Motion as Language: Towards a Situation–Motion Language for Spatio-Temporal Learning

Alejandro Sanchez Guinea **Achref Doula** **Thomas Kreutz**
TU Darmstadt, 4DigitalTwin*
[ale.sanchez.guinea, achref.doula, thomasjonesx22]@gmail.com

Abstract

We argue that progress in spatio-temporal reasoning requires treating motion as language. Existing encodings take different forms: raw coordinates and occupancy grids describe motion metrically or as images, but lack transferable situational meaning and rarely recur across contexts; latent embeddings compress motion histories for prediction, but their semantics are entangled and task-specific. We introduce the idea of a *Situation–Motion Language* (SML), where tokens capture local occupancy situations, sequences form trajectories, and paragraphs describe multi-agent scenes. By collapsing redundant geometry into canonical symbols, SML yields a compact vocabulary that reoccurs across environments, enabling distributional richness akin to natural language. This reframing allows sequence models to learn from "motion text," supports prompting and controllable generation, and offers an interpretable layer that can align with vision–language systems. We argue that a situation-based motion language is a missing modality for scalable and transparent spatio-temporal learning.

1 Introduction & Motivation

Progress in spatio-temporal reasoning requires treating motion as a language problem. Despite advances in vision, language, and multimodal foundation models [Brown et al., 2020, Radford et al., 2021, Alayrac et al., 2022], current encodings remain limited. Raw coordinates, velocity vectors, and occupancy grids describe motion in metric or image-like forms that are precise but lack transferable situational meaning: a coordinate (x, y) is only interpretable relative to a global map, while a raster encodes geometry but not affordances. As a result, these forms provide little recurrence across contexts [Rudenko et al., 2020a, Freksa, 1991a], limiting compositional learning and instruction following. Latent neural embeddings compress motion histories into vectors that are powerful for prediction [Ivanovic and Pavone, 2021a, Giuliari et al., 2021a], but their semantics are entangled and task-specific. Unlike symbolic abstractions, embeddings expose no reusable units or compositional structure [Doshi-Velez and Kim, 2017, Rudin, 2019], making them difficult to prompt, edit, or align with natural language.

We therefore argue for a *Situation–Motion Language* (SML): a symbolic modality where tokens encode local occupancy situations, sequences represent trajectories, and paragraphs capture multi-agent scenes. By collapsing redundant geometry into canonical symbols, SML produces a compact vocabulary that reappears across environments, creating the distributional richness that enables language models to scale. This reframing supports prompting and controllable generation, improves interpretability, and provides a natural alignment layer with vision–language systems. Our position is that a situation-based motion language is not optional but necessary: just as pixels underlie vision

*https://www.4digitaltwin.com/

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: SPACE in Vision, Language, and Embodied AI (SpaVLE).

and words underlie language, SML or an equivalent situational modality may underlie scalable and transparent spatio-temporal learning.

2 The Situation–Motion Language (SML)

We formally define the Situation–Motion Language (SML) as a symbolic representation of spatio-temporal behavior. It is agent-centric and compositional, operating at three levels:

- **Words**: symbolic tokens encoding the local spatial situation of an agent at a given timestep.
- **Sentences**: ordered sequences of words representing an agent's trajectory.
- **Paragraphs**: sets of aligned sentences, one per agent, representing a multi-agent scene.

2.1 Word Construction via Sector Collapse

Let $a_t \in \mathbb{Z}^2$ denote the position of a focal agent at time t on a discrete grid. We partition the plane into M angular *sectors* (e.g., $M = 8$ for the compass bearings N, NE, E, SE, S, SW, W, NW). For each sector $d \in D$, we define the discrete ray

$$r_d = \{a_t + k \cdot d \mid 1 \leq k \leq R\},$$

where k is the offset from the agent in direction d, and R is the maximum radius considered. Each sector is then *collapsed* to its nearest occupied cell:

