OmniTraj: Pre-Training on Heterogeneous Data for Adaptive and Zero-Shot Human Trajectory Prediction

Yang Gao, Po-Chien Luan, Kaouther Messaoud, Lan Feng, Alexandre Alahi
Visual Intelligence for Transportation (VITA) laboratory
EPFL, Switzerland
{firstname.lastname}@epfl.ch

Abstract

While large-scale pre-training has advanced human trajectory forecasting, achieving robust zero-shot generalization across diverse datasets remains a critical challenge. Existing models struggle when encountering heterogeneous sensor configurations, such as varying frame rates and observation horizons. In this work, we revisit zero-shot trajectory prediction from the perspective of distribution shifts and distinguish three transfer settings: temporal transfer, scene transfer, and joint scene–temporal transfer. Through systematic experiments, we show that temporal mismatch is a key source of failure in current pre-trained models. By isolating temporal configuration from dataset shift, we demonstrate that explicitly conditioning on temporal metadata provides a simple and highly effective solution. Building on this insight, we propose OmniTraj, a Transformer-based framework pre-trained on large-scale heterogeneous data with explicit temporal-aware design. OmniTraj achieves state-of-the-art zero-shot performance under joint scene–temporal transfer, reducing prediction error by over 70%. Furthermore, it exhibits exceptional robustness in safety-critical edge cases with severely limited observations and maintains high few-shot data efficiency, paving the way for scalable, dataset-agnostic deployment in real-world autonomous systems.

1. Introduction

Anticipating the trajectories of vulnerable road users (VRUs) is a critical capability for autonomous systems. While large-scale representation learning and pre-training paradigms [3] have advanced the field, deploying these models in unseen environments remains a major challenge. A primary bottleneck for true zero-shot generalization is the reliance of existing architectures on fixed temporal configurations. Specifically, uniform frame rates and strict observation/prediction horizons. Consequently, when encountering heterogeneous sensor setups in the real world, both conventional discrete models [1, 4] and recent continuous-time approaches [6] suffer severe performance degradation, requiring labor-intensive fine-tuning to adapt. To characterize this generalization gap, we systematically formalize zero-shot trajectory prediction under distribution shifts into three distinct settings: **Temporal transfer** (unseen frame rates or horizons within the same scene), **Scene transfer** (unseen environments under matched temporal setups), and **Joint scene-temporal transfer** (both scene and temporal configurations are unseen). While the joint transfer setting represents the most realistic deployment scenario, it is largely overlooked in existing literature. Our empirical analysis reveals that *temporal mismatch* is the dominant, under-addressed source of failure in pre-trained models. By isolating temporal effects from dataset shifts, we demonstrate that explicitly conditioning the network on temporal metadata provides a simple yet highly effective mechanism to disentangle underlying motion dynamics from sensor sampling frequencies.

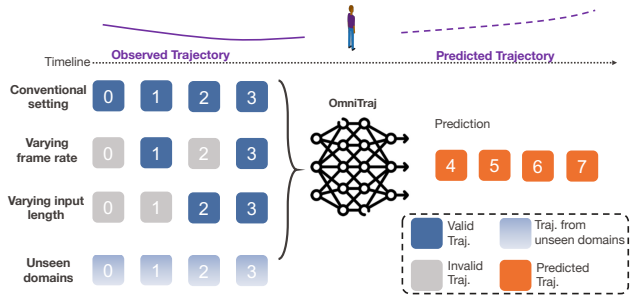


Figure 1. **OmniTraj: A pre-trained trajectory predictor that adapts to varying frame rates and horizons while excelling in zero-shot transfer.**

Building on these insights, we introduce **OmniTraj** (Figure 1), a flexible Transformer-based framework designed for robust zero-shot and few-shot inference across diverse temporal and spatial domains. To train OmniTraj, we develop a frame-rate- and horizon-agnostic data container, **UniHuMotion++**, expanding the pre-training pool

