

# Self-Supervised Multi-Modal World Model with 4D Space-Time Embedding

Lance Legel <sup>\*,1,2</sup> Qin Huang <sup>3</sup> Brandon Voelker <sup>4</sup> Daniel Neamati <sup>5</sup> Jeff Rose <sup>6</sup> Patrick Alan Johnson <sup>7</sup>  
Favyen Bastani <sup>7</sup> James Hennessy <sup>8</sup> Robert Guralnick <sup>2</sup> Douglas Soltis <sup>2</sup> Pamela Soltis <sup>2</sup> Shaowen Wang <sup>9,10</sup>

<sup>1</sup> Ecological Intelligence, Inc. <sup>2</sup> University of Florida <sup>3</sup> Arizona State University <sup>4</sup> University of Houston <sup>5</sup> Stanford University <sup>6</sup> SpatialLogic.com  
<sup>7</sup> Allen Institute for AI <sup>8</sup> Georgia Institute of Technology <sup>9</sup> University of Illinois Urbana-Champaign <sup>10</sup> NSF I-GUIDE \* lance@ecodash.ai



## DeepEarth Architecture

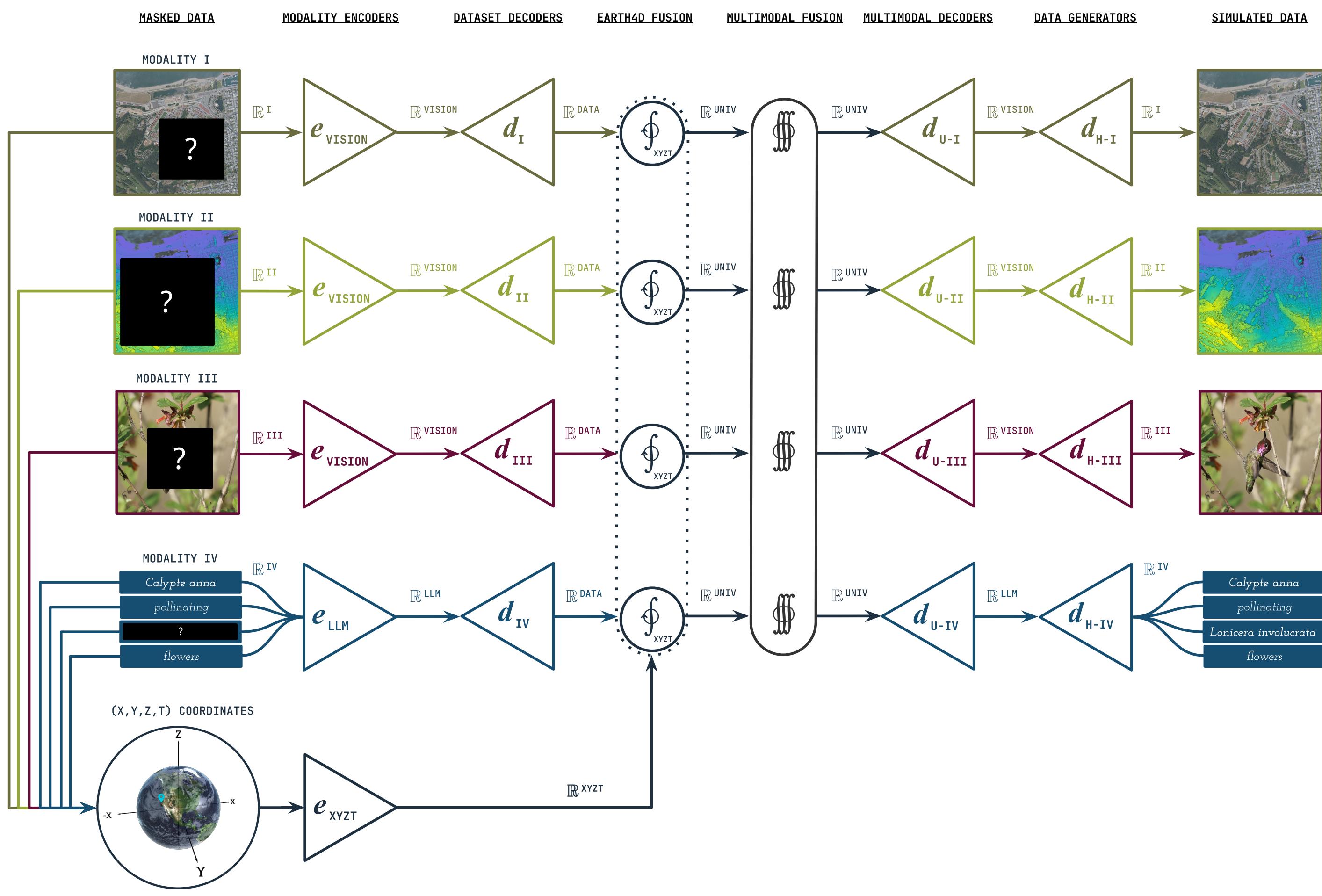


Figure 1. Self-supervised multi-modal world model for planetary science, simulation, & planning.

