Semantically-informed Hierarchical Event Modeling

Shubhashis Roy Dipta

Mehdi Rezaee

Frank Ferraro

UMBC

Supported by: NSF DEVCOM ARMY RESEARCH LABORATORY DARPA

Complex events and situations can be hierarchical.

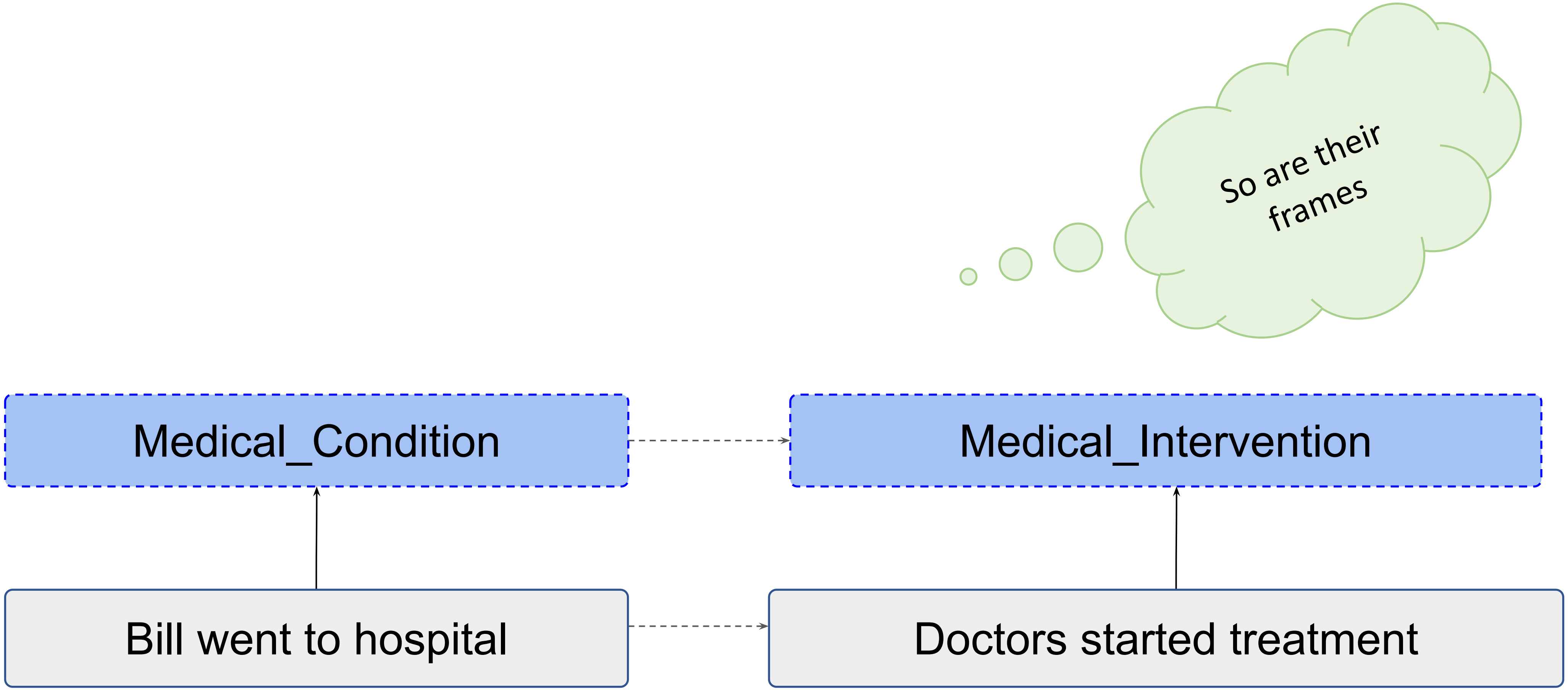
This hierarchy presents difficulties.

