VARSCENE: A Deep Generative Model for Realistic Scene Graph Synthesis

Tathagat Verma

Joint work with

Abir De, Yateesh Agrawal,
Vishwa Vinay and Soumen Chakrabarti

Supported by an Adobe Research grant

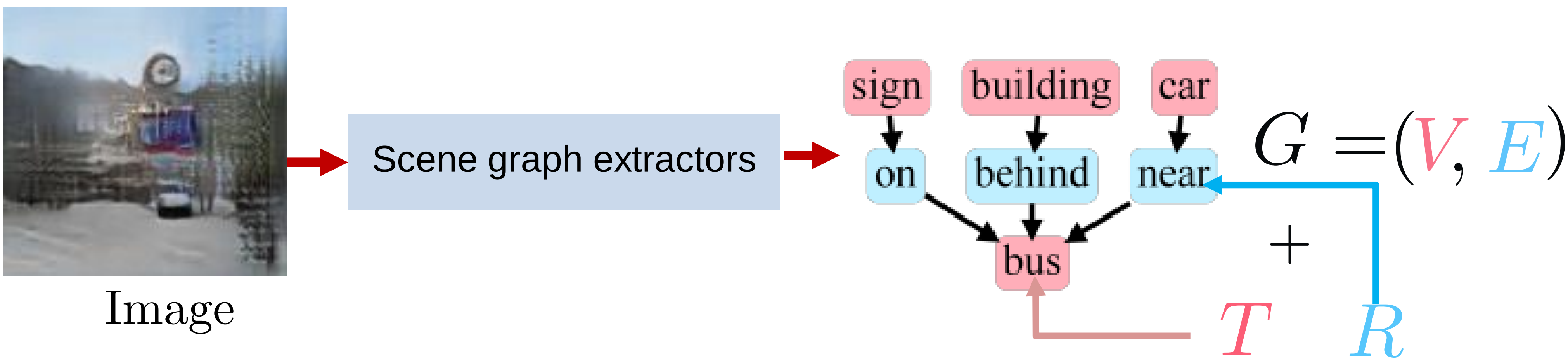
Our work

Generate scene graphs from scene graphs (not from images)

- **Preserve semantic relationships between objects**
- **Wide spectrum of applications:**
  - **Structured query based Image retrieval** (Schnoder et al 2020)
  - **Image editing** (Dhamo et al 2020)
  - **Image captioning** (Milewski et al 2020)

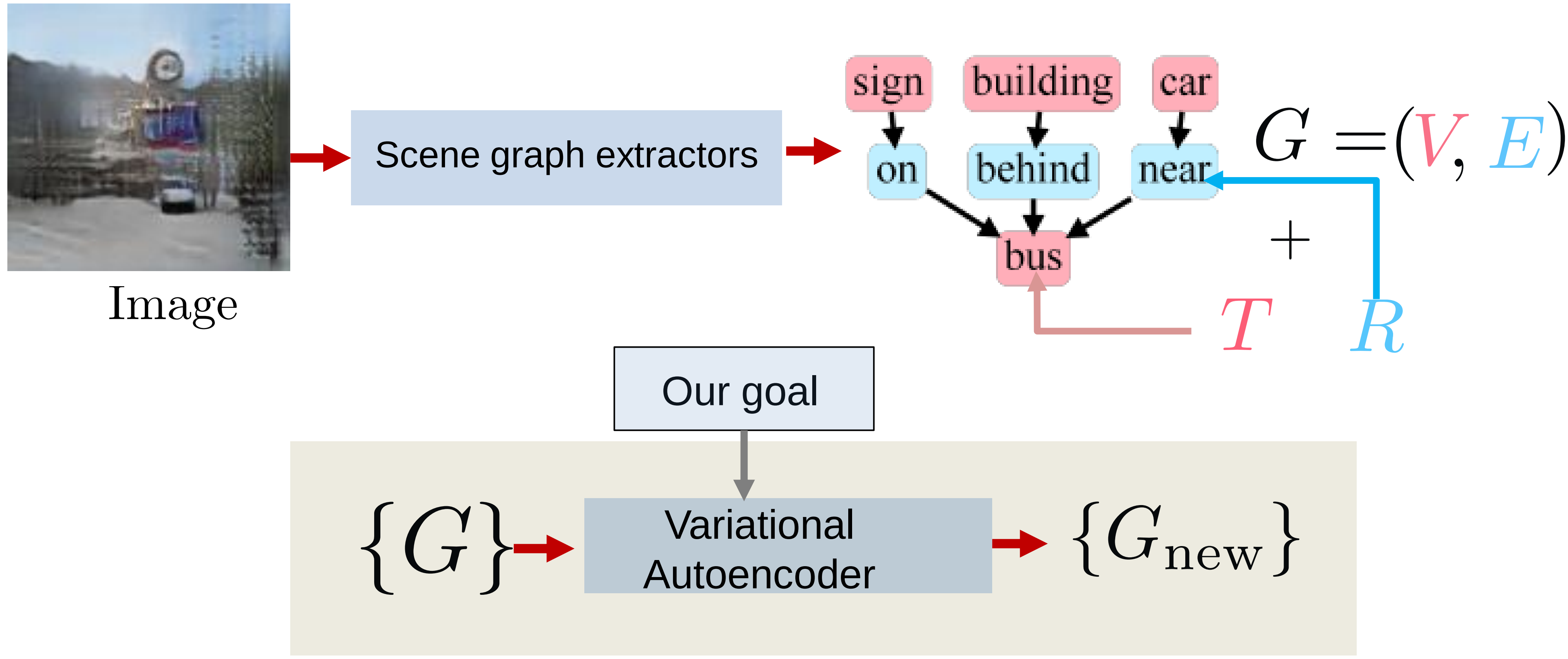
Our work

Generate from scene graphs instead of images



Our work

Generate scene graphs instead of images



VarScene: A generative model for scene graphs

Encoder: Encode **stars around nodes** (instead of nodes) into representations

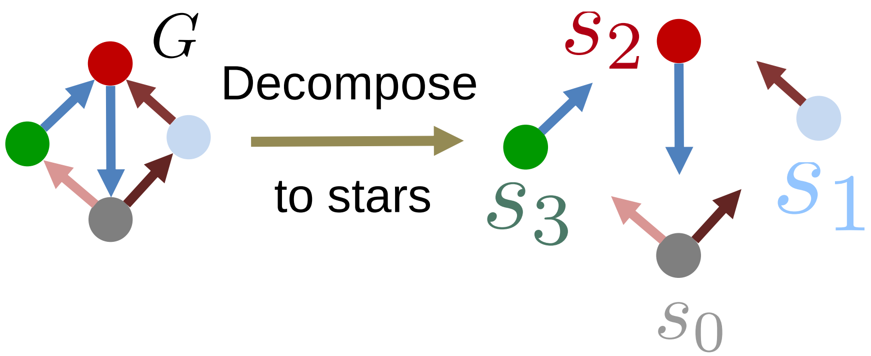
Decoder: **Generates stars** (instead of nodes/edges)