$$k^* = \min\{k \mid a_t + k \cdot d \text{ is occupied}\}.$$

If such a k^* exists, we add the symbolic token k^*d to the word; otherwise, the sector contributes $\emptyset$. Here k^*d is a *symbol*, not a product: it denotes "blocked at distance k^* along direction d." For compactness, $1d$ is written simply as d (e.g., $1n \mapsto n$). The resulting word w_t is the set of all such tokens across sectors. We refer to this reduction as *sector collapse*: per angular sector, only the nearest blocker is retained while all geometry behind it is discarded. This abstraction drops fine-grained detail but preserves the motion-relevant constraint that each sector is either free up to distance k^* or blocked at that distance.

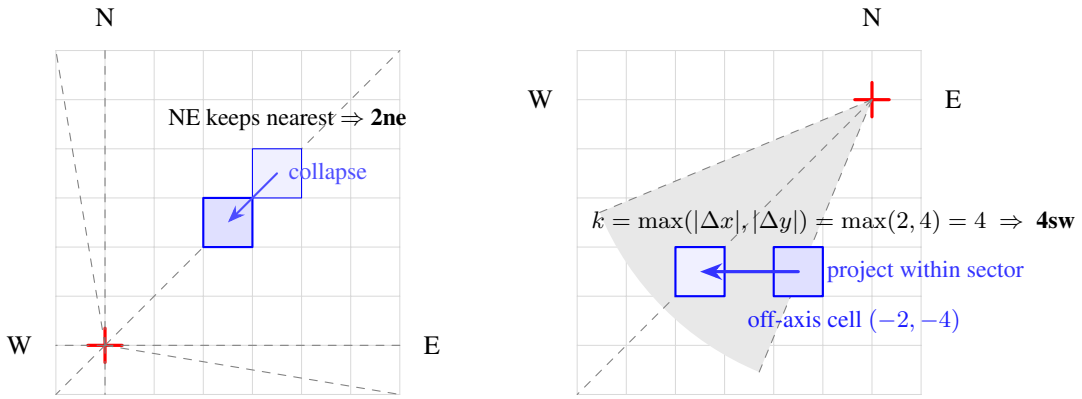
This design mirrors long-standing approaches in robotics such as the Vector Field Histogram and polar occupancy grids [Borenstein and Koren, 1991, Ulrich and Borenstein, 1998, 2000, Geng and Langelaan, 2017], where each sector retains only its nearest obstacle. Our reinterpretation differs in purpose: rather than enabling real-time control, sector collapse builds a symbolic vocabulary of reusable motion situations. This avoids the combinatorial sparsity of full-neighborhood encodings: a radius-R neighborhood already contains $(2R+1)^2$ cells (e.g., ~ 440 for $R = 12$), implying $2^{(2R+1)^2}$ possible patterns with almost no recurrence. By contrast, sector collapse yields only $M \times R$ canonical tokens, a compact linear vocabulary with high recurrence across environments. Crucially, by sparsity we mean not the number of slots per word but the *statistical sparsity of the vocabulary*: high-resolution rays (e.g., 64 thin lines) produce highly specific tokens that rarely repeat, whereas coarser sectors merge offsets into frequently recurring tokens, enabling distributionally dense statistics that support language-model-style learning.

2.2 Example

Before the example, we emphasize that SML is a *symbolic* language: tokens here represent motion situations, not natural-language words. For this reason, repetitions such as "$\cdots$" (stalling in place) carry semantic meaning in motion modeling, even though they would appear unusual in natural language.

