TOWARDS SCENE GRAPH ANTICIPATION

Rohith Peddi
The University of Texas at Dallas
rohith.peddi@utdallas.edu

Saksham Singh
Indian Institute of Technology Delhi
mt1200841@maths.iitd.ac.in

Saurabh Srivastava
Indian Institute of Technology Delhi
csy227518@cse.iitd.ac.in

Parag Singla
Indian Institute of Technology Delhi
parags@cse.iitd.ac.in

Vibhav Gogate
The University of Texas at Dallas
vibhav.gogate@utdallas.edu

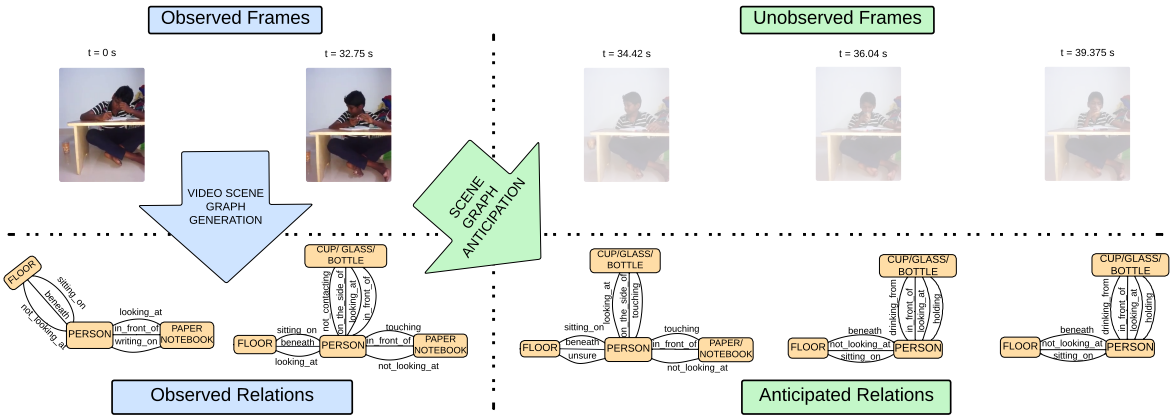

Figure 1: Pictorially contrasts the difference between the tasks of video scene graph generation and scene graph anticipation.

ABSTRACT

Spatio-temporal scene graphs represent interactions in a video by decomposing scenes into individual objects and their pair-wise temporal relationships. Long-term anticipation of the fine-grained pair-wise relationships between objects is a challenging problem. To this end, we introduce the task of Scene Graph Anticipation (SGA). We adopt a state-of-the-art scene graph generation method as a baseline to anticipate future pair-wise relationships between objects. In our proposed approaches, we leverage object-centric representations of relationships to reason about the observed video frames and model the evolution of relationships between objects. We take a continuous time perspective and model the latent dynamics of the evolution of object interactions using concepts of NeuralODE and NeuralSDE, respectively. We infer representations of future relationships by solving an Ordinary Differential Equation and a Stochastic Differential Equation, respectively. Extensive experimentation on the Action Genome dataset validates the efficacy of the proposed methods.

1 Introduction

We focus on spatio-temporal scene graphs [1], which is a widely used framework for representing the evolving spatial and temporal relationships among objects. These graphs contain information about the objects present in a video, including their categories, positions, sizes, and spatial dependencies. Simultaneously, they illustrate how these relationships evolve over time, revealing the movement, interactions, and configuration changes of objects across consecutive frames in a video sequence. They facilitate our understanding of dynamic scenes and serve as a valuable tool for addressing downstream tasks in applications such as action recognition and video analysis, where the temporal dynamics of object interactions play a crucial role.

This paper introduces a novel task known as *Scene Graph Anticipation (SGA)*, which, given a video stream, aims to forecast future interactions between objects, as shown in Figure 1. The Scene Graph Anticipation (SGA) task holds significance across diverse domains due to its potential applications and relevance to several downstream tasks. For instance, it contributes to *enhanced video understanding* by predicting spatiotemporal relationships within video scenes, facilitating improved video analysis and the interpretation of complex object interactions over time. SGA plays a crucial role in *activity recognition*, enabling systems to predict future object interactions for more accurate classification and advanced surveillance. Anticipation aids in *anomaly detection* by identifying deviations from expected object relationships, thereby enhancing the detection of abnormal events in video sequences. *Intelligent surveillance systems* can also benefit from SGA, allowing systems to predict and respond to security threats by understanding evolving object relationships in monitored environments. Finally, for applications in *robotics and autonomous systems*, SGA is essential for predicting object movements and interactions, contributing to safer and more efficient navigation and decision-making processes.

To tackle the challenges of SGA, we introduce two innovative approaches that extend a state-of-the-art scene graph generation method [2]. These approaches utilize object-centric representations of relationships, allowing us to analyze observed video frames and model the dynamic evolution of interactions between objects effectively. This detailed understanding of temporal dynamics serves as the basis for our proposed methods.

In a departure from traditional sequential modelling, our two approaches adopt a continuous-time perspective. Drawing inspiration from Neural Ordinary Differential Equations (NeuralODE) [3] and Neural Stochastic Differential Equations (NeuralSDE) [4], respectively, we develop methods that capture the latent dynamics governing the evolution of object interactions. By formulating the anticipation problem as solving Ordinary Differential Equations (ODE) and Stochastic Differential Equations (SDE) and thus using a continuous representation of the anticipated relationships, our hope is to significantly expand the fidelity of our predictions. We rigorously validated our proposed methods as well as a strong generation-based baseline on the Action Genome dataset [1], a benchmark for spatio-temporal scene understanding. Our experimental results demonstrate the superior performance of our approaches in accurately anticipating fine-grained pair-wise relationships between objects.

2 Related Work

There has been significant interest in using structured representations of visual content to bridge the gap between vision and language.