Maintains semantic relationship between nodes and edges

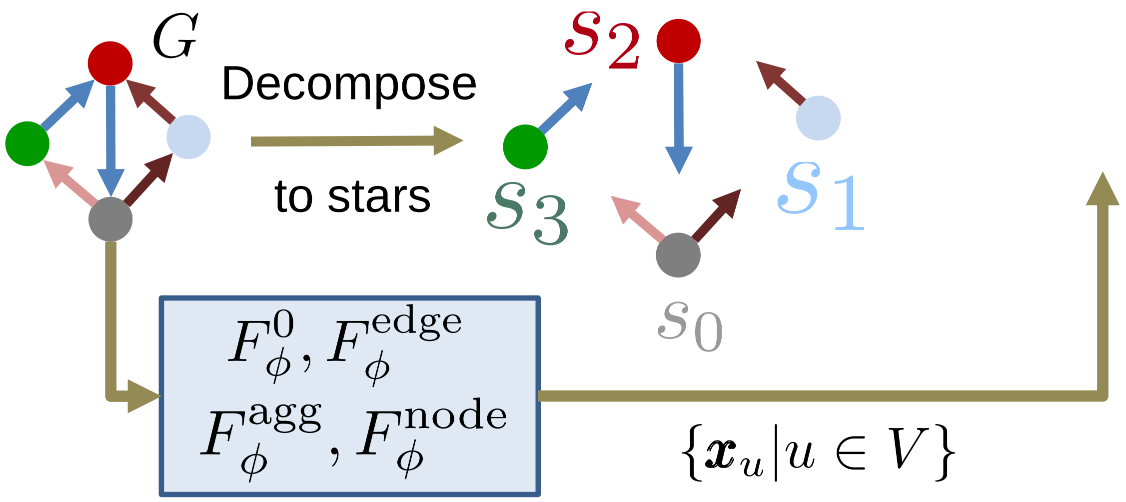
MMD optimized decoder: Final decoder is re-trained to mimic the true distribution