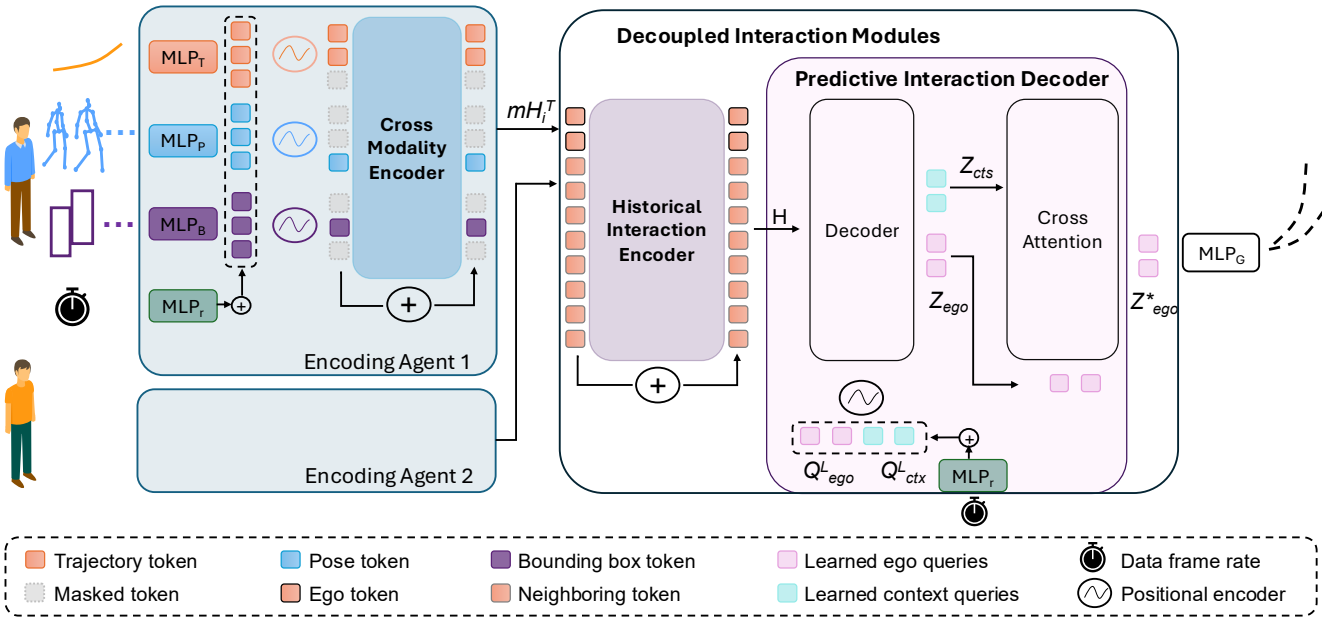


Figure 2. **OmniTraj Framework.** The model applies an explicit FPS embedding for temporal conditioning. The architecture decouples social reasoning: the Historical Interaction Encoder (HIE) processes observed dynamics, while the Predictive Interaction Decoder (PID) models future trajectories using partitioned ego/context queries and an ego-centric cross-attention mechanism.

to **859 hours** of multimodal motion data across **12 diverse datasets**. OmniTraj leverages this massive heterogeneous data through novel FPS-aware embeddings and Decoupled Interaction Modules, which explicitly inform the model of temporal configurations. In summary, our main contributions are:

- **Novel problem framing and empirical insights:** We formalize zero-shot trajectory prediction through three distinct distribution shifts. We identify temporal mismatch as the primary failure mode in SOTA predictors and prove that explicit temporal conditioning is a highly effective, architecture-agnostic solution.
- **OmniTraj framework:** We propose a temporally-aware, pre-trained Transformer framework equipped with FPS embeddings and Decoupled Interaction Modules to effectively capture complex spatial-temporal dynamics across heterogeneous setups.
- **SOTA generalization and unified data:** OmniTraj achieves a remarkable 70% error reduction in zero-shot joint scene-temporal transfer and exhibits exceptional robustness in sparse-observation edge cases (e.g., two-frame prediction). To facilitate future research, we introduce UniHuMotion++, the largest unified trajectory data framework natively supporting heterogeneous temporal configurations.

