We would like to thank all reviewers for their valuable and helpful suggestions. We first respond to the common request on ablation study of our proposed encoder as shown in Table 1. Due to space limit, we will report results on other datasets, more details about model description and related work, and revise typos in our final version as suggested.

Ablation Study. Our proposed encoder contains two modules: dynamic node representation network followed by a temporal self-attention. The goal of the first module is to simultaneously capture spatial-temporal dependency among nodes. We achieve this by introducing temporal dependency to spatial-based GNN with learnable positional encoding and attention mechanism. To test the efficiency of each component, we remove them separately (LG-ODE-no att, LG-ODE-no PE) and find the performances drop. This suggests that distinguishing the importance of nodes w.r.t time and incorporating temporal information via learnable positional encoding would benefit model performance. Additionally, to test the performance of adding learnable parameters and nonlinearity in positional encoding, we compare with manually-designed positional encoding [*15*] (LG-ODE-fixed PE) and find our method more flexible which produces more efficient temporal encoding. Secondly, to test the efficiency of temporal self-attention, we adopt different sequence aggregation methods (LG-first, LG-mean) and find our method performs the best. This suggests that nodes at different timestamp would represent different semantic meanings towards the initial state of the whole system.

Table 1: Mean Square Error (MSE) $\times 10^{-2}$ of Ablation Study and Baselines on Spring Dataset.

Model		LG-ODE	LG-ODE -no att	LG-ODE -no PE	LG-ODE -fixed PE	LG-ODE -first	LG-ODE -mean	Latent -ODE	NRI+RNN -imputation
Interpolation	40%	**0.3350**	0.5145	0.4431	0.4285	1.3017	0.3896	0.5454	2.0743
	60%	**0.3170**	0.4198	0.4278	0.4445	1.1918	0.3901	0.5036	1.9857
	80%	**0.2641**	0.4510	0.3879	0.4083	1.0796	0.3268	0.4290	1.9573
Extrapolation	40%	**1.7839**	2.3847	1.7943	1.7905	6.5742	2.2499	6.6023	3.8966
	60%	1.8084	2.1216	1.8172	**1.7634**	6.3243	2.1165	4.2478	3.8749
	80%	**1.7139**	1.9634	1.7332	1.7545	5.7788	2.2516	4.3192	3.5762

Reviewer 1.

A1. Interaction aspect of the model and Fig.2 explanation. To show the importance of graph interaction, we compare with Latent-ODE which processes each timeseries individually. Our model outperforms it over two tasks as shown in Table 1. For Fig.2, we would like to clarify the terminology "interpolation" following the existing work [*7*]. We try to fit a curve using observed time points with a goal to minimize the MSE. Fig.2 plots the predicted curve as interpolation results and these plotted prediction values may differ from the truth values that are conditioned on.

A2. Experiments and limitations. Thanks for your advice on experiments! We will add additional mocap sequences and provide video link later as rebuttal allows no links. The limitation of our model is that we assume the graph structure is fixed, but in reality graph structure also changes w.r.t time. We will leave it as a future work to further explore.

Reviewer 2. To the best of our knowledge, we are the first to handle irregularly-sampled partial observations with known graph structure. The two papers you mentioned do not consider graph structure. The second paper only handles irregularly-sampled data but not partially-observed dynamic system. It assumes all agents' observations are aligned.

Reviewer 3.

A1. To make our model comparable with existing ones, we compare with baselines from two problem variants. Firstly, we employ RNN-imputation [*R1*] where the graph structure is not considered. It jointly imputes missing values (interpolation) for all agents by simple concatenation of feature vectors. As shown in Table 1, the performance drops which shows that such graph structure is essential for predicting interacting systems. Secondly, to show the effectiveness of our way to handle irregularly-sampled partial observations, we combine RNN-impuation with NRI [*2*] where we first impute each timeseries into regular-sampled one to make it a valid input for NRI, and then predict trajectories jointly with graph structure (extrapolation). As shown in Table 1, the prediction error is large and one possible reason is that we use estimated imputation values for missing data which would cause noise to NRI. Also the two-step process separates imputation with prediction, whereas our approach is an end-to-end framework for both two tasks.

A2. For Eqn2, we adopt the GNN model in [*2*] to capture the interaction among agents. It firstly employs a shared relation function f_R to compute pair-wise influence, then employ a shared object function f_O for influence aggregation. Such weights sharing mechanism is commonly utilized in various GNN models [*2,3,20*].

Reviewer 4. We respectfully disagree with your comment that our model is incremental by extending Latent ODE. In multi-agent system, despite each timeseries can be irregularly-sampled, such system can be only partially observed (timeseries are not aligned). Also as agents continuously influence each other, how to combine such interaction with irregular partial observations to make predictions remains challenging. Latent ODE only deals with single timeseries and is not able to solve these problems. We therefore design a novel encoder that extracts spatial-temporal pattern from irregularly-sampled partial observations and graph structure, and use it to infer all initial states simultaneously. The whole system is then driven by a GNN that models continuous interaction among agents along time.

[R1] *Che, Z. et al. "Recurrent Neural Networks for Multivariate Time Series with Missing Values." Scientific reports vol. 8,1 6085. 2018*