Enhanced generalization

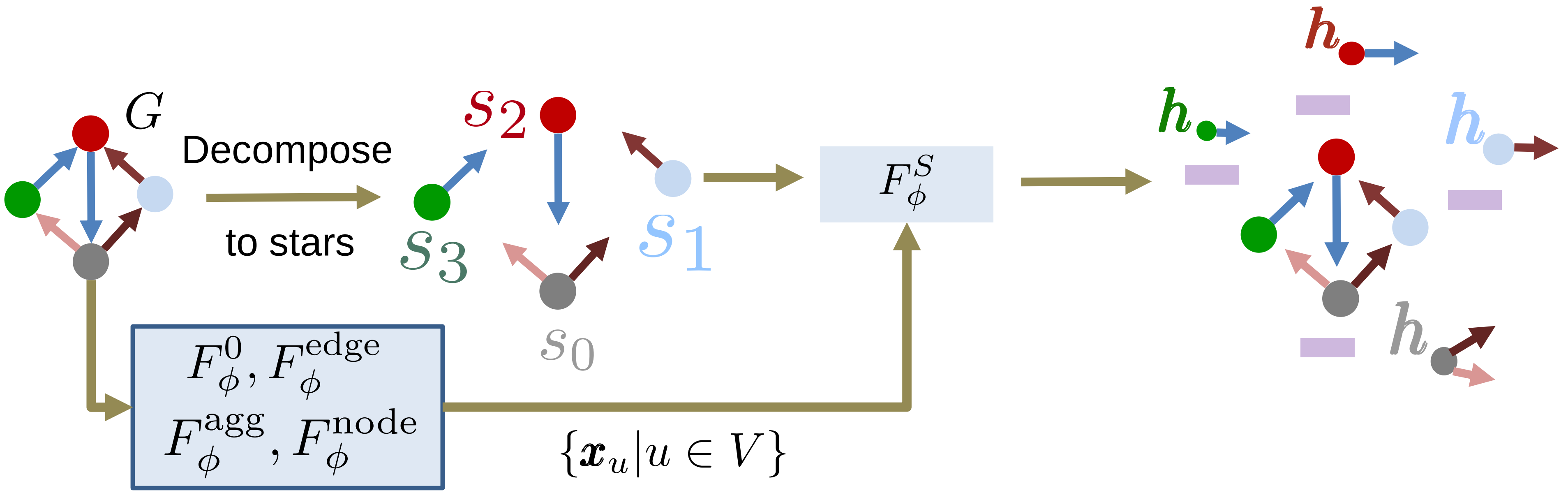
The probabilistic encoder



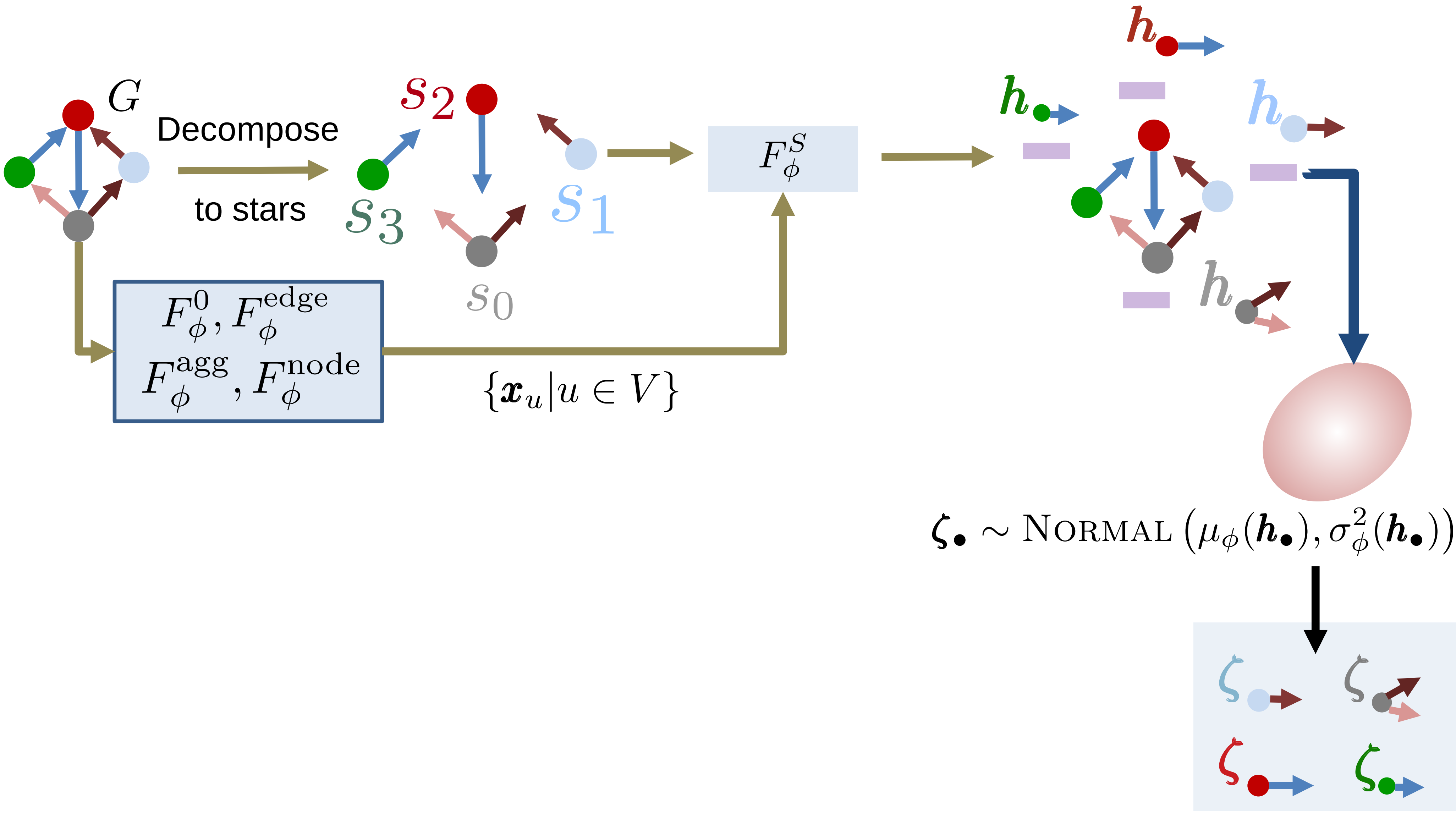
The probabilistic encoder



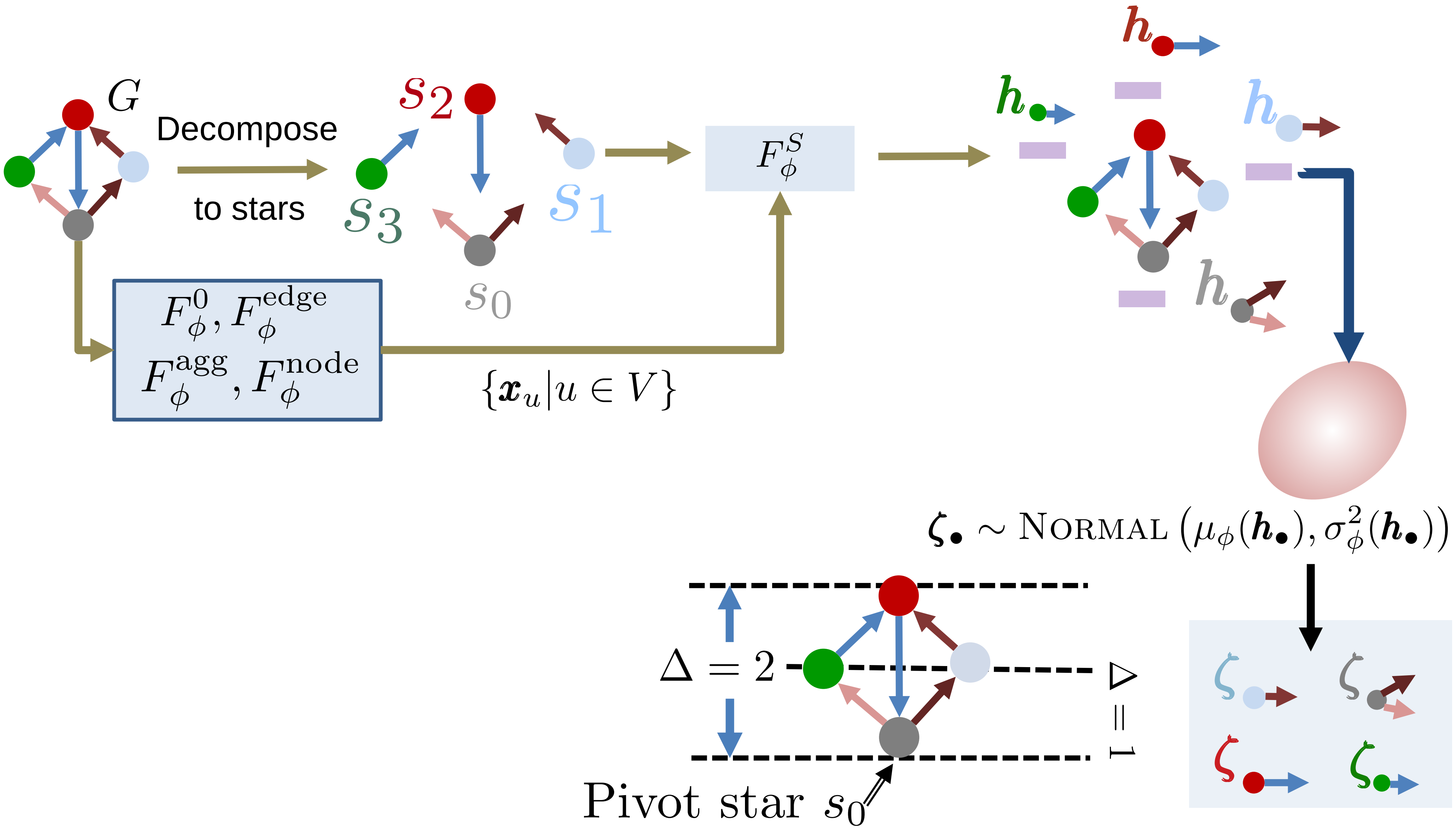