2. Method

OmniTraj is a temporally-aware, decoupled interactive Transformer designed to predict pedestrian trajectories across diverse real-world settings (Figure 2). Instead of relying on fixed temporal configurations, our architecture dynamically adapts to varying frame rates and missing observations through explicit temporal conditioning and cross-modal representation learning.

Formulation and Input Embeddings. Given an agent i over T_{obs} time-steps, we observe its multimodal cues, including historical trajectory, 3D/2D pose, and 3D/2D bounding boxes. To process these heterogeneous inputs, OmniTraj embeds each raw cue x_i^c via a cue-specific MLP, combined with positional encodings, agent identity, and keypoint-type embeddings. This transforms raw inputs into a set of foundational motion tokens.

Explicit Frame Rate Encoding. A core limitation of prior pre-trained models is their inability to transfer across different dataset sampling frequencies. To address this, OmniTraj explicitly conditions the network on the temporal metadata. Given an observation sequence captured at frame rate r, we compute a latent temporal embedding:

$$E_r = \text{MLP}_r(r),$$

This FPS embedding E_r is integrated via element-wise addition into the multimodal tokens. By injecting this explicit temporal signal, OmniTraj normalizes diverse temporal configurations into a shared latent space, effectively disentangling intrinsic human motion dynamics from dataset-specific sampling frequencies.

Cross-Modality Encoder (CME). Rather than processing modalities in isolation, OmniTraj uses a shared CME to

fuse the temporal-aware embeddings. The CME leverages trajectory-associated tokens directly to construct a unified, motion-centric representation (mH_i^T). Concurrently, auxiliary visual cues (e.g., pose and bounding boxes) are processed and mapped into the same common latent space, allowing the network to leverage complementary visual data when available or default to trajectory-only inference when sensors are limited.

Decoupled Interaction Modules. To accurately capture complex multi-agent dynamics, we introduce a decoupled architecture that separates historical and future social reasoning. First, the **Historical Interaction Encoder (HIE)** processes the refined CME tokens to capture past social interactions among agents, yielding a unified encoder context H.

Next, the **Predictive Interaction Decoder (PID)** models future interactions using a novel dual-query design. We partition learned queries into two distinct groups: Ego Queries (Q_{ego}^L), which extract the intrinsic motion features of the primary agent, and Contextual Queries (Q_{ctx}^L), which capture complementary scene-wide information. Processing these queries through the decoder yields separated feature representations:

$$Z_{\text{ego}} = D\big(H,\, Q_{\text{ego}}^L\big), \quad Z_{\text{ctx}} = D\big(H,\, Q_{\text{ctx}}^L\big).$$

To integrate these representations, we apply an ego-centric cross-attention mechanism where the ego tokens attend to the contextual tokens, yielding a refined predictive representation: $Z_{\text{ego}}^* = \text{CA}\big(Z_{\text{ego}},\, Z_{\text{ctx}}\big)$. This decoupling allows OmniTraj to explicitly model future ego-centric interactions, significantly reducing prediction errors.

Spatial-Temporal Masking and Prediction. To ensure robustness against missing frames, OmniTraj is pre-trained using a spatial-temporal masking strategy. We apply modality masking to encourage cross-modal representation learning and temporal masking to force the network to interpolate missing dynamics. Finally, an MLP-based prediction head projects the refined tokens Z_{ego}^* into the predicted future trajectory Y_1.

3. Experiments

Experimental Setup. To pre-train OmniTraj, we developed **UniHuMotion++**, a unified framework containing 859 hours of multimodal human motion data across 12 diverse datasets (e.g., NBA, JRDB, WOMD), agnostic to specific frame rates or horizons. We exclude trajnet++, ETH-UCY, and SDD for zero-shot evaluation and use the others for pre-training. We use the standard ADE/FDE and MinADE_{20}/MinFDE_{20} metrics.