Difficulty of Complex Events

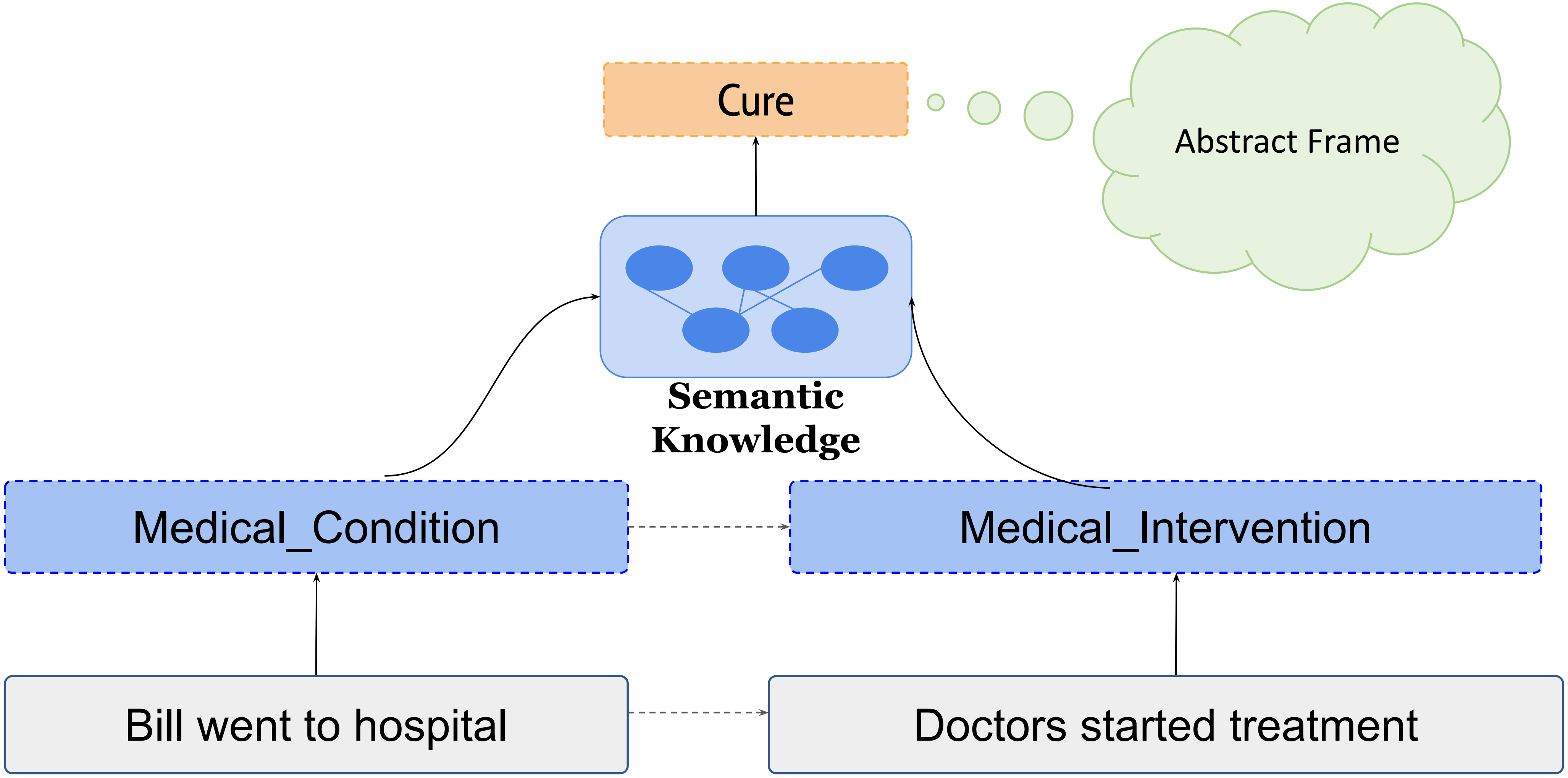
Events are inter-connected

Bill went to hospital → Doctors started treatment

Difficulty of Complex Events