The probabilistic encoder



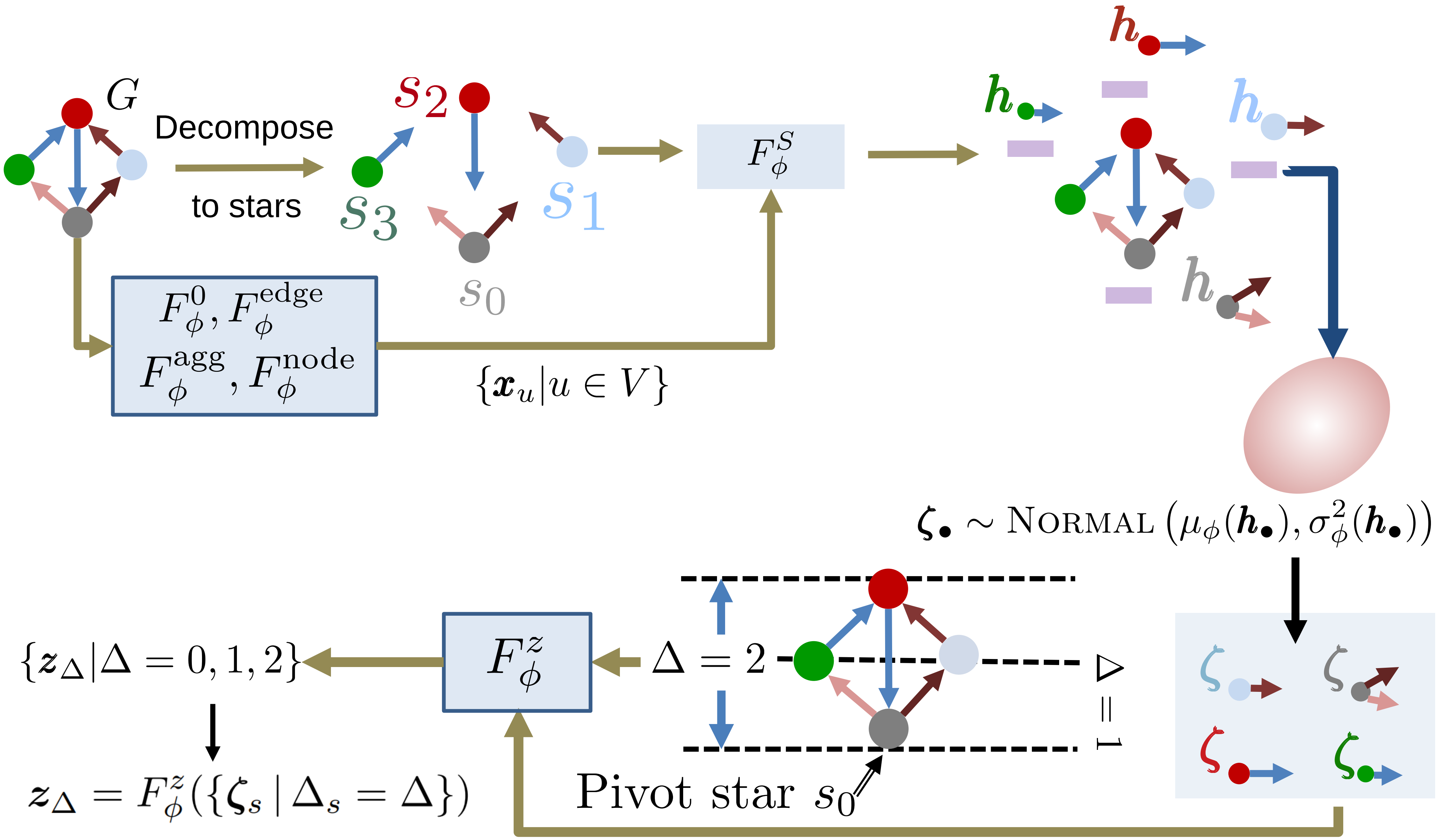
The probabilistic encoder



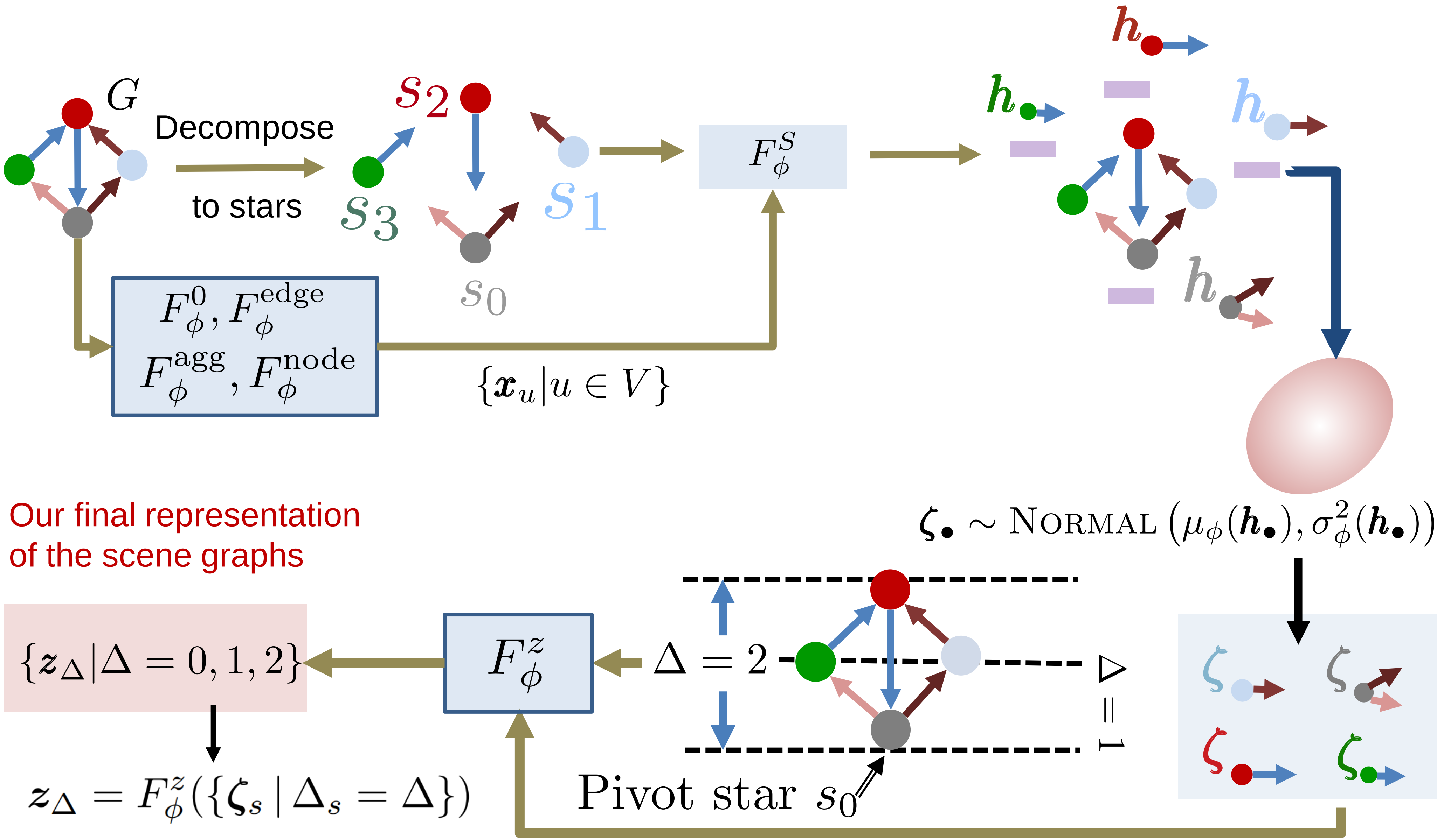
The probabilistic encoder



The probabilistic encoder



The probabilistic encoder

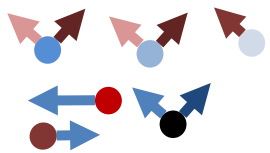


Our final representation of the scene graphs

$\{\boldsymbol{z}_\Delta | \Delta = 0, 1, 2\}$

$\boldsymbol{z}_\Delta = F_\phi^z(\{\boldsymbol{\zeta}_s \,|\, \Delta_s = \Delta\})$