- **Earth4D** encodes  $(x, y, z, t)$  coordinates into learnable space-time embeddings
- DeepEarth fuses multi-modal (e.g. vision-language) and space-time embeddings
- Model trains by masked reconstruction of multi-modal data across space-time

## Key Contributions

1. **Space-Time Positional Encoder:** Unify all kinds of physical data modalities
2. **Multi-Scale 4D Geospatial Simulator:** Bridge planetary-to-cellular dynamics
3. **4D Learned Hash Probing:** Differentially map  $(x,y,z,t) \rightarrow$  embedding indices

## State-of-the-Art Ecological Prediction Benchmark

Live Fuel Moisture Content (LFMC) measures vegetation water for wildfire risk.  
**Training:** Earth4D encodes  $(x, y, z, t)$  coordinates of LFMC metrics to 192D, concatenates with learnable plant species embeddings, MLP predicts LFMC.  
**Result:** Earth4D outperforms pre-trained Vision Transformer with less input data.

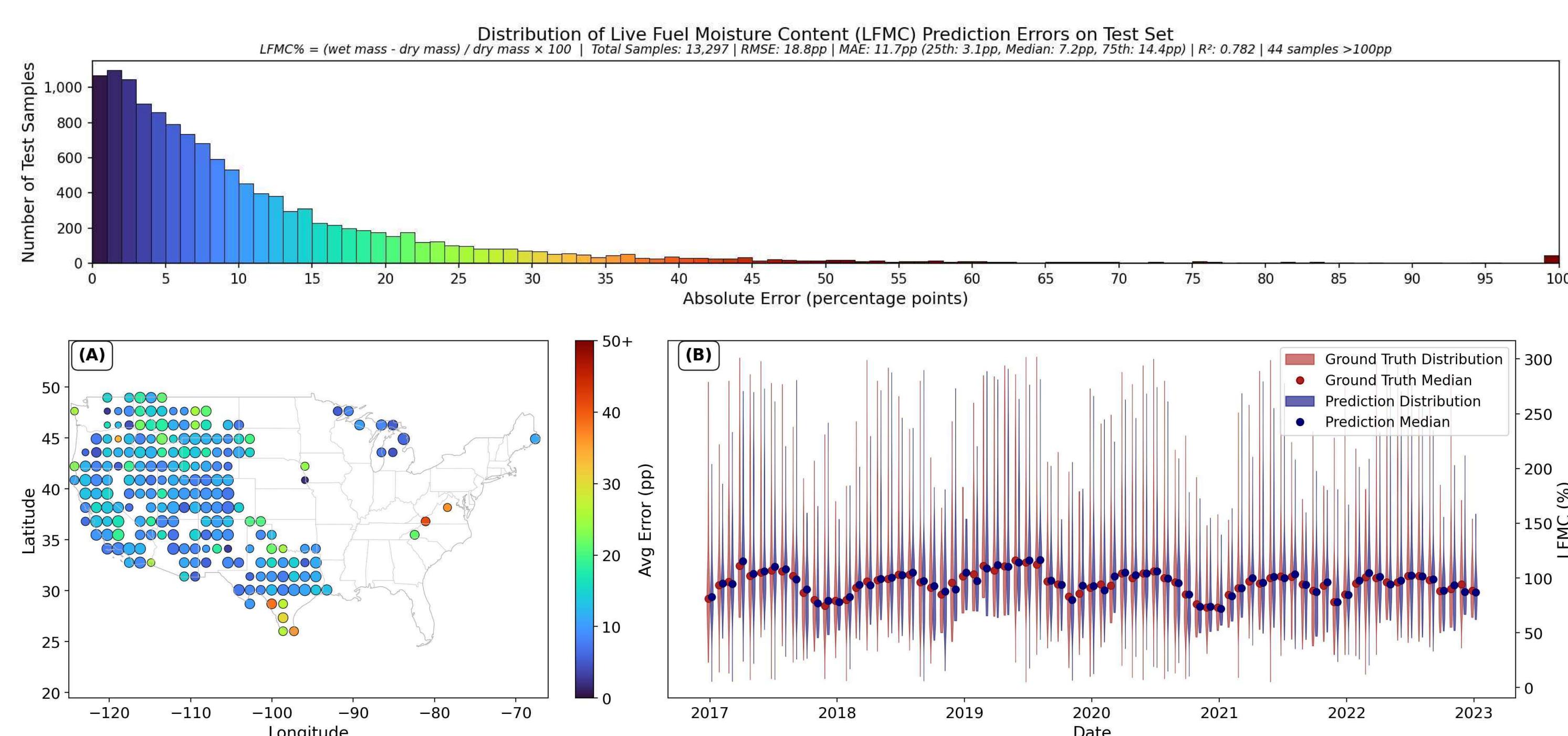