Figure 1 shows a focal agent (red cross) surrounded by obstacles (blue squares) under an 8-sector scheme.[2] Following §2.1, a cell at offset $(\Delta x, \Delta y)$ is assigned to sector $d = (\mathrm{sgn}(\Delta x), \mathrm{sgn}(\Delta y))$ with radial index $k = \max(|\Delta x|, |\Delta y|)$, yielding the token kd. Along the northeast sector, both cells

[2]The same construction extends to M sectors by partitioning the plane into M equal angular slices (nearest by angle via $\mathrm{atan2}$) and using the sector's unit label d. We keep the Chebyshev (L_∞) index $k = \max(|\Delta x|, |\Delta y|)$ as the radial token. Increasing M (e.g., 16, 32, 64) reduces directional quantization while growing the symbolic space; in practice, M is a hyperparameter trading geometric fidelity for recurrence and sample efficiency.



(a) Nearest-obstacle *sector* collapse (NE)

(b) Sector assignment & projection (SW)

Figure 1: **Sector collapse** (8-sector case). Each compass direction represents an angular *sector*, not an infinitesimally thin line. **(a)** Within a sector (NE), only the nearest occupied cell is kept (3ne collapses to 2ne). **(b)** Off-axis cells inside the SW sector (e.g., $(-2, -4)$) are projected to the sector's canonical token using the Chebyshev index $k = \max(|\Delta x|, |\Delta y|)$, yielding **4sw**.

$(2, 2)$ and $(3, 3)$ lie on the same ray; only the nearest ($(2, 2)$) survives, giving the token $2\mathrm{ne}$. In the southwest sector, the cell at $(-4, -4)$ lies on the SW diagonal and contributes $4\mathrm{sw}$. An off-axis cell at $(-2, -4)$ (i.e., $4s2w$) is projected onto the southwest sector with $k = \max(2, 4) = 4$, which also maps to the token $4\mathrm{sw}$. Empty sectors contribute nothing to the word. After collapsing each sector to at most one token and ordering directions in a fixed sequence (N, NE, E, SE, S, SW, W, NW), the canonical word for this timestep is

$$w_t = \{\, 2\mathrm{ne},\ 4\mathrm{sw} \,\}.$$

Over time, an agent's trajectory is represented as a sentence $S = (w_t, w_{t+1}, \ldots, w_{t+T-1})$, while multi-agent interactions form a paragraph $P = \{S^{(1)}, \ldots, S^{(N)}\}$. Words act like morphemes or phoneme-like units of motion: individually limited in meaning, but gaining semantics and functional richness through sequence and composition. This sector-based abstraction thus trades exact geometry for a compact, recurring vocabulary of motion situations, allowing tokens to reappear across agents and environments and enabling distributional learning in the spirit of natural language models.

3 Conceptual Insights

The Situation–Motion Language (SML) embodies several insights that distinguish it from prior representations. **Directionality as sequence.** Motion direction is not explicitly encoded as a vector but emerges from the order of words in a sentence, making trajectories naturally suited to sequence models [Giuliari et al., 2021b, Yuan et al., 2021]. **Abstract vs. instantiated.** Each word is defined relative to the focal agent, supporting coordinate-invariant generalization [Freksa, 1991b, Cui and Cohn, 1991] while still permitting instantiation on maps for planning or evaluation [Rudenko et al., 2020b, Mo et al., 2022]. **Start, middle, and end semantics.** The first and last tokens capture intent, while the middle reflects situational reactivity, echoing ideas in goal-conditioned trajectory models [Salzmann et al., 2020, Ivanovic and Pavone, 2021b]. **Distributional richness.** As in natural language, where words *reoccur* across many contexts, SML words *reappear* across diverse trajectories and environments. This repeated appearance creates dense co-occurrence statistics that models can exploit to infer affordances and long-horizon associations, an advantage unavailable to raw coordinates, which rarely repeat exactly. **Pacing vs. content.** Temporal pacing can be controlled orthogonally to spatial content via *tempo markers* (speed-up/slow-down) and an optional *stall token* $\cdot$ that repeats the current situation without spatial change. This separates *what* surrounds the agent from *how fast* it evolves, improving controllability and analysis. Markers act as metadata: they do not alter the canonical SML vocabulary and can be ignored during training or used at decoding time for controllable timing (see §2.2 for a short clarification).