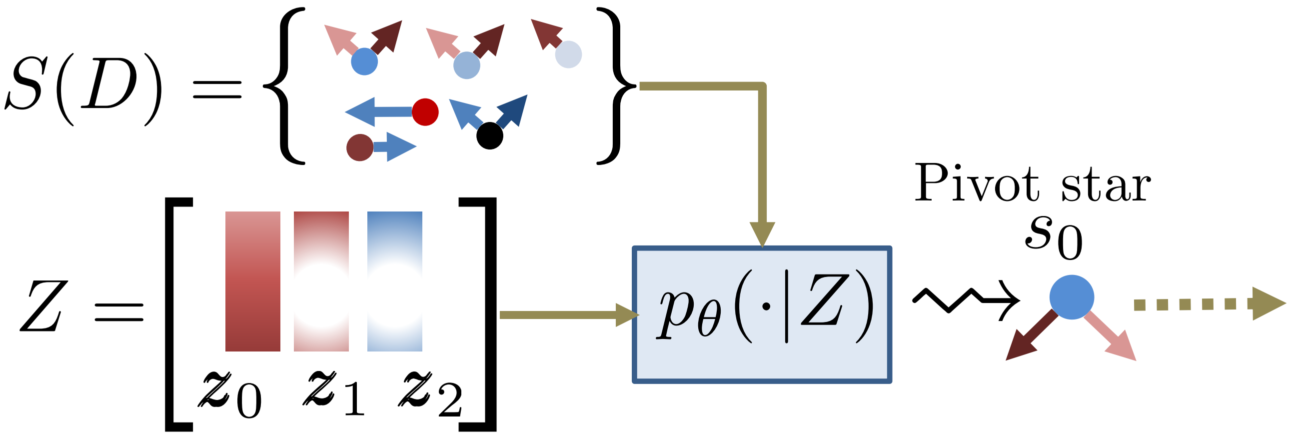
The probabilistic decoder

$S(D) = \{$  $\}$

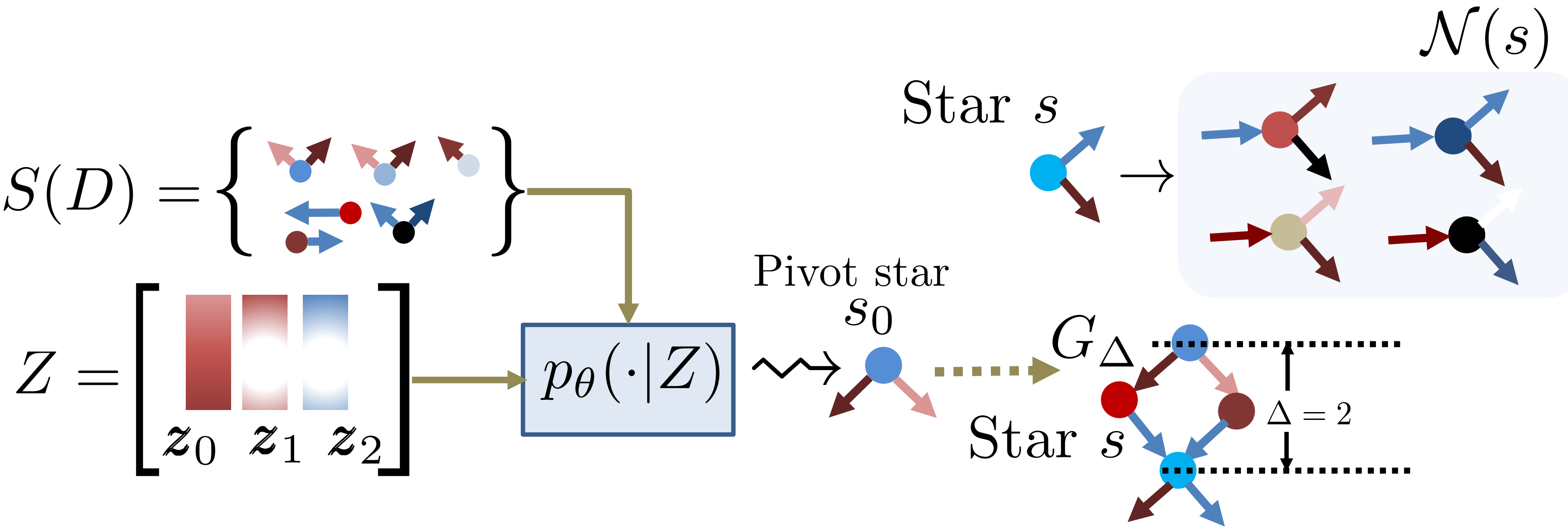
$Z =$ 

$\boldsymbol{z}_0 \quad \boldsymbol{z}_1 \quad \boldsymbol{z}_2$

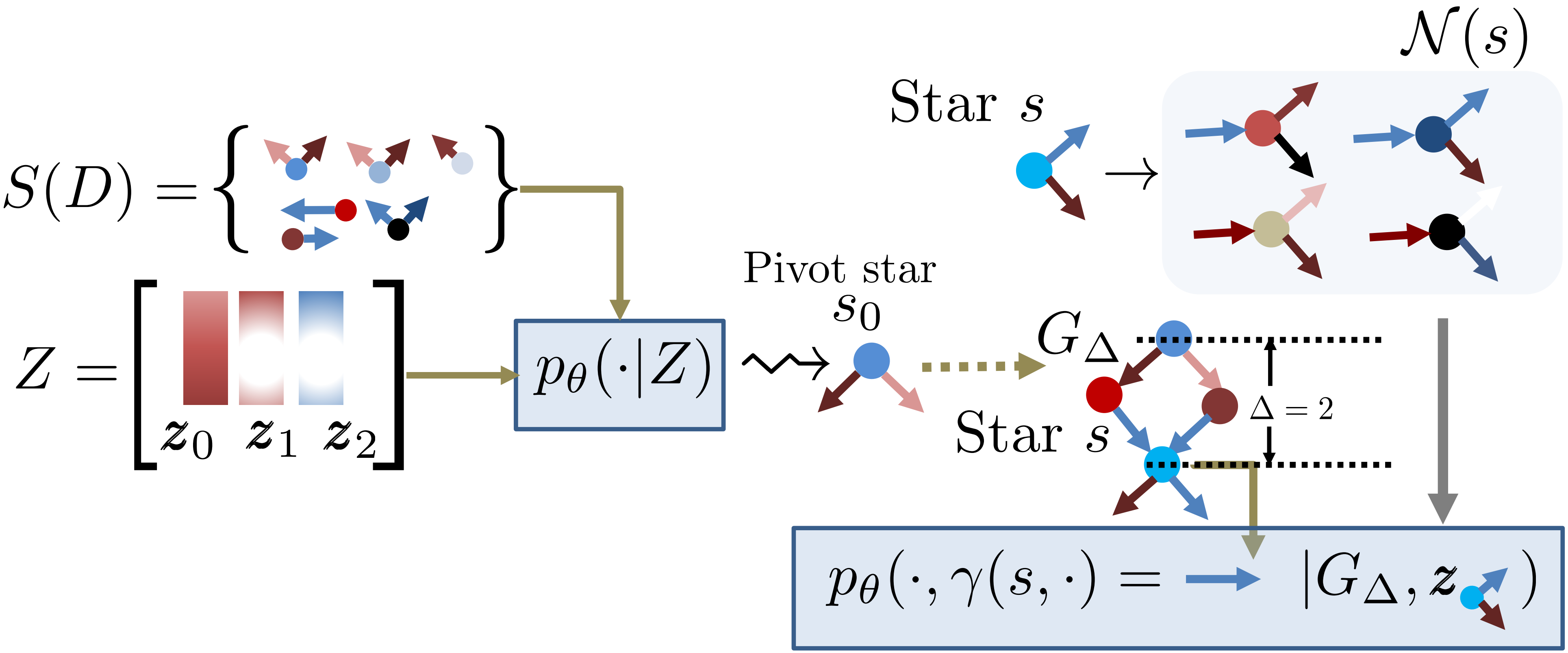
The probabilistic decoder



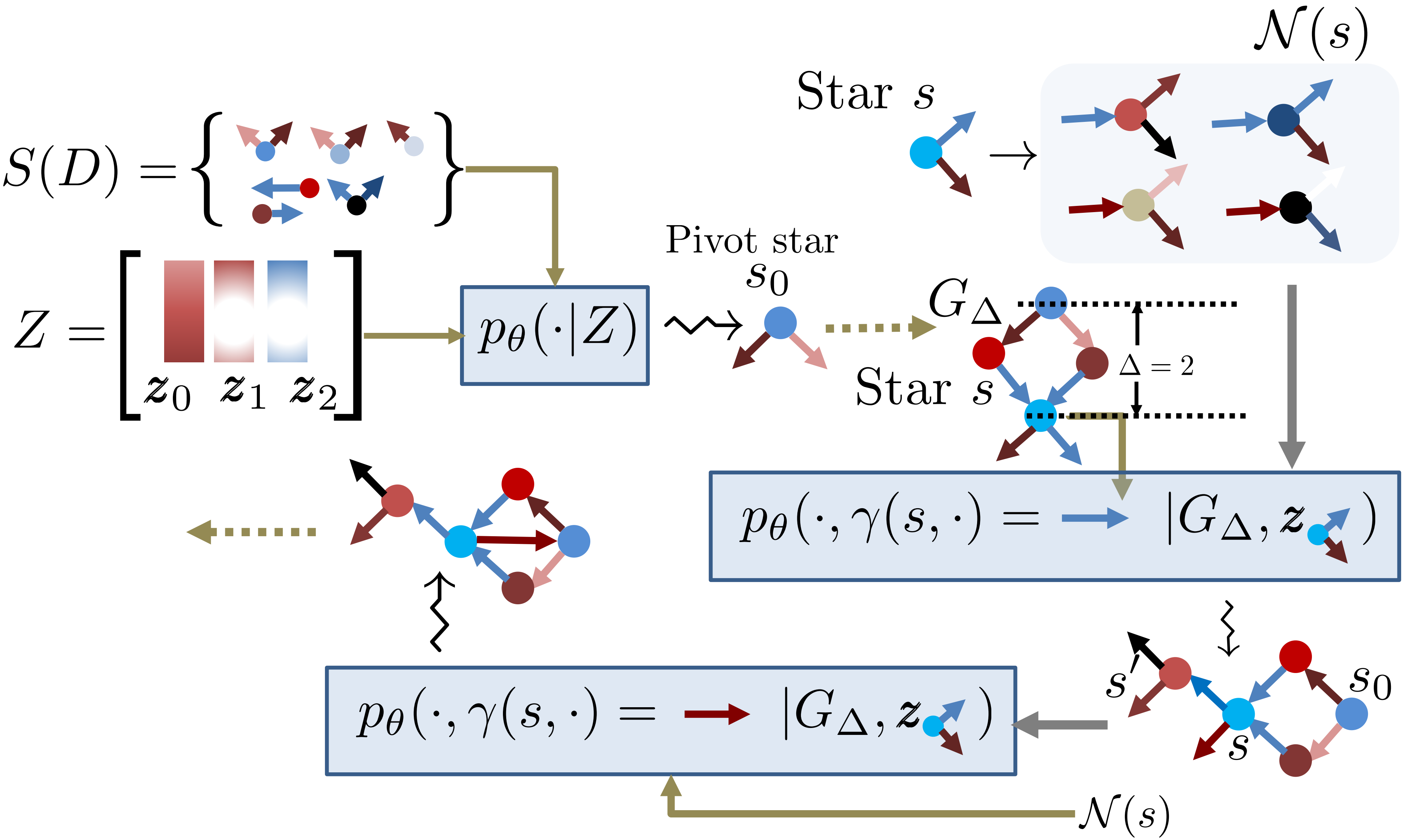
The probabilistic decoder



The probabilistic decoder



The probabilistic decoder



Training the VAE

Maximize ELBO

$$\max_{\phi,\theta} \sum_{G\in D} \Big[\mathbb{E}_{Z\sim q_{\phi}(.|G)}\big[\log p_{\theta}(G|Z)\big] + KL\big(q_{\phi}(Z|G)||p_0(Z)\big)\Big]$$

Training the VAE

Maximize ELBO

$$\max_{\phi,\theta} \sum_{G \in D} \Big[\mathbb{E}_{Z \sim q_{\phi}(.|G)} \big[\log p_{\theta}(G|Z) \big] + KL\big(q_{\phi}(Z|G) || p_0(Z)\big) \Big]$$

Approximate Posterior

Lower bound on the true objective

Training the VAE

Maximize ELBO

$$\max_{\phi,\theta} \sum_{G\in D} \Big[\mathbb{E}_{Z\sim q_\phi(.|G)}\big[\log p_\theta(G|Z)\big] + KL\big(q_\phi(Z|G)||p_0(Z)\big)\Big]$$

Approximate Posterior

Lower bound on the true objective

Training is incognizant to underlying distribution

MMD optimized decoder design

Re-train decoder using the MMD between generated graphs and validation graphs