Figure 4. Earth4D LFMC test predictions on Allen Institute for AI (AI2) Globe-LFMC 2.0 benchmark.

Model	Data Inputs	MAE	RMSE	R <sup>2</sup>
AI2 (Pre-trained ViT)	$(x,y,z,t)$ + Species + Vision (Remote Sensing)	12.6pp	18.9pp	0.72
<b>Earth4D</b>	$(x,y,z,t)$ + Species	<b>11.7pp</b>	<b>18.7pp</b>	<b>0.783</b>

**Key Finding:** Earth4D surpasses AI2's foundation model without satellite, weather, or topography data. Only coordinates and species names are used.

## Earth4D Space-Time Positional Encoding

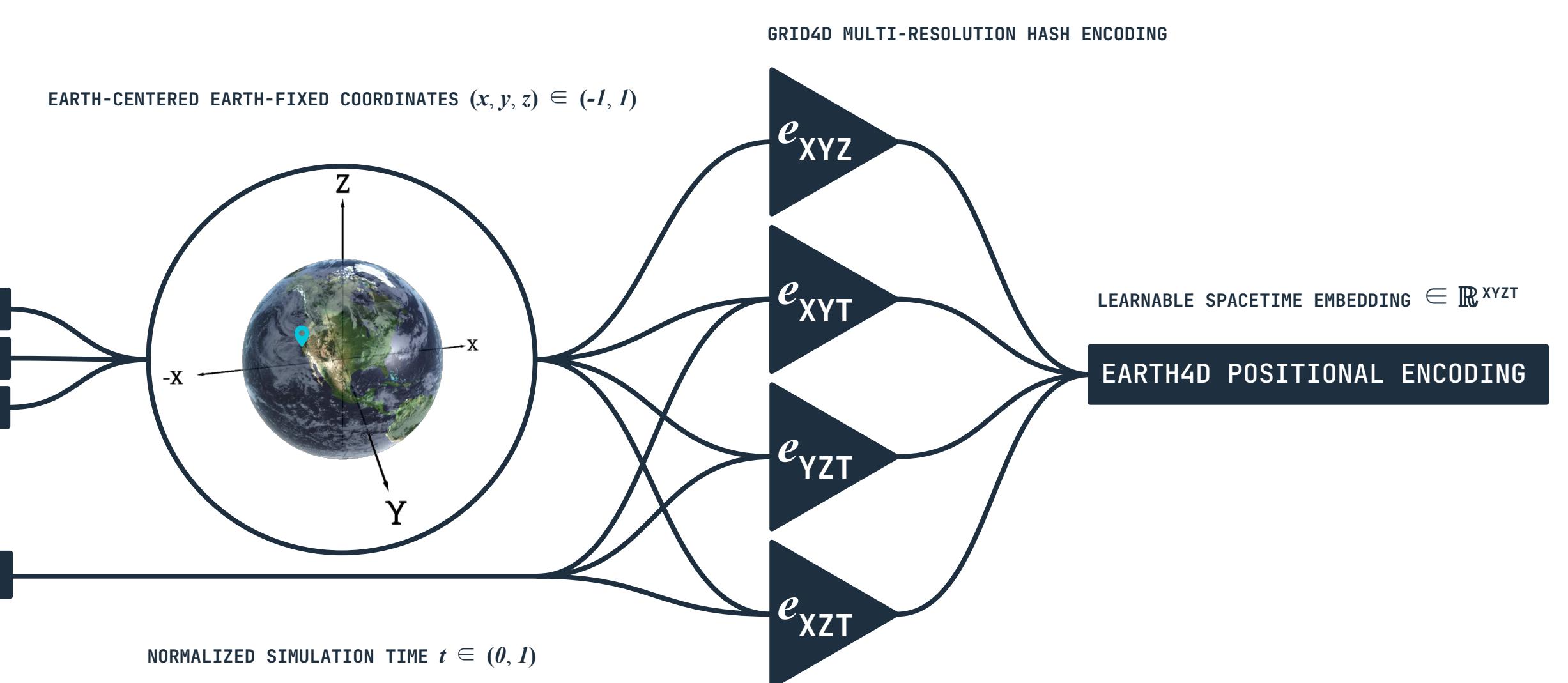


Figure 2. Multi-resolution hash encoding extended to 4D space-time across the planet over years. Earth4D: **multi-resolution hash encoder** with four parallel 3D grids. **Spatial** ( $xyz$ ): static structure. **Spatio-temporal** ( $yzt$ ,  $xzt$ ): dynamics. **Geographic:** Maps ( $latitude$ ,  $longitude$ ,  $elevation$ ,  $time$ )  $\rightarrow (x, y, z, t)$