2.1 Structured Representation

Spatial Graphs. Learning to represent visual content in static data such as 2D/3D images as a spatial graph where the objects act as nodes and edges to describe the visual relation between the objects is called *Image Scene Graph Generation* (ImgSGG). There has been extensive research in 2D ImgSGG direction following the seminal work of Visual Genome [5]. Following this, the task has been extended to 3D by the introduction of [6]. Recently, a surge of ImgSGG methods explored the role of foundation models in open-vocabulary ImgSGG [7], weakly supervised ImgSGG [8], panoptic ImSGG [9]

Spatio-Temporal Graphs. Dynamic visual content, such as videos, provides a more natural set of contextual features describing dynamic interaction between objects. Encoding such content into a structured spatiotemporal graphical representation for frames where nodes describe objects and edges describe the temporal relations is called Video Scene Graph Generation (VidSGG). Early approaches on VidSGG extended previously proposed ImSGG-based methods to the temporal domain. We refer to [10] for a comprehensive survey on earlier work. Recent work explored learning better representations using architectures like Transformers [2, 11] and unbiased representations [12, 13] owing to the long-tailed nature of the benchmark VidSGG datasets Action Genome [1], VidVRD [14].

The task introduced in the paper is related to [15], wherein they use relation forecasting as an intermediate step for action prediction. However, we aim for accurate anticipation of future relations, present a learning methodology, and propose evaluation metrics. Scene Graph Anticipation is closely related to [16], which constructed a dataset from Action Genome by sampling frames and truncating the videos; instead, we train models on the complete dataset. We also note that the approach in [16] is similar to the baseline proposed in this paper.

2.2 Dynamics Models

Scene Graph Anticipation is also closely related to the task of Video Prediction. Early video prediction methods treat dynamics modelling as a sequential prediction problem in pixel space using image-level features [17]. Later approaches proposed include using external knowledge as priors [18, 19], better architectural design to model contextual information [20, 21], and focusing on object-centric representations [22]. Recently, there has been a surge in methods proposed that use diffusion models to estimate the distribution of a short future video clip [23, 24]. Unlike dense pixel-based generation in video prediction, in SGA, we are primarily interested in learning representations that aid in predicting long-term interactions and their fine-grained relationship representations. Our experiments estimate relationship representations more than 30 seconds into the future.

After [3] introduced NeuralODE, several methods were proposed to explore latent dynamics models using the framework of NeuralODE. These include methods on trajectory prediction [25], traffic forecasting [26], video generation [27], and multi-agent dynamical systems [28, 29, 30].

3 Background

Ordinary Differential Equations. An initial value problem (IVP) formulated as an instance of the ODE is given by

$$\frac{d\mathbf{z}(t)}{dt} = \mathbf{f}(t, \mathbf{z}(t)), \mathbf{z}\left(t_0\right) = z_0 \tag{1}$$

Here, $\mathbf{f} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ represents time-varying smooth vector field and function $\mathbf{z}(t)$ is a solution to the IVP. Recently, [3] presented the framework of NeuralODE, wherein they relaxed the time-variance of the vector field $\mathbf{f}$ and learned it from data by parameterizing it with a neural network. In their work, they presented learning a Latent ODE that models the dynamics of a latent state.

Numerical Methods. Finding analytical solutions for complex ordinary differential equations is infeasible. We use numerical methods to approximate the solution. The key idea employed by ODE solvers is discretizing the time domain into small steps and approximating the solution at each step. Thus, they trade off precision with computation time. On the faster, less accurate side, we have a single-step method, such as Euler's method, while on the slower, more accurate side, we have multistep methods, such as Adams-Bashforth.

Stochastic Differential Equations. An initial value problem formulated as an instance of the SDE is given by

$$\mathrm{d}\mathbf{z}(t) = \mu(\mathbf{z}(t), t)\mathrm{d}t + \sigma(\mathbf{z}(t), t)\mathrm{d}\mathbf{W}(t), \mathbf{z}\left(t_0\right) = z_0 \tag{2}$$

Here, $\mathbf{W}(t)$ denotes a Wiener Process and the terms $\sigma(\mathbf{z}(t), t) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d \times m}$, $\mu : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ represents diffusion and drift respectively. As NeuralODE, we can parameterise both drift and diffusion terms using neural networks and learn them from data called a NeuralSDE [31, 4].

Numerical Methods. Similar to ODEs, solving SDEs analytically can often be challenging or impossible. There are many interpretations of a stochastic differential equation, but Ito and Stratonovich are popular. The choice of interpretation is connected to the conceptual model underlying the equation and impacts the numerical solution of these equations. Stratonovich's interpretation of SDE is commonly used to model physical systems subjected to noise. In ablation, we provide the impact of the type of interpretation of SDE on our learning problem.

4 Notation & Problem Description

Notation. We use $[,]$ to denote stacking operation and use $\langle, \rangle$ to denote concatenation operation. We use the representation $\{.^t\}_{t=0}^N$ to denote a set of N vectors. We denote an observed key-frame at time t as $\mathcal{I}^t$ and an object

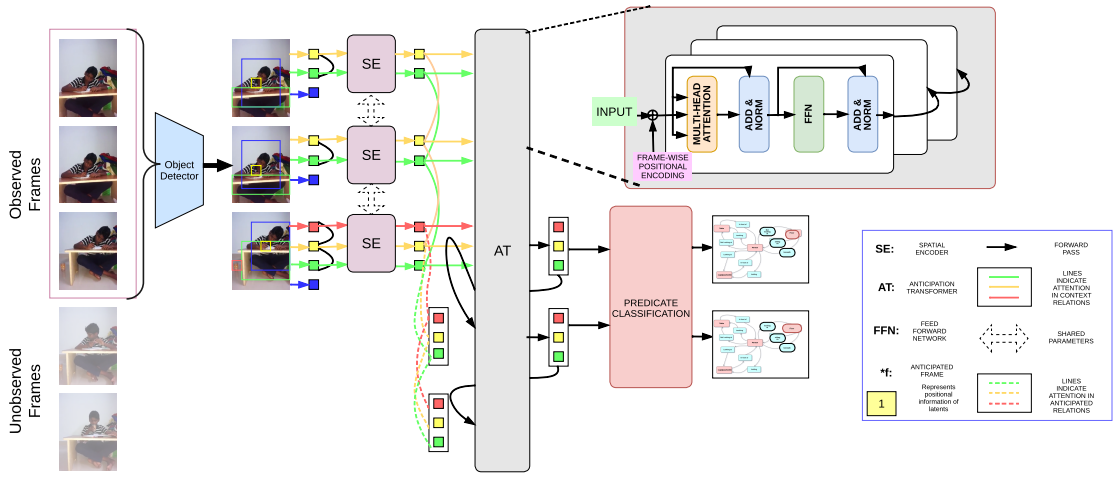

Figure 2: Architecture diagram for the proposed method SgatTran. Given the context frames (observed) in the red box, relationship predictions are made for the greyed-out (unobserved) frames. Different colours are used to distinguish between objects. The arrows between objects after the spatial encoder block depict the temporal attention across frames.