Zero-Shot Distribution Shifts. We systematically evaluate zero-shot generalization under three distribution shifts.

1) Temporal Transfer: Evaluated on the NBA dataset with completely unseen temporal configurations (trained on (obs=10,pred=20,FPS=5) and (obs=4,pred=8,FPS=2.5), then tested on (obs=3,pred=3,FPS=1). As shown in Table 1 (left), OmniTraj outperforms both data-unaware discrete models (Multi-Transmotion) and continuous-time models (TrajSDE) by 30%, validating our temporal conditioning.

Method	ADE_{20}/FDE_{20}
M-Trans. [3]	1.68/2.15
TrajSDE [6]	1.71/2.05
OmniTraj	**1.18/1.22**

(a) Temporal transfer

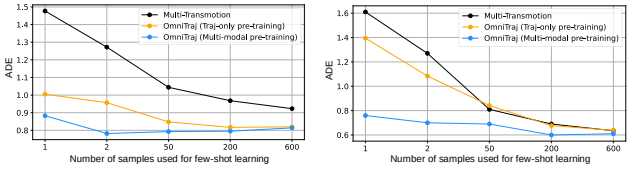
Method	Avg (A-E)
TrajCLIP [10]	0.68/1.36
OmniTraj	**0.49/0.92**

(b) Scene transfer

Table 1. **Zero-Shot temporal or scene transfer.** OmniTraj excels in isolated temporal shifts (NBA) and scene shifts (ETH-UCY).

Model	Trajnet++ ADE	SDD ADE
Multi-Transmotion [3]	3.40	3.58
OmniTraj (Traj-only)	1.57 (↓53%)	1.91 (↓46%)
OmniTraj (Multi-modal)	**1.01** (↓**70%**)	**0.93** (↓**74%**)

Table 2. **Zero-Shot Joint Transfer.** OmniTraj yields massive improvements on unseen datasets and heterogeneous frame setups.



(a) Trajnet++ (b) SDD

Figure 3. **Few-shot efficiency.** OmniTraj adapts to new domains with fewer samples than prior models.

2) Scene Transfer: Following the cross-scene protocol on ETH-UCY (where subsets A-E represent different spatial domains). Table 1 (right) shows OmniTraj outperforms state-of-the-art contrastive approaches like TrajCLIP, reducing average MinADE_{20} by 28%.

3) Joint Scene-Temporal Transfer: The most realistic and challenging setting. As shown in Table 2, when evaluating models on completely unseen datasets (Trajnet++ and SDD) with unseen temporal configurations, OmniTraj achieves a massive **>70% error reduction** over Multi-Transmotion. Multi-modal pre-training further enhances this generalization gap.

Few-Shot Data Efficiency. When target domain data is extremely limited, OmniTraj exhibits remarkable sample efficiency. As shown in Figure 3, OmniTraj fine-tuned with merely 2 samples consistently outperforms Multi-Transmotion fine-tuned with 200 samples on both Trajnet++ and SDD.

Ablation: Explicit Temporal Conditioning. Table 3 ablates the FPS-encoder under the temporal transfer setting. Implicitly learning temporal shifts (e.g., via velocity encoding) couples time discretization with agent-specific motion, yielding suboptimal results (1.54/2.00). Conversely, our explicit MLP-based FPS encoder with latent-

FPS Encoder Variant	MinADE$_{20}$/FDE$_{20}$
w/o FPS-encoder	1.87/2.49
w/ FiLM-based [7]	1.62/2.26
w/ MLP Vel-encoder, Latent sum	1.54/2.00
w/ MLP FPS-encoder, Latent sum (Ours)	**1.18/1.22**

Table 3. **FPS-Encoder Ablation (Temporal Transfer).**

Models	Full Frames ADE/FDE	Two-Frame ADE/FDE
Social-Force [5]	0.83/1.49	0.85/1.53
Social-Transmotion [8]	0.69/1.45	1.08/2.03
OmniTraj (Ours)	**0.66/1.37**	**0.73/1.48**

Table 4. **Precognition under sparsity (NuScenes).** OmniTraj maintains high accuracy even when given only two historical frames.

space summation effectively disentangles motion dynamics from sampling frequencies, achieving the lowest errors (**1.18/1.22**). Furthermore, ablations on our decoupled architecture confirm that separating Historical Interaction Encoding (HIE) from Predictive Interaction Decoding (PID) improves MinADE$_{20}$ by over 18% compared to a unified backbone.

