WM@BOOTH 2026 • Workshop on World Models

Learning to Hear Motion Before Naming It

Katerina Vinciguerra
University of Parma • katerina.vinciguerra@unipr.it

1 THE IDEA: LEARN TO REPRESENT MOTION, THEN TO NAME IT

Motivation:

- An agent perceiving a continuous sensory stream should first represent **how the stream is changing** before naming or describing it.
- Following the principle of **understanding the consequences of an action** [1] by being able to predict the next latent representation of acoustic motion for self-driving perception applications.

Why self-supervised, then supervised: Inspired by human learning abilities (Spelke 1990 [2]; Bertenthal et al. 1984 [3]), we first learn a representation of how the acoustic scene is **changing**, *without labels*, and then learn to **name** the motion.

Implementation: Two-stage approach on 8-channel raw audio of vehicles approaching an occluded junction.

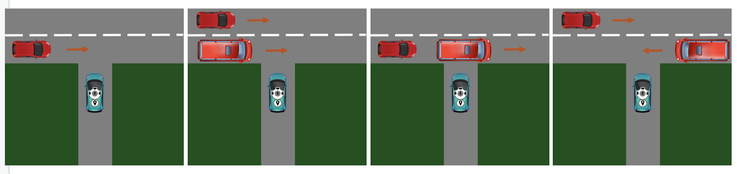
- First stage trains a predictor on the latent representation of multichannel waveforms with no labels.
- Second stage attaches a temporal model with three-task classification heads (**direction**, **count**, and **type** of vehicles being detected).

Contributions:

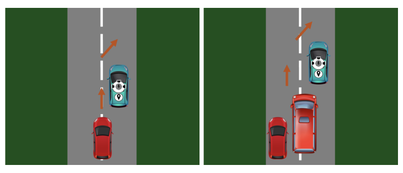
- An architecture that beats state-of-the-art models on our dataset.
- We isolate what **self supervised learning (SSL)** contributes and what **supervised learning** changes.
- The SSL representation jointly organises **motion** and **source identity**.