Table 1: Conceptual comparison of SML with existing paradigms.

	Coordinates	Latent Embeddings	Occupancy Grids	SML
Unit-level interpretability	Low	Very low	Medium	High
Transferability	Limited	Limited	Moderate	High
Data efficiency	Low	Moderate	Moderate	High
Integration with LMs	Weak	Weak	Weak	Natural

Together, these insights frame SML as a symbolic layer that captures purpose, constraint, and compositionality in motion, and suggest extensions such as enriching the vocabulary with semantic tokens (e.g., `wall`, `vehicle`) for higher-level reasoning.

4 Architectures & Use Cases

SML's symbolic structure makes sequence models natural and suggests inductive biases. **SML-Biased Transformer.** A Transformer over SML words can place global attention on start/end tokens (intent) and local windows in between (situational coherence), mirroring trajectory transformers [Giuliari et al., 2021b, Yuan et al., 2021] but over symbolic rather than metric inputs. **SML-Intent2Trajectory.** An encoder–decoder can generate trajectories autoregressively from start/goal tokens, echoing goal-conditioned planning [Salzmann et al., 2020, Ivanovic and Pavone, 2021b], with diversity from stochastic decoding or prompts. **SML-LatentPlanner.** A global latent z can summarize high-level intent (route class, avoidance mode, tempo), enabling controlled multimodality and planning-time search. In short: Intent2Trajectory *conditions*; LatentPlanner *plans via a latent*. Beyond these sketches, SML applies broadly: replacing raw coordinates with SML tokens may address issues of interpretability and data efficiency noted in [Rudenko et al., 2020b]; in *planning/simulation*, sentences can be composed or edited for controllability; and in *instruction following*, SML aligns with natural language to ground embodied agents [Anderson et al., 2018, Shridhar et al., 2020].

5 Discussion & Outlook

Existing paradigms for trajectory modeling fall into three categories: coordinates, latent embeddings, and occupancy grids. SML contributes explicit invariances (agent-centric, ray canonicalization) and a recurrent symbolic vocabulary that reappears across scenes. Table 1 summarizes the contrasts. Coordinates are precise but context-bound: a tuple (x, y) has meaning only relative to a map and rarely reappears across datasets. Embeddings compress histories into fixed-size vectors, but their axes are entangled, so they lack unit-level semantics and demand large-scale pretraining to generalize. Occupancy grids are somewhat more interpretable since they visualize local geometry, yet they are tied to arbitrary grid resolution and introduce large, mostly empty inputs that reduce efficiency. By contrast, SML tokens directly encode *situational units* such as "blocked north," which are discrete, human-readable, and composable. These units recur across agents and environments, allowing models to exploit dense co-occurrence statistics. In this sense, Table 1 reflects not just a conceptual distinction but a practical one: SML offers tokens that are simultaneously compact, interpretable, and transferable, while existing paradigms typically satisfy at most one of these criteria. Unlike recent LLM–prompt baselines [Xue et al., 2022, Lan et al., 2024, Haydari et al., 2024], which textualize coordinates, SML offers a principled symbolic interface: short geometry-aware tokens that enforce nearest-obstacle semantics and support controllability and planner integration.

Evaluation can derive SML sequences from datasets such as ETH/UCY or TrajNet++ and compare to coordinate-, grid-, and embedding-based models. Metrics should probe predictive accuracy, sample efficiency, and interpretability (e.g., mutual information between tokens and outcomes, or counterfactual editing).

Looking forward, SML may serve as a symbolic motion modality that integrates with vision–language models, enriching multimodal agents with a structured interface for motion. Extending the vocabulary with semantic tokens (e.g., `wall`, `vehicle`) could bridge geometry and semantics. More broadly, symbolic motion is not optional but necessary: just as pixels underlie vision and words underlie language, SML or an equivalent scheme may underlie spatio-temporal foundation models. Beyond

trajectories, it points to motion itself as a new modality: a building block we must recognize today for the next generation of multimodal AI.

