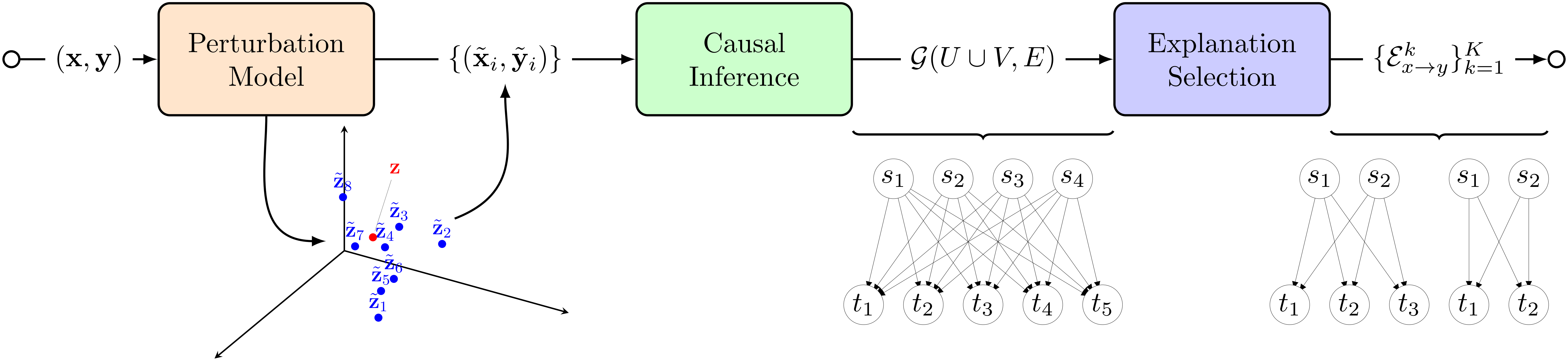
Interpretability for Black-Box Seq2Seq Models

AM+ Jaakkola, *EMNLP’17*

- **Perturb:** Encode -> perturb -> decode -> query

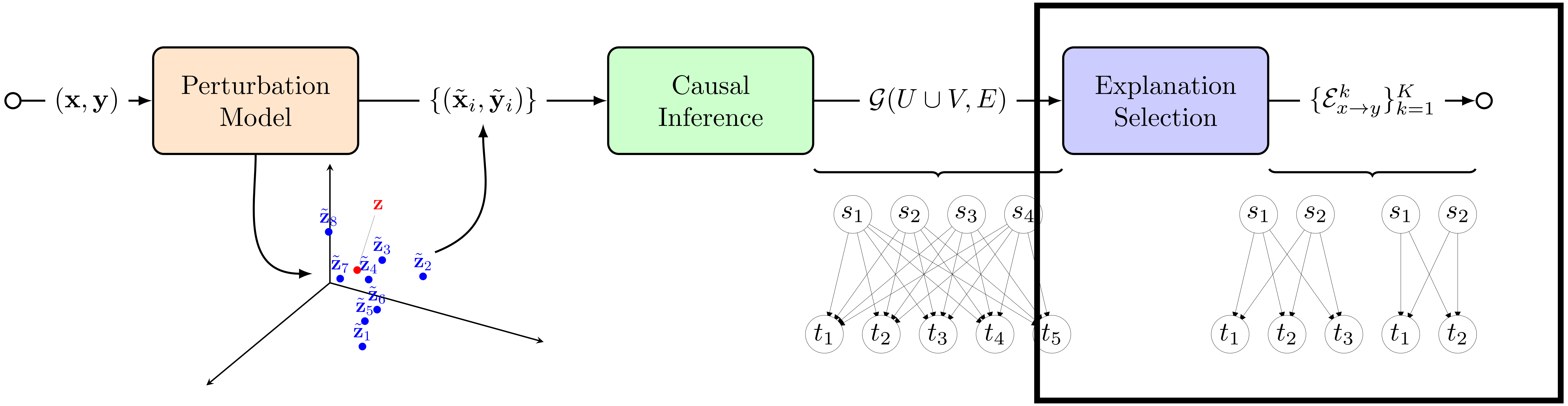
- **Infer:** Logistic regression to infer causal dependencies

- **Select:** Partition dependency graph into *explanation chunks*


$(\mathbf{x}, \mathbf{y})$
Perturbation Model
$\{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i)\}$
Causal Inference
$\mathcal{G}(U \cup V, E)$
Explanation Selection
$\{\mathcal{E}^k_{x \to y}\}_{k=1}^K$
$\mathbf{z}$
$\tilde{\mathbf{z}}_8$
$\tilde{\mathbf{z}}_3$
$\tilde{\mathbf{z}}_7$
$\tilde{\mathbf{z}}_4$
$\tilde{\mathbf{z}}_2$
$\tilde{\mathbf{z}}_6$
$\tilde{\mathbf{z}}_5$
$\tilde{\mathbf{z}}_1$
s_1 s_2 s_3 s_4
t_1 t_2 t_3 t_4 t_5
s_1 s_2
t_1 t_2 t_3
s_1 s_2
t_1 t_2

Interpretability for Black-Box Seq2Seq Models

AM+ Jaakkola, *EMNLP'17*



- **Perturb:** Encode -> perturb -> decode -> query
- **Infer:** Logistic regression to infer causal dependencies
- **Select:** Partition dependency graph into *explanation chunks*

Interpretability for Bias Detection in Machine Translation