2 TASK & DATA: DETECT AND TRACK OCCLUDED VEHICLES

Static (T-junction) dataset



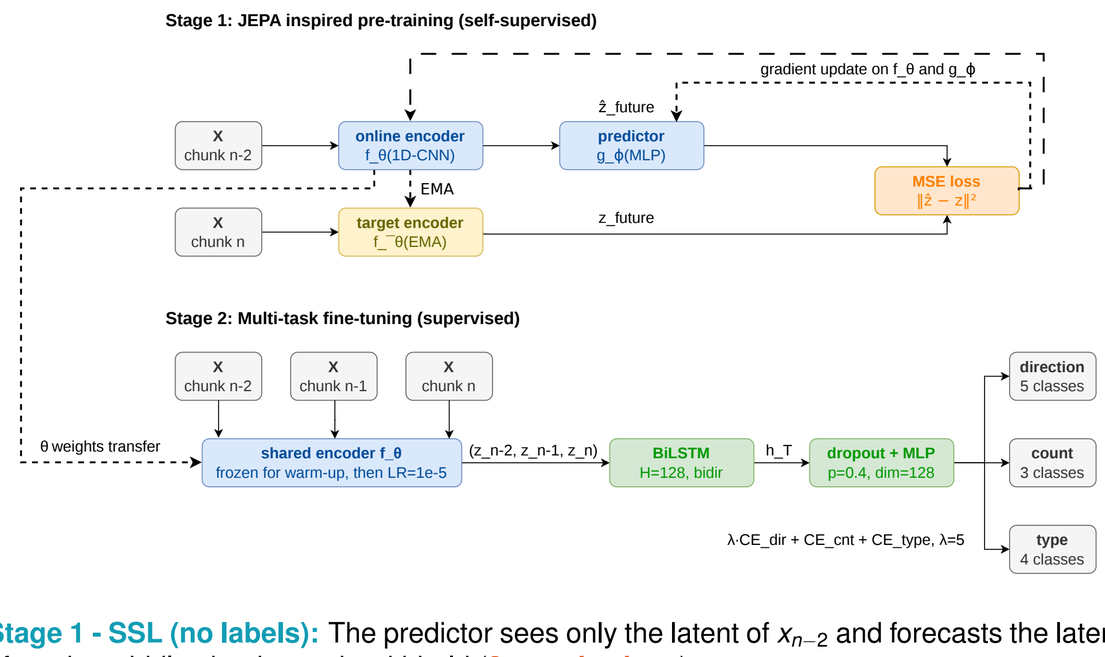
Driving (Take-over) dataset



- The ego-vehicle hears the **direction**, **number** and **type** of vehicles it cannot see around a corner.
- **9,800** labelled 1 s chunks across 6 scenarios and 4 acoustic environments;
- Train and test on the **T-junction** set, with representation transfer to the held-out **Take-over** set.

3 METHOD: 2-STAGE LEARNING

Input: 1 s chunk of 8-channel raw audio, grouped into 3 s windows $W_n = (x_{n-2}, x_{n-1}, x_n)$, $x_k \in \mathbb{R}^{8\times 48000}$.



Stage 1 - SSL (no labels): The predictor sees only the latent of x_{n-2} and forecasts the latent of x_n; the middle chunk x_{n-1} is withheld (**2-step horizon**).

$$\mathcal{L}_{\text{SSL}} = \left\| g_\phi(f_\theta(x_{n-2})) - f_{\bar{\theta}}(x_n) \right\|_2^2, \qquad \bar{\theta} \leftarrow m\,\bar{\theta} + (1-m)\,\theta,\ m{=}0.99.$$

Stage 2 - Supervised (three heads): θ transfers; a BiLSTM aggregates the 3-chunk latent sequence; its final state feeds three linear heads.

$$\mathcal{L} = \lambda\,\mathcal{L}_{\text{CE}}^{\text{dir}} + \mathcal{L}_{\text{CE}}^{\text{cnt}} + \mathcal{L}_{\text{CE}}^{\text{type}}, \qquad \lambda = 5.$$

The 2-step horizon is the key design choice: withholding the intermediate chunk forces forecasting and captures longer-range motion structure.

4 RESULTS: SSL BEATS STATE-OF-THE-ART (SOTA)

The SSL representation reaches multi-task accuracy **above supervised SOTA** - T-junction validation accuracy (%):

Method / Config.	Direction	Count	Type
Related works (single-task / monitoring)			
SRP-PHAT + SVM [4]	53.4	–	–
DOA + TF + CNN [6]	61.9	–	–
Mel spec. + CNN [5]	75.0	–	–
DCASE baseline [7]	67.5	89.5	66.9
Encoder initialization (probe = BiLSTM)			
BiLSTM only (no encoder)	33.7	47.0	40.0
Random + BiLSTM (frozen)	66.3	79.7	76.2
Random + BiLSTM (fine-tuned)	70.4	83.3	80.3
SSL + BiLSTM (frozen)	79.9	94.1	93.2
SSL + BiLSTM (fine-tuned, ours)	**81.8**	**94.5**	**93.7**
Ablation — prediction target			
Interpolation	79.3	89.0	87.3
One-step	80.8	94.1	92.5
Two-step (ours)	**81.8**	**94.5**	**93.7**

- SSL over a random frozen encoder adds **+23.4%, +18.5%, +23%** for direction, count, type respectively; fine-tuning adds $\leq$**2.4%**: the representation is learnt *at the SSL stage*.
- SSL vs. the multi-task baseline: **+21.2%** direction, **+5.6%** count, **+40.1%** type.