observed in frame $\mathcal{I}^t$ as $\mathcal{O}^t$. We use $\mathcal{C}$ and $\mathcal{P}$ to denote object class set and predicate class set, respectively. Each observed object $\mathcal{O}^t$ is identified by its bounding box $\mathbf{b}$ and object class $\mathbf{c}$, where $\mathbf{b} \in [0,1]^4$ and $\mathbf{c} \in \mathcal{C}$. We denote the relationship tuple between objects, $\mathcal{O}_i^t$ and $\mathcal{O}_j^t$, using $\mathcal{R}_{ij}^t$, where $\mathcal{R}_{ij}^t = \left(\mathcal{O}_i^t, p^t, \mathcal{O}_j^t\right)$ and predicate $p^t \in \mathcal{P}$. Each pair of objects $\left(\mathcal{O}_i^t, \mathcal{O}_j^t\right)$ can have multiple relationships, and we denote them using $\mathbf{p}_{ij}^t$, where $\mathbf{p}_{ij}^t \in \{0,1\}^{|\mathcal{P}|}$, thus a relationship triplet can be expressed as $\mathcal{R}_{ij}^t = \left(\mathcal{O}_i^t, {}^k p_{ij}^t, \mathcal{O}_j^t\right)$, where $\mathbf{p}_{ij}^t[k] = 1$. We denote the distribution of the object classes and predicate classes as $\hat{\mathbf{c}} \in [0,1]^{|\mathcal{C}|}$ and $\hat{\mathbf{p}}^t \in [0,1]^{|\mathcal{P}|}$, where $\sum_i \hat{\mathbf{c}}[i] = 1$.

Scene Graph Anticipation (SGA). Given a video segment, SGA predicts future objects and their relationships based on observed interactions (see Fig 1). We note that anticipation and localization of new objects that appear in the future are significantly harder problems. In this paper, we have relaxed the problem, redefined it and presented a methodology.

Task Description. Given a video segment identified by observation of a set of key frames, we assume the presence of observed objects in future frames and predict the evolution of fine-grained relationships between them in future frames. Throughout the paper, we will use the terms generation to refer to Video Scene Graph Generation (VidSGG) and anticipation to refer to Scene Graph Anticipation (SGA). Formally, we define the tasks of generation and anticipation as follows:

Generation. Given $\mathbf{V} = \{\mathcal{I}^t\}_{t=-T}^{0}$ as the set of observed frames in a video, generation entails detecting $N(t)$ objects $\{\mathcal{O}_i^t\}_{i=0}^{N(t)}$ in key-frame $\mathcal{I}^t$, where $N(t)$ is the total number of objects in frame $\mathcal{I}^t$ and predicting relationships $\left(\mathcal{O}_i^t, {}^k p_{ij}^t, \mathcal{O}_j^t\right)_{ijk}$ between them.

Anticipation. Given $\mathbf{V} = \{\mathcal{I}^t\}_{t=-T}^{0}$ as the set of observed frames in a video, in anticipation, we assume the presence of objects observed at time $t = 0$ in all future frames $\{\mathcal{I}^t\}_{t=0}^{T}$ i,e $\{\mathcal{O}_i^t\}_{i=0}^{N(t)} = \{\mathcal{O}_i^0\}_{i=0}^{N(0)}, \forall t > 0$ and predict the relationships between the objects $\left(\mathcal{O}_i^t, {}^k p_{ij}^t, \mathcal{O}_j^t\right)_{ijk}$.

Graph Generation Strategies. In order to generate anticipated graphs, we utilize established scene graph generation strategies that have been widely adopted in VidSGG literature. These include (a) **With Constraint** permits a single edge between a pair of objects in the anticipated graph; specifically $\mathbf{p}_{ij}$ will be a one-hot vector with $\mathbf{p}_{ij}[k] = 1$ and $\mathbf{p}_{ij}[l] = 0, \forall l \neq k$. (b) **No Constraint** allows the anticipated graph to contain multiple edges between a pair of objects; specifically, we include all relationship triplet predictions $\left(\mathcal{O}_i^t, {}^k p_{ij}^t, \mathcal{O}_j^t\right)_{ijk}$ in the graph.

5 Technical Approach

We begin by explaining the local predicate embedding generator, a block that is common to all the approaches presented in the paper. Then, we will describe the proposed baseline (SgatTran) and proceed to present proposed approaches using the concepts of NeuralODE and NeuralSDE.

5.1 Local Predicate Embedding Generator

Our proposed approaches were evaluated on the Action Genome benchmark [1]. This dataset primarily encompasses videos with a single actor interacting with objects. Without loss of generality, we motivate our approach from the lens of a subject-object pair $(\mathcal{O}_s^t, \mathcal{O}_o^t)$ in the consequent sections. Thus, our goal is to predict future relationship triplets $\left(\mathcal{O}_s^t, {}^k p_{so}^t, \mathcal{O}_o^t\right), \forall t > 0$, given $\mathbf{V} = \{\mathcal{I}^t\}_{t=-T}^{0}$

Object Detection. We employ a pre-trained Faster R-CNN [32] to extract visual features $\{\mathbf{v}_i^t\}_{i=0}^{N(t)}$, bounding boxes $\{\mathbf{b}_i^t\}_{i=0}^{N(t)}$ and object category distributions $\{\mathbf{d}_i^t\}_{i=0}^{N(t)}$ of $N(t)$ object proposals for the observed frames.

Object Classification. Following [2], for observed frames, we use a combination of visual features and object category distributions provided by an object detector and feed it through a two-layer neural network to compute the object distribution. We use the standard cross-entropy loss $\mathcal{L}_{(o)}^t$ to compute the loss.