## Joint Embeddings Across Spatio-Temporal Scales

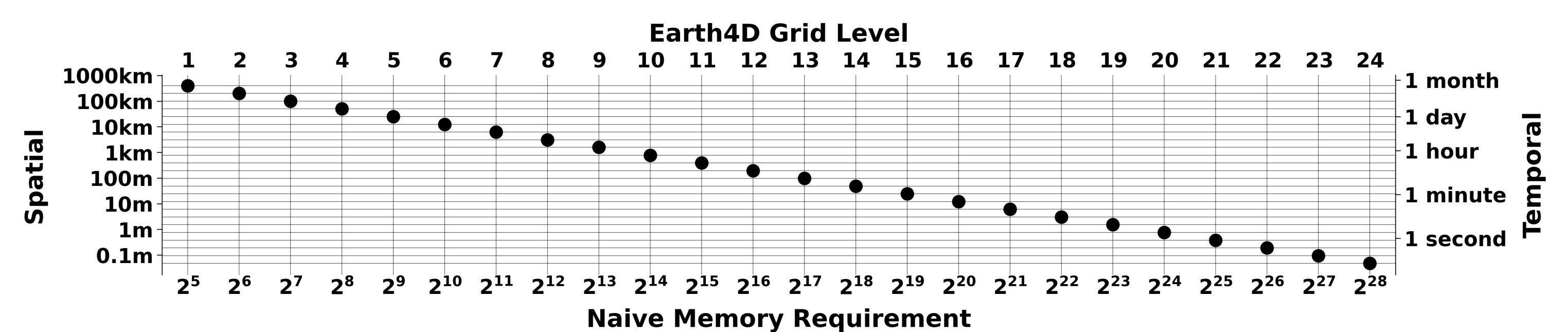


Figure 3. 24 resolution levels per grid ( $xyz$ ,  $yzt$ ,  $xzt$ ,  $xyz$ ), up to  $2^{22}$  entries each. Output: 192D trainable embedding per  $(x,y,z,t)$  coordinate.

## Learned Hash Probing

Hash encoding compresses features into fixed memory, but collisions hurt accuracy. **Learned hash probing** optimizes memory allocation end-to-end.

**Performance:** 33% fewer collisions; MAE improved 27%; R<sup>2</sup> improved 30%.

## Code Demo: Space-Time Positional Encoding

```
# https://github.com/legel/deepearth
from deepearth.encoders.xyzt.earth4d import Earth4D
world_model = Earth4D()
embeddings = world_model(
    # Bletchley Park (Turing breaks Enigma, 1941)
    (51.9976, -0.7416, 110, "1941-06-01 09:00 GMT"),
    # Carnegie Mellon (Hinton invents Boltzmann Machines, 1985)
    (40.4433, -79.9436, 270, "1985-01-15 10:00 ET"),
    # CERN (Berners-Lee invents WWW, 1989)
    (46.2330, 6.0557, 430, "1989-03-12 10:00 CET"),
    # Mila, Quebec (World Modeling Workshop 2026)
    (45.5308, -73.6128, 63, "2026-02-04 11:00 ET"),
)
# embeddings.shape: [4, 192] -- trainable space-time features
```

## References

1. Müller et al. "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding." ACM SIGGRAPH, 2022.
2. Takikawa et al. "Compact Neural Graphics Primitives with Learned Hash Probing." SIGGRAPH Asia, 2023.
3. Xu et al. "Grid4D: 4D Decomposed Hash Encoding for High-Fidelity Dynamic Gaussian Splatting." NeurIPS, 2024.
4. Yebra et al. "Globe-LFMC 2.0, an enhanced and updated dataset for live fuel moisture content research." Scientific Data, 2024.
5. Tseng et al. "Galileo: Learning Global and Local Features of Many Remote Sensing Modalities." ICML, 2025.
6. Johnson et al. "High-Resolution LFMC Maps for Wildfire Risk From Multimodal Earth Observation Data." PMLR, 2025.