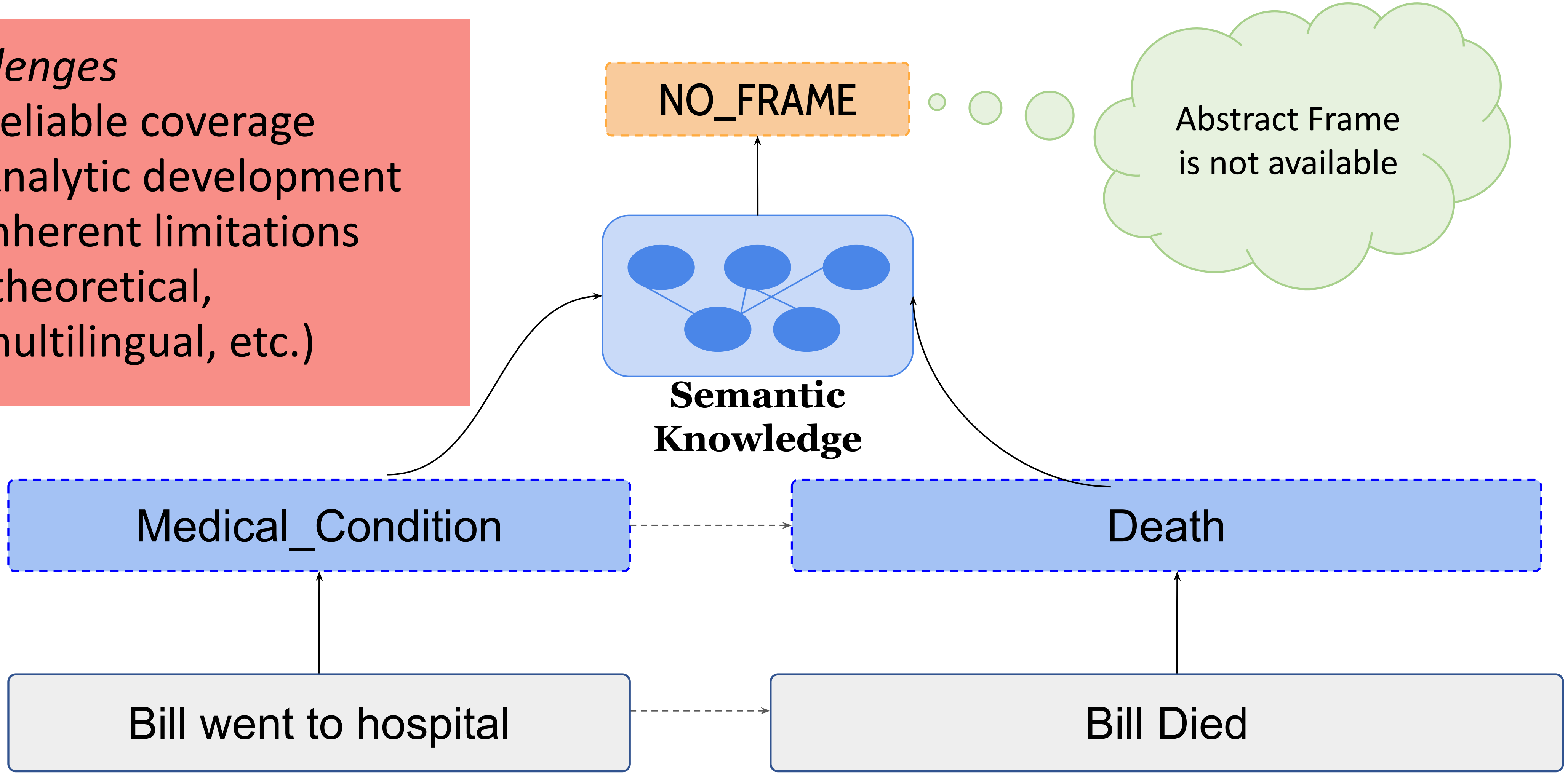
Semantic Associations can help…



…unless you don't have them…

Challenges

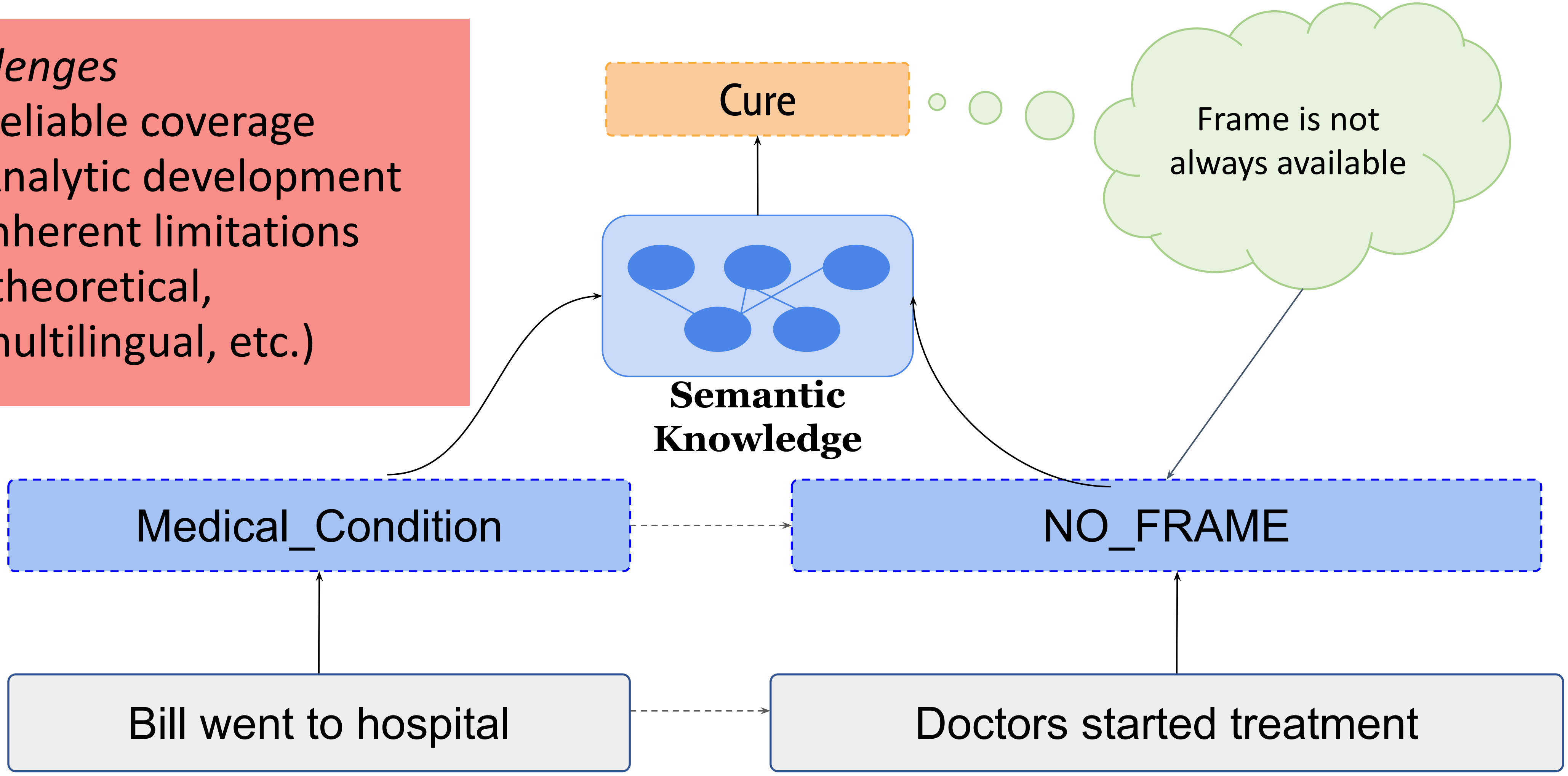
- Reliable coverage
- Analytic development
- Inherent limitations (theoretical, multilingual, etc.)



…or if you are missing "lower" semantics

Challenges

- Reliable coverage
- Analytic development
- Inherent limitations (theoretical, multilingual, etc.)