Relationship Representation. Following [2], we build subject-object relationship representation $\mathbf{z}_{so}^t$ by concatenating visual and semantic features of subject and object proposals obtained from the detector as follows:

$$\mathbf{z}_{so}^t = \left\langle \mathbf{W}_s \mathbf{v}_s^t, \mathbf{W}_o \mathbf{v}_o^t, \mathbf{W}_u \mathcal{U}_{ij}^t, \mathcal{S}_s^t, \mathcal{S}_o^t \right\rangle \tag{3}$$

$\mathbf{W}_s, \mathbf{W}_o$ and $\mathbf{W}_u$, represent learnable weights of linear layers used for compressing visual features, $\mathcal{S}_s^t, \mathcal{S}_j^t$ denote semantic embedding vectors [33] of subject and object categories predicted by the detector.

Spatial Encoder. Following VidSGG methods, we employ a spatial transformer that aggregates information from the spatial context by attending to all the relationship representations observed in a frame. Specifically, for each observed frame $\mathcal{I}^t$, we construct $\mathbf{Z}^t = [\{\mathbf{z}_{so}^t\}_{so}]$ use it as input for the spatial transformer. Here, the queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ share the same input and the output of the n-th encoder layer is presented as:

$${}^n\mathbf{Z}^t = \text{SpatialEncoder}\left(\mathbf{Q} = \mathbf{K} = \mathbf{V} = {}^{n-1}\mathbf{Z}^t\right) \tag{4}$$

We omit n from output representation ${}^n\mathbf{Z}^t$ and write it as $\mathbf{Z}^t$ for the next stages of the proposed approach.

5.2 Approach: SgatTran

We propose a baseline for the task of anticipation by employing a transformer to anticipate the relationship representations of the interacting objects. Figure 2 presents the architecture proposed for the method SgatTran.

Anticipatory Transformer. Following [34], we rearrange the output representations $\{\mathbf{Z}^t\}_{t=-T}^{0}$ for observed frames $\{\mathcal{I}^t\}_{t=-T}^{0}$ according to their subject and object classes $\mathbf{z}_{so}^t$. Now, we stack relationship representations $\mathbf{z}_{so}^t$ corresponding to a subject-object classes (s, j) for all observed frames into a matrix given by $\mathbf{Z}_{so}^{(-T:0)} = \left[\{\mathbf{z}_{so}^t\}_{t=-T}^{0}\right]$. Now, for a given horizon H, we generate future relationship representations $\{\mathbf{z}_{so}^t\}_{t=1}^{H}$ autoregressively. Specifically, we first feed in the matrix $\mathbf{Z}_{so}^{(-T:t)}$ into an Anticipatory Transformer that performs temporal reasoning and predicts the next future representation $\mathbf{z}_{so}^{t+1}$. Then, we concatenate the prediction to the matrix to get $\mathbf{Z}_{so}^{(-T:t+1)}$ and feed it back to the Transformer to generate a new prediction until the end of the horizon. Our approach can be given as follows

$$\begin{aligned} \mathbf{z}_{so}^{t+1} &= \text{AnticipatoryTransformer}(\mathbf{Z}_{so}^{(-T:t)}) \\ \mathbf{Z}_{so}^{(-T:t+1)} &= \left[\mathbf{Z}_{so}^{(-T:t)}, \mathbf{z}_{so}^{t+1}\right] \end{aligned}$$

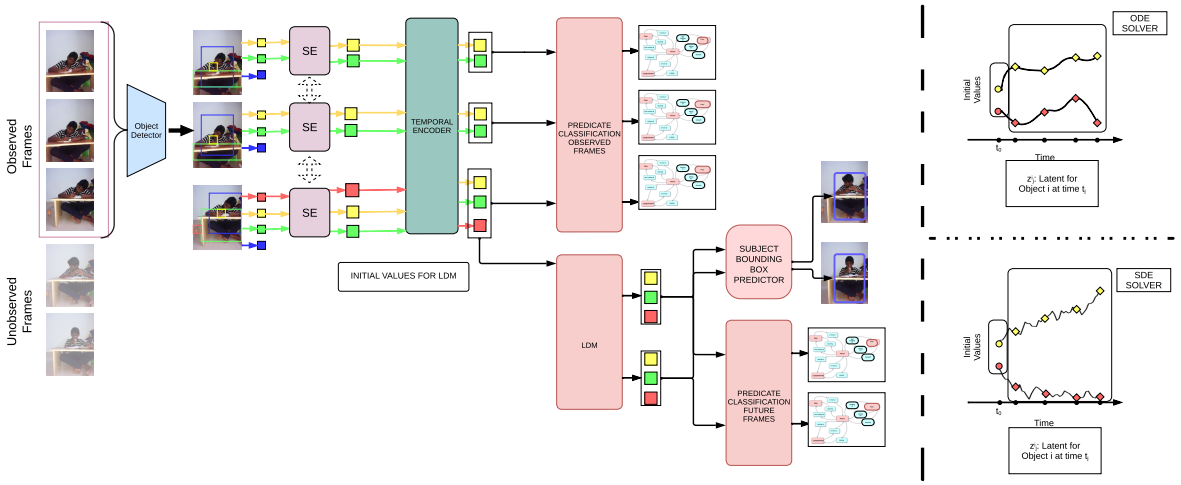

Figure 3: Architecture diagram for the proposed methods SgatODE and SgatSDE. The leftmost part of the model comprises a local predication embedding generator. The Neural ODE/SDEs are used to extrapolate the relationship representations to the timestamps of the future frames.On the right is a visualization of potential trajectories in the latent space modelled by Neural ODE/SDEs.