Full-Data Fine-Tuning & LLM Comparison. When fully fine-tuned, OmniTraj sets a new state-of-the-art on standard benchmarks. On the highly interactive NBA dataset, it achieves **0.73/0.91** (MinADE/FDE$_{20}$), surpassing NMRF [2]. On JTA, our trajectory-only model (**0.90/1.81** ADE/FDE) outperforms previous SOTA models (e.g., EmLoco [9]) that rely on complex 3D pose inputs. In zero-shot evaluations against LLMs on ETH-UCY, OmniTraj (0.24 MinADE$_{20}$) massively outperforms GPT-4 (0.39 MinADE$_{20}$), showing the gap between the pre-trained trajectory model and the general foundation model.

Robustness in Safety-Critical Scenarios (Two-Frame). In real-world driving, pedestrians often appear suddenly from occlusions, leaving minimal historical context. Table 4 evaluates this extreme sparsity on NuScenes using only two observed frames. While data-driven models like Social-Transmotion suffer a 57% ADE degradation under sparse inputs, OmniTraj's temporally-aware pre-training inherently robustifies it against missing frames, achieving the lowest errors and demonstrating high reliability for autonomous navigation.

4. Conclusion

We address the challenge of deploying trajectory predictors under real-world distribution shifts. By formalizing zero-shot generalization into temporal, scene, and joint transfers, we identify temporal mismatch as the primary bottleneck in existing models. To overcome this, we introduce OmniTraj, a temporally-aware Transformer pre-trained on heterogeneous data (UniHuMotion++). Explicit temporal conditioning enables OmniTraj to disentangle intrinsic motion dynamics from dataset-specific sampling frequencies. Consequently, OmniTraj achieves state-of-the-art zero-shot generalization, reducing prediction errors by $>70\%$ under joint scene-temporal shifts. Its exceptional few-shot efficiency and robustness to extreme observational sparsity establish OmniTraj as a scalable, reliable foundation model for predictive vision and autonomous navigation.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016. 1

[2] Zilin Fang, David Hsu, and Gim Hee Lee. Neuralized markov random field for interaction-aware stochastic human trajectory prediction. In *The Thirteenth International Conference on Learning Representations*, 2025. 4

[3] Yang Gao, Po-Chien Luan, and Alexandre Alahi. Multi-transmotion: Pre-trained model for human motion prediction. In *8th Annual Conference on Robot Learning*, 2024. 1, 3

[4] Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. In *International Conference on Learning Representations (ICLR)*, 2022. 1

[5] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 4

[6] Daehee Park, Jaewoo Jeong, and Kuk-Jin Yoon. Improving transferability for cross-domain trajectory prediction via neural stochastic differential equation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10145–10154, 2024. 1, 3

[7] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4

[8] Saeed Saadatnejad, Yang Gao, Kaouther Messaoud, and Alexandre Alahi. Social-transmotion: Promptable human trajectory prediction. In *International Conference on Learning Representations (ICLR)*, 2024. 4

[9] Hiromu Taketsugu, Takeru Oba, Takahiro Maeda, Shohei Nobuhara, and Norimichi Ukita. Physical plausibility-aware trajectory prediction via locomotion embodiment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12324–12334, 2025. 4

[10] Pengfei Yao, Yinglong Zhu, Huikun Bi, Tianlu Mao, and Zhaoqi Wang. Trajclip: Pedestrian trajectory prediction method using contrastive learning and idempotent networks. *Advances in Neural Information Processing Systems*, 37: 77023–77037, 2024. 3