References

Jean-Baptiste Alayrac et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Peter Anderson, Qi Wu, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018.

Johann Borenstein and Yoram Koren. The vector field histogram – fast obstacle avoidance for mobile robots. *IEEE Transactions on Robotics and Automation*, 7(3):278–288, 1991. doi: 10.1109/70.88137.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Zhishi Cui and Anthony G. Cohn. Qualitative simulation of large-scale space. In *Proceedings of AAAI*, pages 679–684. AAAI, 1991.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. URL `https://arxiv.org/abs/1702.08608`.

Christian Freksa. Qualitative spatial reasoning. In *Cognitive and Linguistic Aspects of Geographic Space*, pages 361–372. Springer, 1991a. doi: 10.1007/978-94-011-2606-9_17.

Christian Freksa. Qualitative spatial reasoning. In *Cognitive and Linguistic Aspects of Geographic Space*, pages 361–372. Springer, 1991b.

Shuo Geng and J. Willem Langelaan. Polar histogram based path planning for small fixed-wing uavs in complex environments. In *AIAA Information Systems-AIAA Infotech @ Aerospace*, pages 1–10, 2017. doi: 10.2514/6.2017-0611.

Francesco Giuliari, Armin Hasan, Marco Cristani, and Alessandro Del Bue. Transformer networks for trajectory forecasting. *Pattern Recognition*, 120:108135, 2021a. doi: 10.1016/j.patcog.2021.108135.

Francesco Giuliari, Armin Hasan, et al. Transformer networks for trajectory forecasting. *Pattern Recognition*, 2021b.

A. Haydari et al. Mobilitygpt: Enhanced human mobility modeling with a foundation model. arXiv:2402.03264, 2024. URL `https://arxiv.org/abs/2402.03264`.

Boris Ivanovic and Marco Pavone. Heterogeneous-agent trajectory forecasting incorporating class-specific behavior. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2053–2060, 2021a. doi: 10.1109/IROS51168.2021.9636203.

Boris Ivanovic and Marco Pavone. Heterogeneous-agent trajectory forecasting incorporating class-specific behavior. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021b.

Z. Lan et al. Traj-llm: A new exploration for empowering trajectory prediction with llms. arXiv:2405.04909, 2024. URL `https://arxiv.org/abs/2405.04909`.

Kaixin Mo et al. Recap: Data-efficient learning for human trajectory prediction. In *ECCV*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

Andrey Rudenko, Luigi Palmieri, Matthias Herman, Kris M. Kitani, Dariu M. Gavrila, and Kai O. Arras. Human motion trajectory prediction: A survey. *International Journal of Robotics Research*, 39(8):895–935, 2020a. doi: 10.1177/0278364920917446.

Andrey Rudenko, Luigi Palmieri, Matthias Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *International Journal of Robotics Research (IJRR)*, 39(8):895–935, 2020b.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Tim Salzmann, Boris Ivanovic, Pratik Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision (ECCV)*, pages 683–700. Springer, 2020.

Mohit Shridhar et al. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *CVPR*, 2020.

Iwan Ulrich and Johann Borenstein. Vfh+: Reliable obstacle avoidance for fast mobile robots. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1572–1577, 1998. doi: 10.1109/ROBOT.1998.677362.

Iwan Ulrich and Johann Borenstein. Vfh*: Local obstacle avoidance with look-ahead verification. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2505–2511, 2000. doi: 10.1109/ROBOT.2000.846405.

Hao Xue, Bhanu Prakash Voutharoja, and Flora D. Salim. Leveraging language foundation models for human mobility forecasting. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM) Workshops*, 2022. URL `https://arxiv.org/abs/2209.05479`.

Ye Yuan, Xin Weng, and Kris Kitani. Agentformer: Agent-aware transformers for social trajectory prediction. In *ICCV*, 2021.