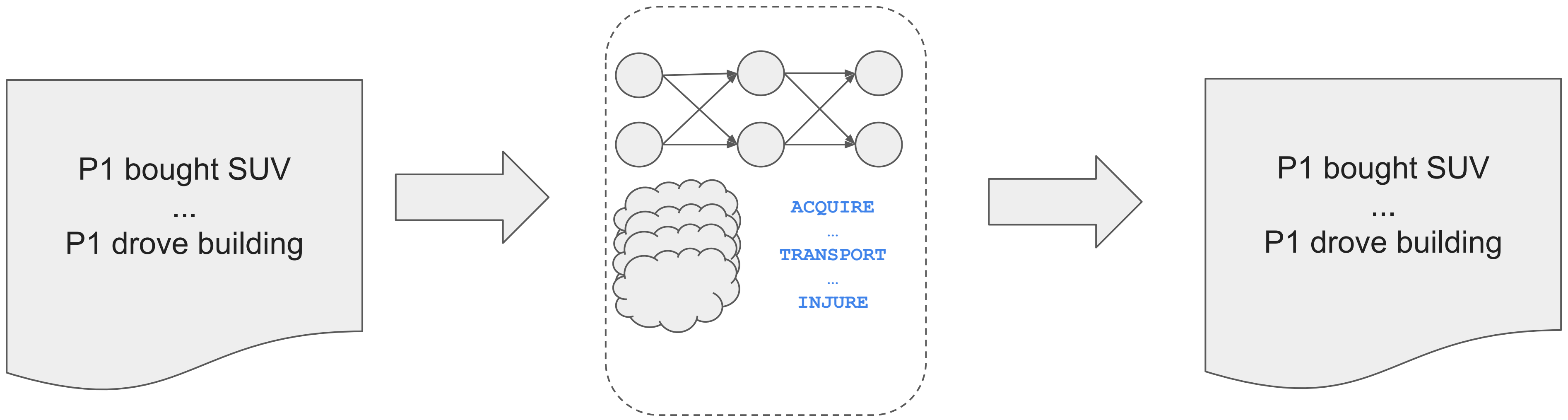
Complex events and situations are hierarchical. How can we better capture this hierarchy…

- via modeling improvements?
- utilizing (existing) semantic resources?

We introduce SHEM:

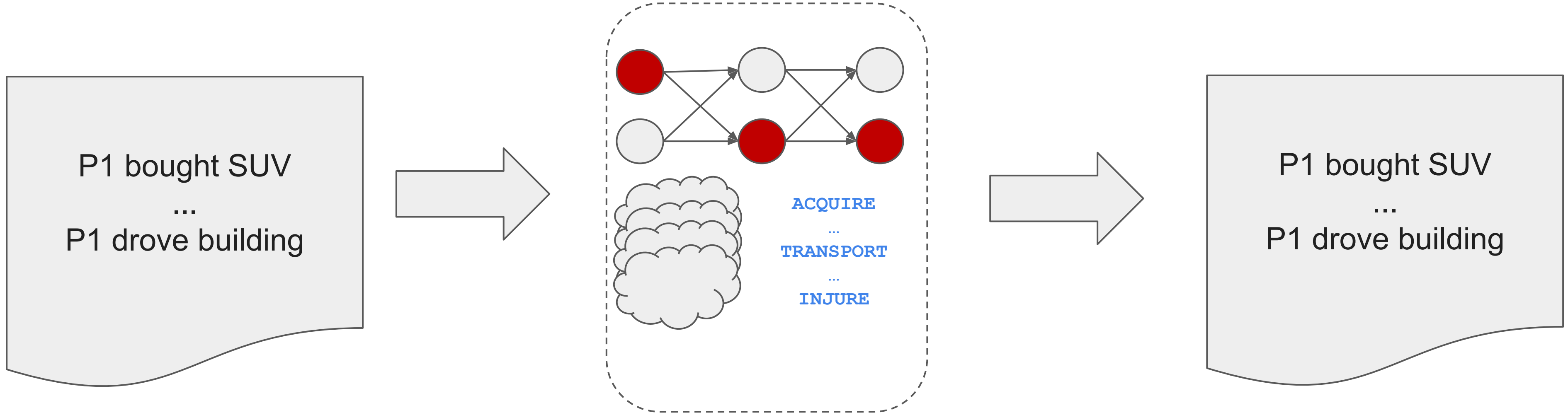
- A Semi-supervised, Hierarchical Event learning Model
- Use of existing FrameNet resource for extracting side knowledge and abstract concept about the events
- Hierarchical model with combination of InfoNCE loss to provide better event representation
- Our model shows better performance in multiple tasks
  - What event comes next
  - Generating missing events
  - Identify similar or related events

