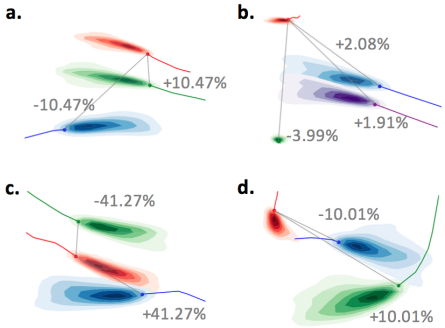
R1, R2: Technical novelty. While our work builds on top of previously proposed methods including GAN, attention mechanisms, and GAT, their adaptation, integration and application to the complex problem of human trajectory prediction is a non-trivial technical challenge. Our main technical contribution is a novel neural model that 1) better encodes social cues (a crucial factor to predict human trajectories) based on neural graph structures, and 2) enables multimodal predictions (an intrinsic property of future human motion) inspired by BicycleGAN architecture. Learning an optimal representation of social behavior is a non-trivial task, as humans follow unwritten social rules and juggle a variety of implicit factors. As we discuss in our literature review and experimental evaluation, prior works that model social interactions fall short by either limiting the expressiveness of the models (eg: by using a pooling mechanism that is unable to capture interactions in large scenes), or by imposing human-defined constraints rather than learning from the data. Further, applying a multi-modal distribution to represent future human trajectories is not easy since it requires solving a delicate trade-off between increasing variance of each model vs. increasing multimodality, which relates to the degradation of the model's predictions as a function of allowed samples.

R1, R3: Analysis on the attention weights. As the reviewers suggested, we have run several experiments exploring the correlation between the weights and different pedestrian features. The figure on the right depicts the attention weights of other pedestrians wrt. the red pedestrian for different scenes. Weight values correspond to the percent decrease/increase compared to without attention: positive indicates more attention was paid to that interaction, and negative indicates less attention was paid. From Scenes (a, b) we can infer that Euclidean distance is one feature the network implicitly uses to assign attention. Scenes (c) and (d) show however that the attention further generalizes in learning which agents are important socially: in Scene (c) it pays large attention to the blue agent it may collide with in the future, even though that agent is farther away from it than the green one, and in Scene (d) it ignores the blue agent for the farther green agent with whom it might collide with. The diversity of social awareness that the model displays validates the choice of GAT over prior works. We will include this figure and a visualization of physical attention to the final manuscript, with more samples in the supplementary material.



R1: Improved generalization claim. We thank R1 for their suggestion and have updated the table as shown below. We see that the GAT architecture initially performs better than the BiGAN, but with fewer samples the GAT error increases faster. This aligns with our intuition that the inclusion of the BiGAN in our architecture enables for better capturing of a multimodal distribution, instead of generating samples from a unimodal distribution. We note that BiGAN's error increases slightly slower than Social-BiGAT, but the inclusion of the GAT allows for a lower overall error.

R1: Comparison with SOTA baselines. Yes, Sophie was the state of the art in terms of ADE/FDE at the time of submission.

Model	K=20	K=10	K=5	K=1	% Increase
GAT	0.518 / 1.064	0.529 / 1.127	0.584 / 1.241	0.682 / 1.494	31.6% / 40.4%
BiGAN	0.523 / 1.091	0.531 / 1.144	0.579 / 1.298	0.662 / 1.439	**26.6% / 31.9%**
Social-BiGAT	**0.476 / 0.998**	**0.488 / 1.096**	**0.527 / 1.260**	**0.606 / 1.328**	**27.3% / 33.1%**

R1: Clarifying pooling for joint feature representation. In architectures that use pooling, each pedestrian is encoded into a single vector. Across all pedestrians the vectors are pooled, resulting in a single vector representing the social interactions between all agents in the scene (not specific to any particular agent), which each pedestrian receives. Our choice of GAT serves as a major improvement as it allows for different interactions to be attended to based on their social importance, which the network learns.

R2: Inputting top-view image instead of image sequences. Thanks to the reviewer for pointing out the confusion, we will clarify this in the paper. However, we would like to point out that most prior research/datasets in this field are using a single top-down image input, such as CAR-Net, Desire, and Sophie, and the more recent Social Ways, and Multi Agent Tensor Flow. Moreover, our approach has also been tested on sequences where the scene's angle view is not necessarily perfectly top-view, e.g. UCY dataset. The criticism of not accepting a sequences of images is valid, as we only use a single image. This is done for fair comparison, as prior methods also only use single images. Finally, our work is indeed limited by requiring temporal sequences of pedestrian locations. However, while we agree that performing end-to-end tracking and forecasting directly on sequences of images is a very good future direction , we are currently limited by a lack of proper benchmark data. All existing benchmarks follow the same input modality, and we do the same to make the comparison feasible. In the future we definitely hope to research this same problem end-to-end.

R2, R3: Lack of clarity in notations, architecture. We will clarify the equations and better explain the variables we introduce before using them, along with labelling weights in Fig. 2, which should improve the architecture explanation. We also will add an in-depth ablative analysis and explanation of model components in the supplementary material.

R2: Better visualizations & demos. As suggested by R2, we have already started implementing it by adding several examples to the supplementary work, where we can visualize multiple dimensions of the latent space across multiple pages and better illustrate how our model qualitatively results in multimodality. We will also add figures exploring how the latent noise generated through the encoder changes when varying parts of the scene (such as number of agents, or speed and direction of agents). We also are looking forward to publishing demo videos that we can place in the supplementary material as well. Space permitting, we will add these figures to the main paper as well.