$$p_{\widehat{\theta}} \overset{\text{re-train}}{\longrightarrow} p_{\theta^{\text{MMD}}}$$

MMD optimized decoder design

Re-train decoder using the MMD between generated graphs and validation graphs

$$p_{\widehat{\theta}} \overset{\text{re-train}}{\longrightarrow} p_{\theta^{\text{MMD}}}$$

Validation set

$$\underset{\theta\in\Theta}{\text{minimize}}\ \mathbb{E}_Z\Bigg[\ \mathop{\mathbb{E}}_{D(Z)\underset{iid}{\sim} p_\theta^{\text{MMD}}(\cdot\mid Z)} \underbrace{\text{MMD}\Big(D_V, D(Z)\Big)}_{\text{Difference between generated and existing graphs}} + \rho\,\text{KL}\Big(p_\theta^{\text{MMD}}(\cdot\mid Z)\Big\| p_{\widehat{\theta}}(\cdot\mid Z)\Big)\Bigg]$$

Already trained decoder

Experimental setup

We use three datasets: **Visual Genome (large and small) and Visual Relationship Detection** datasets.

Five baselines (Scene graph generators are rare in literature)

DeepGMG (Li et al. 2018)

MolGAN (De Cao et al. 2018)

GraphRNN (You et al. 2018)

GraphGen (Goyal et al. 2020)

Generic or molecular graph generators

SceneGen (Garg et al. 2021) **Scene graph generator**

Experimental setup

Reliable measures for evaluating generic graph generators are lacking.