Loss Function. In order to train the Anticipatory Transformer, we define a loss function on the decoded, predicted future relationship representations. Specifically, for each subject-object pair (s, o), we feed the predicted future relationship representations $\{\mathbf{z}_{so}^t\}_{t=1}^H$ to three classification heads constructed to predict attention relationship, spatial relationships and contacting relationships respectively. Following VidSGG methods, we apply a multi-label margin loss for the classification of the anticipated relationship representations. The predicate classification loss function is given by:

$$\mathcal{L}_{p_{so}^t}(\hat{\mathbf{p}}_{so}^t) = \sum_{i \in \mathcal{P}^+} \sum_{j \in \mathcal{P}^-} \max(0, 1 - \hat{\mathbf{p}}_{so}^t[j] + \hat{\mathbf{p}}_{so}^t[i]) \tag{5}$$

Thus, the total objective for the proposed method SgatTran is the combination of the object classification loss ($\mathcal{L}_{(o)}^t$) for observed frames and the predicate classification loss ($\mathcal{L}_{p_{so}^t}$) for the anticipated relationship representations along the horizon. It can be written as follows:

$$\mathcal{L}_{\text{SgatTran}} = \sum_{t=-T}^{0} \mathcal{L}_{(o)}^t + \sum_{t=1}^{H} \sum_{so} \mathcal{L}_{p_{so}^t} \tag{6}$$

5.3 Approach: SgatODE

Under the assumption that there exists a differential equation that governs the evolution of the relationship representations between interacting objects. Following the work on NeuralODE, we propose an approach to learn a vector field from data that models this evolution. We cast the anticipation problem as an instance of the ODE Initial Value Problem with the initial condition as a latent feature corresponding to the last observed subject-object relationship representation.

$$\mathbf{z}_{so}(t) = \mathbf{z}_{so}(0) + \int_0^t \mathbf{f}_\theta(\mathbf{z}_{so}(s))\mathrm{d}s \tag{7}$$

Here $\mathbf{f}_\theta$ represents the ODE function that models the dynamics of the evolution of relationships.

Temporal Encoder. We use the output representations $\{\mathbf{Z}^t\}_{t=-T}^0$ of the local predicate embedding generator (see 5.1) to compute matrix $\mathbf{Z}_{so}^{(-T:0)}$ by stacking relationship representations $\mathbf{z}_{so}^t$ for all observed frames. To further enrich relationship representations $\mathbf{z}_{so}^t$ with temporal information, we employ a temporal encoder that aggregates information

from all previously observed relationship representations corresponding to a subject-object pair through attention. Specifically, we feed $\mathbf{Z}_{so}^{(-T:0)}$ as input to a temporal transformer with queries $\mathbf{Q}$, keys $\mathbf{K}$ and values $\mathbf{V}$ sharing the same input and the output of the n-th encoder layer is presented as:

$$ {}^{n}\mathbf{Z}_{so}^{(-T:0)} = \text{TemporalEnc}\left(\mathbf{Q} = \mathbf{K} = \mathbf{V} = {}^{n-1}\mathbf{Z}_{so}^{(-T:0)}\right) $$

We omit n from output representation ${}^{n}\mathbf{Z}_{so}^{(-T:0)}$ and write it as $\mathbf{Z}_{so}^{(-T:0)}$ for the next stages of the proposed approach.

Generation Loss. We feed the output of the temporal encoder $\mathbf{Z}_{so}^{(-T:0)}$ to relationship classification heads to predict the attention, spatial and contacting relationships in the observed frames. We apply a multi-label margin loss to classify the observed relationship representations (see eq. 5). We call this Generation Loss as this corresponds to predicate classification loss of the observed frames.

$$ \mathcal{L}_{\text{gen}} = \sum_{t=-T}^{0} \sum_{so} \mathcal{L}_{p_{so}^{t}} \tag{8} $$

Anticipation Loss. We use the output relationship representation $\mathbf{z}_{so}^{0}$ from the temporal encoder as the initial value for the proposed Initial Value Problem. We use a multi-step Adam's method to extrapolate the relationship representation $\{\mathbf{z}_{so}^{t}\}_{t=0}^{H}$. We feed the anticipated relationship representations $\{\mathbf{z}_{so}^{t}\}_{t=0}^{H}$ to three classification heads constructed to predict attention, spatial and contacting relationships. We apply multi-label margin loss for the anticipated relationship representations, which we call Anticipation Loss.

$$ \mathcal{L}_{\text{ant}}^{\text{ode}} = \sum_{t=0}^{H} \sum_{so} \mathcal{L}_{p_{so}^{t}} \tag{9} $$

Decoder. In the AG benchmark dataset, a single actor interacts with objects. We note that we extrapolate relationship representations corresponding to subject-object pairs. So, we propose an additional loss term to enforce the constraint that a single actor interacts with all the observed objects. Specifically, we feed each of the anticipated relationship representations to a linear layer to predict the bounding box of the subject and Smoothed L1 Loss is applied to the predicted bounding boxes, which we denote using $\mathcal{L}_{\text{boxes}}$

Loss Function. Thus, the total objective for the proposed method SgatODE can be written as:

$$ \mathcal{L}_{\text{SgatODE}} = \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{ant}}^{\text{ode}} + \lambda \mathcal{L}_{\text{boxes}} + \sum_{t=-T}^{0} \mathcal{L}_{(o)}^{t} \tag{10} $$

5.4 Approach: SgatSDE

We aim to model the evolving relationships between interacting objects by analyzing noisy and partial video frames. This necessitates a robust modelling framework that can handle the inherent stochasticity of the task. Following the work on NeuralSDE, we propose an approach to learning the stochastic differential equation that governs the evolution of relationship representations. Specifically, similar to our approach SgatODE, we cast the anticipation problem as an instance of the SDE Initial Value Problem with the initial condition as a latent feature corresponding to the last observed subject-object relationship representation.

$$ \mathbf{z}_{so}(t) = \mathbf{z}_{so}(0) + \int_{0}^{t} \mu_{\theta}(\mathbf{z}(s))\mathrm{d}s + \int_{0}^{t} \sigma_{\phi}(\mathbf{z}(s))\mathrm{d}\mathbf{W}(s) \tag{11} $$

Here, $\mu_{\theta}(\mathbf{z}(t))$ and $\sigma_{\phi}(\mathbf{z}(t))$ represent drift and diffusion terms parameterized by neural networks. We intend to learn these parameters from the data. We employ the same learning pipeline as SgatODE, where we first use a Temporal Encoder to extract temporally aware relationship representations that are (1) used to predict relationships in the observed frames and (2) used as initial values for the proposed SDE initial value problem. Figure 3 presents the architecture proposed for the SgatODE and SgatSDE methods.

