$

Then the raw dot product is:

$$\mathbf{q}_{bus}^\top \mathbf{k}_{bus} = 1.0 \cdot 1.5 + 0.5 \cdot 1.0 + 0.2 \cdot 1.5 = 1.5 + 0.5 + 0.3 = 2.3$$

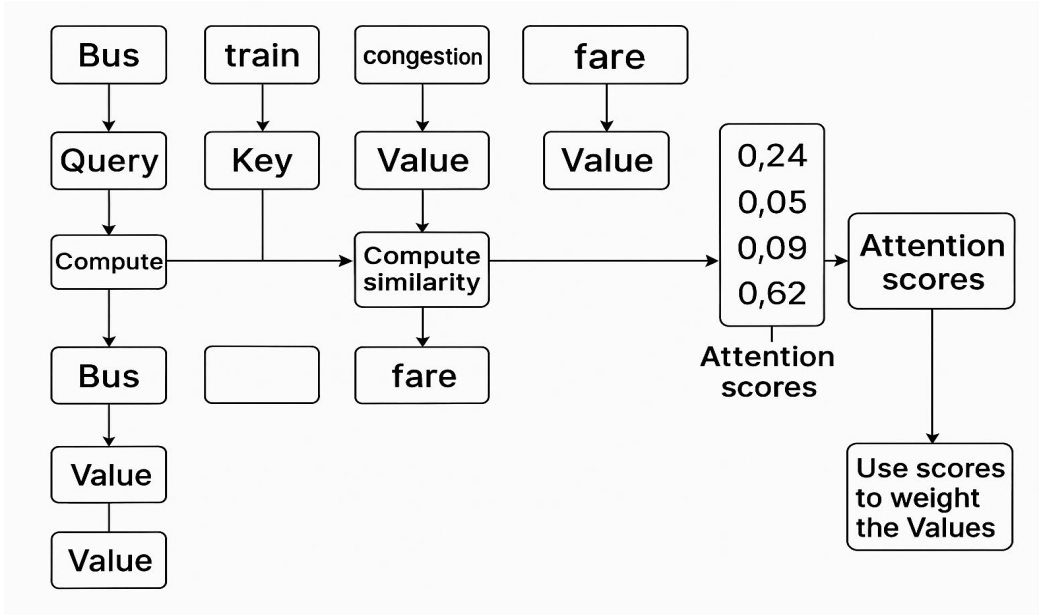
Further, similarity score 2.3 can become an attention weight:

The vector of raw scores for Bus attending to all tokens might be:

$$[e_{bus,bus},\ e_{bus,train},\ e_{bus,congestion},\ e_{bus,fare}]$$

Then attention weights are:

$$\alpha_{bus,j} = \mathrm{softmax}(e_{bus,\cdot})_j = \frac{\exp(e_{bus,j})}{\sum_m \exp(e_{bus,m})}$$