5 TRANSFER: FROZEN SSL ADAPTS

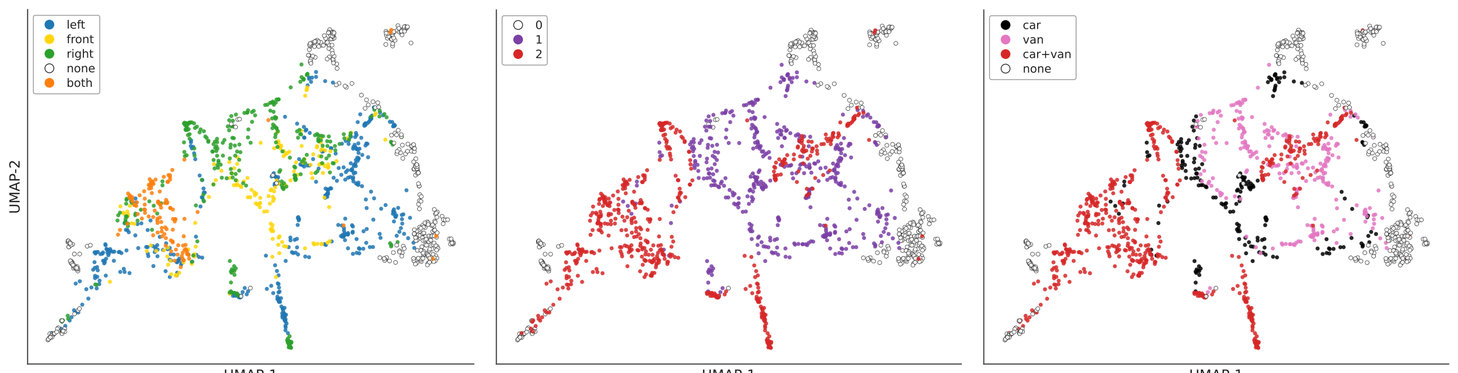
Held-out **driving take-over** set; a 6th direction class **(back)** is added with zero-initialised weights and adapted from only **10** labelled clips, encoder frozen:

10-shot	Direction	Count	Type
Accuracy (%)	70.1 ± 10.1	85.7 ± 3.1	56.8 ± 3.6

Observation:

- Early confusion in type task: ego-motion strengthens the acoustic signature, bringing more confusion between vehicle classification.
- Back $\leftrightarrow$ left boundary is not acoustically grounded - an arbitrary cut.

6 WHAT THE SSL LATENT CONTAINS (NO LABELS USED)



UMAP of frozen-encoder validation latents - **same projection coordinates** across the three panels: **(left)** direction, **(middle)** count, **(right)** type.

- **Silence sits outside the active manifold:** *none*/*0*/*none* share peripheral islands –> stable "no-vehicle" state.
- **Dominant axis = one vs. two vehicles** (count); type-task inherits it (two-vehicle = car+van).
- **Direction regions emerge: left** and **right** on opposite sides; **front** intermixes, encoder's spatial limit.

7 CONCLUSION

- A **JEPA-style latent-prediction** pretext on raw multichannel audio already organises *both* the scene's **motion** and its sources' **acoustic identity** — with *no labels*.
- A small temporal probe reads it out *above* supervised pipelines trained on far larger labelled data.
- The supervised stage adds essentially nothing the pre-training has not already provided: **represent first, name second**.
- *Next:* re-use the predictor g_ϕ at inference to measure how far each future is from expectation — toward a system that reasons about its own anticipations.

REFERENCES

[1] Q. Garrido, T. Nagarajan, B. Terver, N. Ballas, Y. LeCun, M. Rabbat, "Learning latent action world models in the wild," *arXiv:2601.05230*, 2026.
[2] E. S. Spelke, "Principles of object perception," *Cognitive Science*, 14(1):29–56, 1990.
[3] B. I. Bertenthal, D. R. Proffitt, J. E. Cutting, "Infant sensitivity to figural coherence in biomechanical motion," *J. Exp. Child Psychology*, 37(2):213–230, 1984.
[4] Y. Schulz, A. K. Mattar, T. M. Hehn, J. F. P. Kooij, "Hearing what you cannot see: Acoustic vehicle detection around corners," *IEEE RA-L*, 6(2):2587–2594, 2021.
[5] D. Li, "Sound source localization of cars at intersections based on deep learning," *IEEE ICISCAE*, pp. 377–381, 2023.
[6] M. Hao et al., "Acoustic non-line-of-sight vehicle approaching and leaving detection," *IEEE T-ITS*, 25(8):9979–9991, 2024.
[7] S. Damiano, L. Bondi, S. Ghaffarzadegan, A. Guntoro, T. van Waterschoot, "Can synthetic data boost the training of deep acoustic vehicle counting networks?," *IEEE ICASSP*, pp. 631–635, 2024.