Anticipation Loss. We use the temporal encoder's output $\mathbf{z}_{so}^0$ as the initial value for the proposed SDE Initial Value Problem. We use Stratonovich interpretation of the Stochastic Differential Equation and use the Reversible Heun solver to extrapolate the relationship representation $\{\mathbf{z}_{so}^t\}_{t=0}^H$. We feed the anticipated relationship representations $\{\mathbf{z}_{so}^t\}_{t=0}^H$ to three classification heads, and we apply multi-label margin loss for the anticipated relationship representations.

$$\mathcal{L}_{\text{ant}}^{\text{sde}} = \sum_{t=0}^{H} \sum_{so} \mathcal{L}_{p_{so}^t} \tag{12}$$

We also employ a decoder similar to SgatODE (see 5.3).

Loss Function. Thus the total objective for the proposed method SgatSDE can be written as:

$$\mathcal{L}_{\text{SgatSDE}} = \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{ant}}^{\text{sde}} + \lambda \mathcal{L}_{\text{boxes}} + \sum_{t=-T}^{0} \mathcal{L}_{(o)}^t \tag{13}$$

6 Experiments

Dataset. Similar to the VidSGG methods, we chose to train and evaluate our models on the Action Genome benchmark [1] dataset. As we are interested in anticipating future relationships between observed objects, we pre-process the data and filter out the videos that possess less than 3 annotated frames. Thus, we obtained 11.4K videos in total; we adhered to the train and test split provided by the dataset. The dataset encompasses 35 object classes and 25 relationship classes. These 25 relationship classes are grouped into three categories, namely: (1) Attention relations primarily describe the attention of the subject towards the object, (2) Spatial relations describe the spatial relationship between two objects, and (3) Contacting relations indicate different ways the object is contacted. We note that, in AG, for a subject-object pair, there can be multiple spatial and contacting relationships.

Evaluation Metrics. Following ImSGG and VidSGG literature, we formulate three tasks for the evaluation of the proposed approaches. We design tasks by varying the amount of information provided to the model both during training and testing to anticipate future relationships of the observed objects. We name them based on the information they provide to the model and call (1) **Frame-Based Anticipation (FBAT)**: When we provide only observed frames as input to the model to anticipate future relationships (2) **Localized Frame-Based Anticipation (LFBAT)**: When we provide frames along with bounding box information of the objects present in the frames (3) **Labelled Localized Frame-Based Anticipation (LLFBAT)**: When we provide frames, bounding box information of the objects and also the object labels. We use the standard Recall@K where $K \in \{10, 20, 50\}$ metric to evaluate the anticipated relationships of the observed objects. For each of these tasks, we evaluate our model's performance by varying the initial fraction $\mathcal{F}$ [0.3,0.5,0.7,0.9] of video frames supplied as context to our model.

Implementation details. For the spatial and temporal encoder architecture, we use the same number of layers and attention heads as in [34]. The loss scaling factors for the SgatODE and SgatSDE models are kept the same. The weight for the bounding box loss, λ, is set to 0.1. The initial learning rate for all models is set to 10^{-5}.

6.1 Comparision

Recall. The results demonstrate the superior performance of our proposed models, SgatODE and SgatSDE, in comparison to the transformer-based baseline, SgatTran, across all three tasks. In the FBAT task, With Constraint setting, SgatODE and SgatSDE outperform SgatTran by a minimum of $\sim 27\%$ and $\sim 38\%$ in recall@10 as shown in Table 1. Similarly, they exhibit at least $\sim 47\%$ and $\sim 64\%$ better performance for LFBAT task, shown in Table 2 and at least $\sim 8\%$ and $\sim 17\%$ improvement for LLFBAT tasks, shown in Table 3 respectively. This trend persists in the No Constraint setting also, for all tasks with all fractions, with SgatSDE achieving the highest performance. Please note that there is no distinction in the LLFBAT-R@20 and R@50 values under the **With Constraint** setting due to the restricted number of objects and constraints on the number of relation edges. The same observation applies to the LFBAT-R@20 and R@50 values.

Mean Recall. The long-tailed distribution of relationships in the training set [12] can result in the generation of biased scene graphs. While the Recall@K metrics can be dominated by the performance on more common relationships,

Table 1: **Anticipation Results** for FBAT task.

FBAT		With Constraint			No Constraint		
$\mathcal{F}$	Method	R@10	R@20	R@50	R@10	R@20	R@50
0.3	SgatTran	18.1	22.4	23.0	17.2	28.1	41.9
	SgatODE	23.1	29.2	31.4	23.3	32.5	45.1
	SgatSDE	**25.0**	**31.7**	**34.3**	**25.9**	**35.0**	**47.4**
0.5	SgatTran	19.4	24.2	24.9	18.5	30.1	44.9
	SgatODE	25.9	32.6	34.8	26.4	36.6	49.8
	SgatSDE	**27.3**	**34.8**	**37.0**	**28.4**	**38.6**	**51.4**
0.7	SgatTran	21.2	26.5	27.2	19.8	32.9	49.5
	SgatODE	30.3	36.6	38.9	32.1	42.8	55.6
	SgatSDE	**31.4**	**38.0**	**40.5**	**33.3**	**44.0**	**56.4**
0.9	SgatTran	22.7	28.7	29.7	20.8	34.4	53.2
	SgatODE	33.9	40.4	42.6	36.6	48.3	61.3
	SgatSDE	**34.8**	**41.9**	**44.1**	**37.3**	**48.6**	**61.6**

Table 2: **Anticipation Results** for LFBAT task.