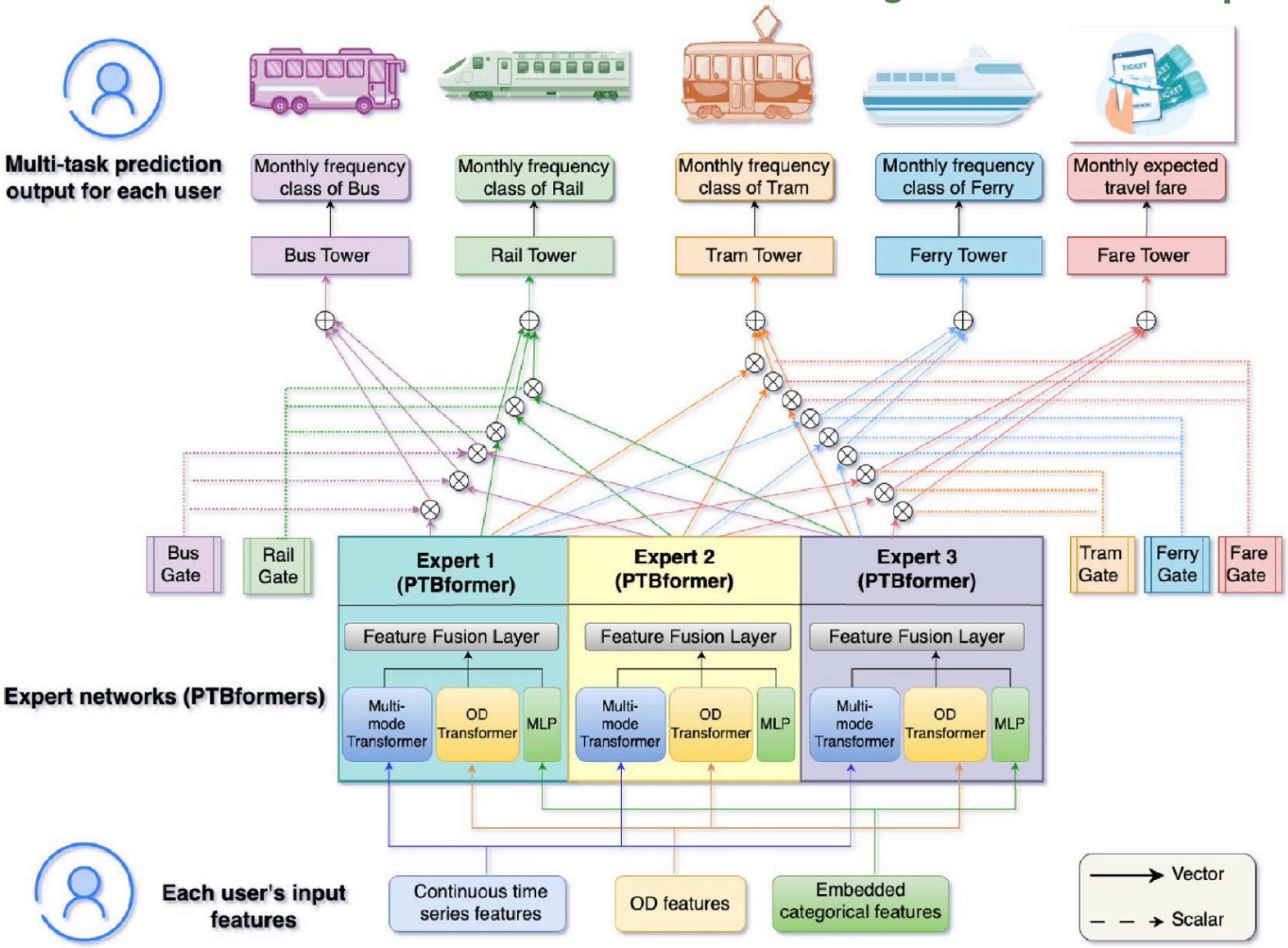


These alphaα values sum to 1 and are what we often visualise.

Finally, the representation for Bus is updated by a weighted sum of value vectors.

A value like 2.3 is simply the dot product between the projected query and key vectors for that sample; it is not manually assigned but results deterministically from the learned parameters and the current input. The model learns W_Q and W_K so that when two signals are predictive together (e.g., Bus and Train), their projected vectors become more aligned in the latent space. That alignment increases $\mathbf{q}^\top \mathbf{k}$, producing a larger score (e.g., 2.3), which after softmax yields a larger weight and therefore a stronger influence of that token's information in the update.

PTBformer-MMoE: PTBformer within the Multi-gate Mixture-of-Experts (MMoE) framework



PTBformer-MMoE tailored for individual travel behaviour predictions

Originally developed for language tasks and widely adopted in large language models, such as generative pre-trained Transformer (chatGPT), Transformers leverage self-attention to effectively integrate diverse user features, including mode-specific, OD-specific, and user attributes.

In this study, we propose a PTBformer-MMoE architecture, a Transformer-based model within MMoE framework, **tailored for individual travel behaviour predictions.**

Each expert network is a PTBformer and the **MMoE framework integrates multiple expert networks into a unified architecture that improves predictive performance for each task**. Gate networks dynamically adjust the weights for each expert network by processing task-specific input features. These weighted combinations of expert outputs are then fed into task-specific tower networks.

Each tower focuses on a particular prediction task, such as monthly travel frequency classification of various modes (e.g., bus, rail, tram, ferry) and expected monthly travel fare.