In addition to Kernels, we also use cosine similarities between various quantities:

$$\cos\left(\mathbb{E}_{G' \sim p_\theta^{\mathrm{MMD}}} \nu(G'), \mathbb{E}_{G \sim D_{\mathrm{test}}} \nu(G)\right)$$

Number of different stars

Number of different edge-bigrams # $\langle r_{(\bullet,u)}, t_u, r_{(u,\bullet)} \rangle$

Number of different node-bigrams # $\langle t_u, r_{(u,v)}, t_v \rangle$

Evaluation Metrics for Graph Generative Models: Problems, Pitfalls, and Practical Solutions. O-Bray et al. ICLR 2022

Experimental results

Model	Star-Sim	Edge-sim	Node-sim	SP-K	WL-K	NSPD-K
Visual Genome (VG)						
DeepGMG	0.69	0.46	0.15	0.01	<u>0.09</u>	0.01
MolGAN	0.00	0.00	0.00	0.00	0.04	0.01
GraphGen	0.66	0.37	0.11	0.00	0.03	0.01
GraphRNN	0.63	0.00	0.03	0.00	0.03	0.01
SceneGen	<u>0.73</u>	<u>0.50</u>	0.32	0.02	0.08	0.01
VarScene$^{\text{unc}}$	0.59	0.45	<u>0.40</u>	**0.22**	**0.11**	0.01
VarScene$^{\text{cond}}$	**0.86**	**0.52**	**0.62**	<u>0.08</u>	0.07	0.01

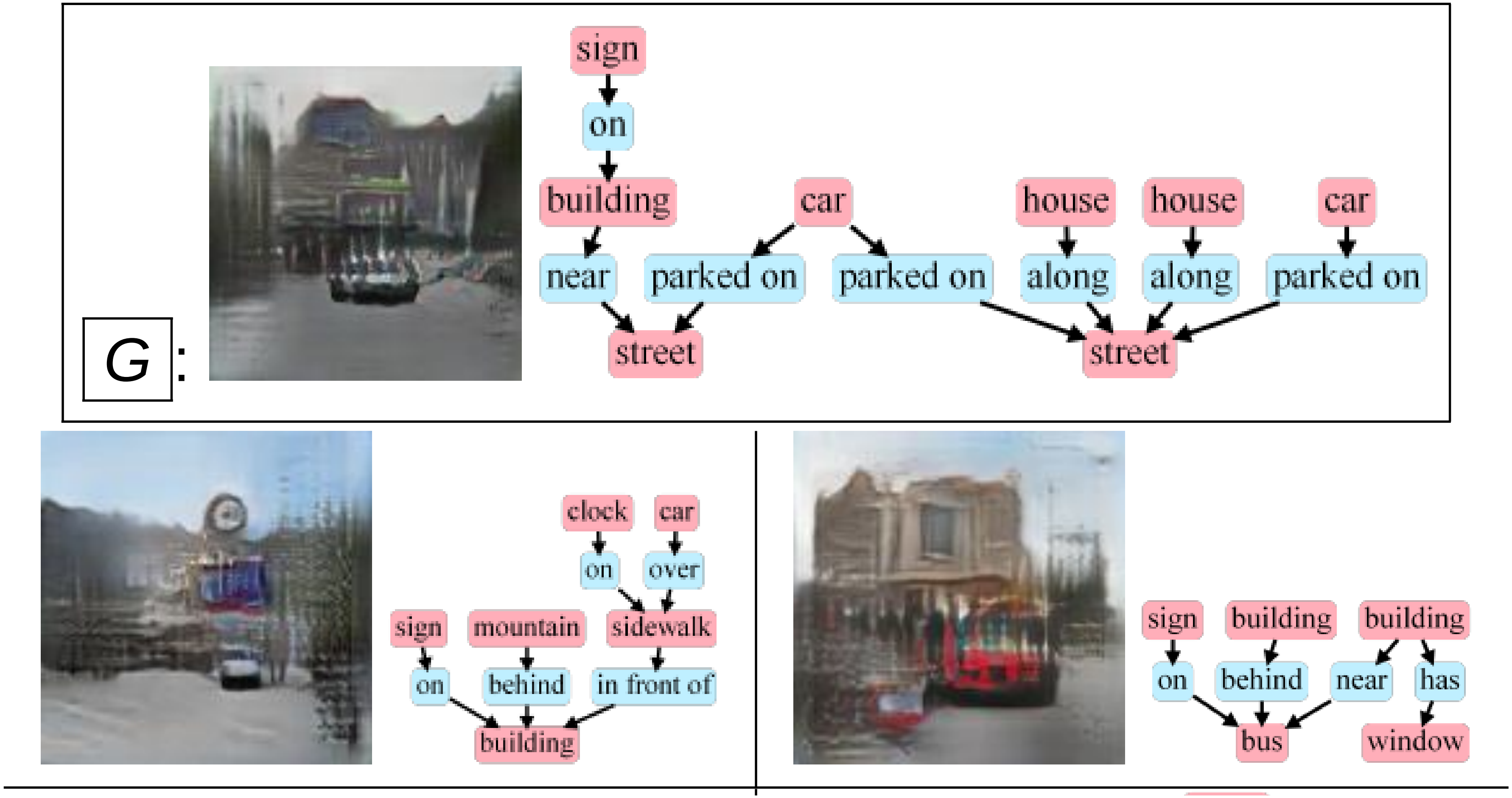
Experimental results

Model	Star-Sim	Edge-sim	Node-sim	SP-K	WL-K	NSPD-K
Visual Relationship Detection (VRD)						
DeepGMG	0.74	0.73	0.60	0.99	1.41	0.20
MolGAN	0.00	0.00	0.00	0.01	0.97	0.21
GraphGen	0.64	0.75	0.64	0.31	0.79	0.17
GraphRNN	0.54	0.29	0.71	0.21	0.76	0.18
SceneGen	0.81	**0.94**	**0.95**	0.60	1.12	0.21
VarScene^{unc}	**0.91**	0.93	0.94	1.03	1.56	**0.23**
VarScene^{cond}	**0.91**	0.93	0.93	**1.45**	**1.92**	0.22

Effect of MMD optimization

	VG		SVG		VRD	
	Star	Edge	Star	Edge	Star	Edge
p_θ^{MMD}	**0.8660**	**0.5268**	**0.9182**	**0.6964**	**0.9140**	**0.9372**
p_θ	0.5867	0.2588	0.7120	0.4195	0.8988	0.9339

Qualitative results



Sg2im: Image Generation from Scene Graphs. Johnson et al. CVPR 2018

Conclusions

We have introduced a variational autoencoder (VAE) for scene graphs which, thanks to several technical innovations, beats the state of the art.

There are many interesting questions for **future work**:

1. Image editing after generating scene graphs

2. Controlling tradeoff between diversity and quality