LFBAT		With Constraint			No Constraint		
$\mathcal{F}$	Method	R@10	R@20	R@50	R@10	R@20	R@50
0.3	SgatTran	16.9	17.1	17.1	25.1	34.0	37.5
	SgatODE	25.1	26.1	26.1	30.5	39.3	46.3
	SgatSDE	**28.9**	**29.9**	**29.9**	**35.3**	**42.8**	**46.8**
0.5	SgatTran	18.9	19.1	19.1	28.0	37.9	41.6
	SgatODE	29.0	30.0	30.0	35.0	44.5	51.5
	SgatSDE	**32.2**	**33.3**	**33.3**	**39.0**	**47.5**	**52.3**
0.7	SgatTran	22.6	22.8	22.8	33.6	45.2	48.9
	SgatODE	36.2	37.0	37.0	44.0	54.1	60.0
	SgatSDE	**38.5**	**39.4**	**39.4**	**46.4**	**55.8**	**60.3**
0.9	SgatTran	26.2	26.4	26.4	38.6	50.3	54.0
	SgatODE	42.4	43.1	43.1	53.1	62.3	66.6
	SgatSDE	**43.6**	**44.2**	**44.2**	**54.2**	**62.7**	**66.7**

Table 3: **Anticipation Results** for LLFBAT task.

LLFBAT		With Constraint			No Constraint		
$\mathcal{F}$	Method	R@10	R@20	R@50	R@10	R@20	R@50
0.3	SgatTran	31.8	33.2	33.2	39.1	56.5	64.9
	SgatODE	34.9	37.3	37.3	40.5	54.1	63.9
	SgatSDE	**39.7**	**42.2**	**42.3**	**46.9**	**59.1**	**65.2**
0.5	SgatTran	37.4	39.0	39.0	43.9	63.6	73.0
	SgatODE	40.7	43.4	43.4	47.0	62.2	72.4
	SgatSDE	**45.0**	**47.7**	**47.7**	**52.5**	**66.4**	**73.5**
0.7	SgatTran	44.3	45.9	45.9	51.2	72.8	82.8
	SgatODE	49.1	51.6	51.6	58.0	74.0	82.8
	SgatSDE	**52.0**	**54.5**	**54.5**	**61.8**	**76.7**	**83.4**
0.9	SgatTran	50.8	52.1	52.1	61.0	83.5	92.8
	SgatODE	58.1	59.8	59.8	72.6	86.7	93.2
	SgatSDE	**60.3**	**61.9**	**61.9**	**74.8**	**88.0**	**93.5**

the mean recall metric introduced in [35] is a more balanced metric that scores the generalisation of the model to all predicate classes. Again, a similar trend emerges, as shown in Table 4, SgatSDE outperforms SgatODE and the baseline on all tasks. In the With Constraint setting, for the FBAT task, SgatSDE shows an improvement of $\sim 60\%$ over the baseline and $\sim 9\%$ over the SgatODE model in the mR@10 metric. On the same metric, the improvement jumps to about $\sim 200\%$ and $\sim 10\%$ for the LFBAT task and $\sim 100\%$ and $\sim 15\%$ for the LLFBAT task, over the baseline and SgatODE models respectively. The consistent gap between the SgatODE and SgatSDE model confirms our hypothesis that modelling the noise explicitly results in better generalisation for relationship prediction for both head and tail classes. The qualitative results for both models are shown in Figure 4.

6.2 Conclusion

We introduced a novel task, Scene Graph Anticipation (SGA) and proposed approaches that model the latent dynamics of the evolution of relationship representations between interacting objects. Our work opens up several avenues for future work. First, an exciting direction is to include localization into the framework. Second, SGA as a tool can be used to develop methods for Surveillance Systems, Robotics etc.

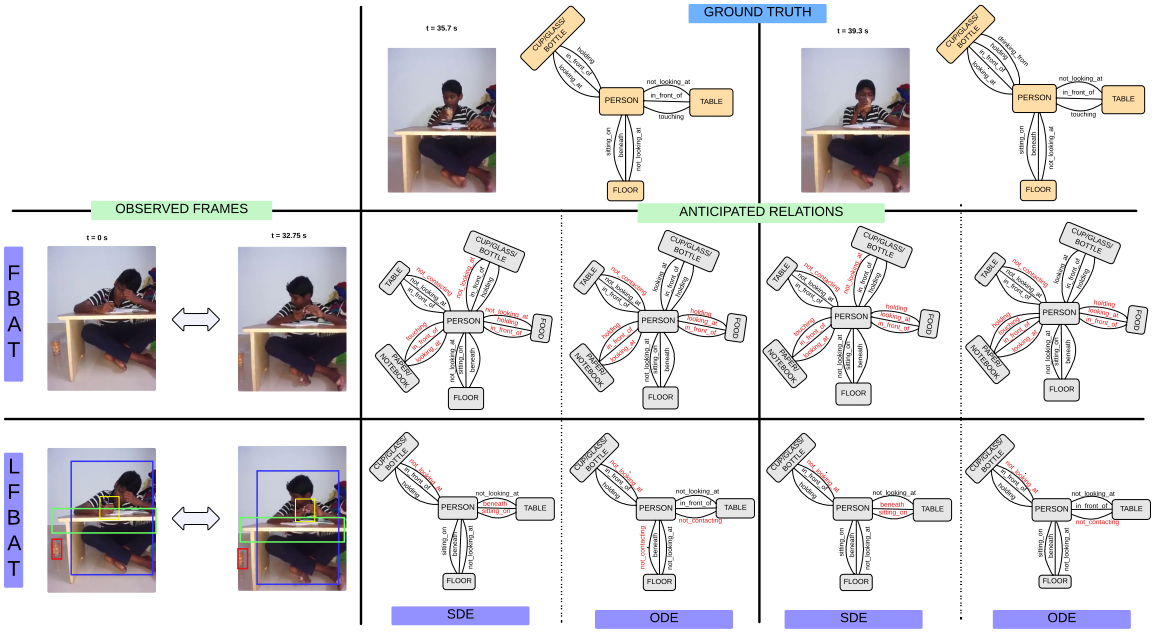

Figure 4: The top row's columns display the ground truth image and scene graph of future frames. In the following two rows, the first two columns depict the observed frames for the FBAT and LFBAT tasks, respectively. The subsequent columns showcase the anticipated relations generated by the SgatSDE and SgatODE models in their corresponding positions. The black-colored relations depict correctly anticipated relations by the model, while the red-colored relations highlight instances where the model deviates.

Table 4: With Constraint Mean Recall anticipation results for tasks FBAT, LFBAT, LLFBAT

With Constraint		FBAT			LFBAT			LLFBAT		
Context	**Method**	**mR@10**	**mR@20**	**mR@50**	**mR@10**	**mR@20**	**mR@50**	**mR@10**	**mR@20**	**mR@50**
0.3	Baseline	6.81	8.15	8.33	4.69	4.78	4.78	8.67	9.2	9.21
	SgatODE	10.55	13.75	15.02	12.92	13.38	13.38	15.1	16.59	16.61
	SgatSDE	**11.42**	**15.3**	**16.92**	**14.21**	**14.74**	**14.74**	**18.35**	**20.51**	**20.54**
0.5	Baseline	7.28	8.78	9.0	5.27	5.36	5.36	10.07	10.68	10.68
	SgatODE	11.59	15.19	16.41	14.93	15.45	15.46	17.39	19.24	19.31
	SgatSDE	**12.39**	**16.61**	**17.96**	**15.82**	**16.56**	**16.57**	**20.74**	**23.01**	**23.08**
0.7	Baseline	7.81	9.49	9.71	6.06	6.15	6.15	11.86	12.47	12.47
	SgatODE	12.8	16.39	17.8	16.9	17.33	17.33	20.96	22.89	22.94
	SgatSDE	**13.82**	**17.74**	**19.32**	**18.35**	**19.05**	**19.05**	**24.08**	**26.47**	**26.47**
0.9	Baseline	8.23	10.08	10.39	6.83	6.96	6.96	13.76	14.3	14.31
	SgatODE	14.01	18.13	19.31	18.97	19.37	19.37	24.96	26.37	26.37
	SgatSDE	**15.14**	**19.44**	**21.03**	**20.65**	**21.07**	**21.07**	**28.47**	**29.8**	**29.81**

Acknowledgments

References

[1] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, Fei-Fei Li, and Juan Carlos Niebles. Action genome: Actions as composition of spatio-temporal scene graphs. *arXiv: Computer Vision and Pattern Recognition*, 2019.

[2] Yuren Cong, Wentong Liao, H. Ackermann, M. Yang, and B. Rosenhahn. Spatial-temporal transformer for dynamic scene graph generation. *IEEE International Conference on Computer Vision*, 2021.

[3] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *Neural Information Processing Systems*, 2018.

[4] Patrick Kidger, James Foster, Xuechen Li, and Terry Lyons. Efficient and accurate gradients for neural sdes, 2021.

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.

[6] Ue-Hwan Kim, Jin-Man Park, Taek jin Song, and Jong-Hwan Kim. 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE Transactions on Cybernetics*, 50:4921–4933, 2019.

[7] Zuyao Chen, Jinlin Wu, Zhen Lei, Zhaoxiang Zhang, and Changwen Chen. Expanding scene graph boundaries: Fully open-vocabulary scene graph generation via visual-concept alignment and retention, 2023.

[8] Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. Llm4sgg: Large language model for weakly supervised scene graph generation, 2023.

[9] Zijian Zhou, Miaojing Shi, and Holger Caesar. Vlprompt: Vision-language prompting for panoptic scene graph generation, 2023.

[10] Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, and Mohammed Bennamoun. Scene graph generation: A comprehensive survey, 2022.

[11] Suprosanna Shit, Rajat Koner, Bastian Wittmann, Johannes Paetzold, Ivan Ezhov, Hongwei Li, Jiazhen Pan, Sahand Sharifzadeh, Georgios Kaissis, Volker Tresp, and Bjoern Menze. Relationformer: A unified framework for image-to-graph generation, 2022.

[12] Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K Roy-Chowdhury. Unbiased scene graph generation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22803–22813, 2023.

[13] Anant Khandelwal. Correlation debiasing for unbiased scene graph generation in videos, 2023.

[14] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1300–1308, New York, NY, USA, 2017. Association for Computing Machinery.

[15] Xinxiao Wu, Jianwei Zhao, and Ruiqi Wang. Anticipating future relations via graph growing for action prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):2952–2960, May 2021.

[16] Li Mi, Yangjun Ou, and Zhenzhong Chen. Visual relationship forecasting in videos. *arXiv.org*, 2021.

[17] Alex X. Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction, 2018.

[18] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders, 2016.

[19] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction, 2016.

[20] Wei Yu, Wenxin Chen, Songhenh Yin, Steve Easterbrook, and Animesh Garg. Modular action concept grounding in semantic video prediction, 2022.

[21] Haozhi Qi, Xiaolong Wang, Deepak Pathak, Yi Ma, and Jitendra Malik. Learning long-term visual dynamics with region proposal interaction networks, 2021.

[22] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models, 2023.

[23] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling, 2022.

[24] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation, 2022.

[25] Yuxuan Liang, Kun Ouyang, Hanshu Yan, Yiwei Wang, Zekun Tong, and Roger Zimmermann. Modeling trajectories with neural ordinary differential equations. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1498–1504. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[26] Zibo Liu, Parshin Shojaee, and Chandan K Reddy. Graph-based multi-ode neural networks for spatio-temporal traffic forecasting, 2023.

[27] Sunghyun Park, Kangyeol Kim, Junsoo Lee, Jaegul Choo, Joonseok Lee, Sookyung Kim, and Edward Choi. Vid-ode: Continuous-time video generation with neural ordinary differential equation, 2021.

[28] Zijie Huang, Yizhou Sun, and Wei Wang. Learning continuous system dynamics from irregularly-sampled partial observations, 2020.

[29] Michael Poli, Stefano Massaroli, Junyoung Park, Atsushi Yamashita, Hajime Asama, and Jinkyoo Park. Graph neural ordinary differential equations, 2021.

[30] Zijie Huang, Yizhou Sun, and Wei Wang. Coupled graph ode for learning interacting system dynamics. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 705–715, New York, NY, USA, 2021. Association for Computing Machinery.

[31] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. *Computer Vision and Pattern Recognition*, 2021.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[33] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[34] Shengyu Feng, Hesham Mostafa, Marcel Nassar, Somdeb Majumdar, and Subarna Tripathi. Exploiting long-term dependencies for generating dynamic scene graphs. *IEEE Workshop/Winter Conference on Applications of Computer Vision*, 2021.

[35] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation, 2019.