Encoding into Primitives and Back-Again



A neural generative model that uses
partially-observed, semantic (ontology-based) knowledge
to explain and predict events

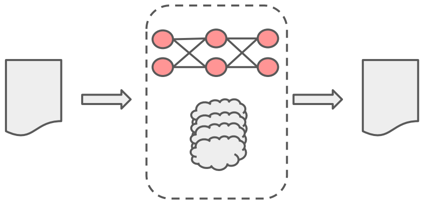
Encoding into Primitives and Back-Again

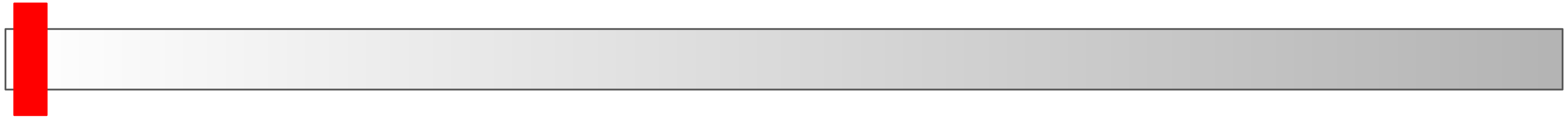


A neural generative model that uses partially-observed, semantic (ontology-based) knowledge to explain and predict events

Represent latent variables with Gumbel-Softmax and ***softly*** inject the information into them

Encoding into Primitives and Back-Again





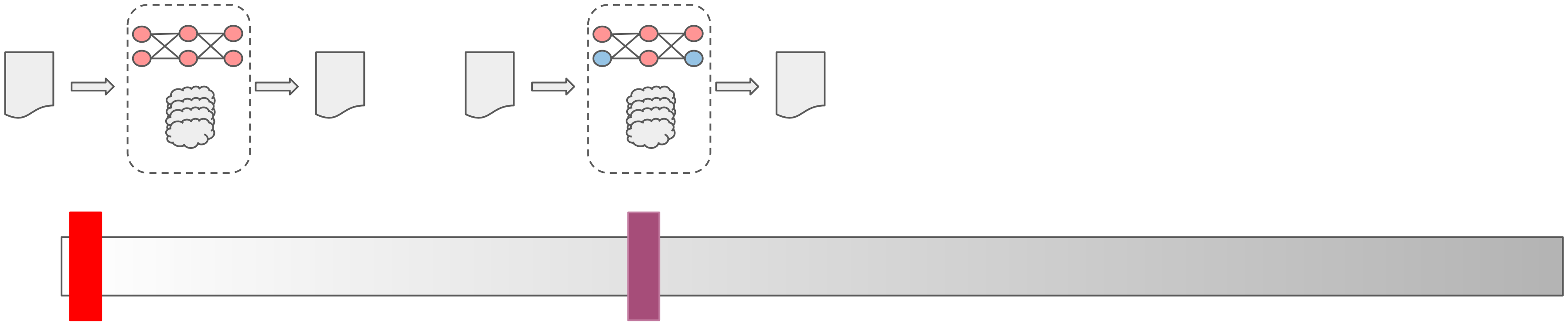
0%

Semantic knowledge should be there. **None** are known during learning.

A neural generative model that uses partially-observed, semantic (ontology-based) knowledge to explain and predict events

Represent latent variables with Gumbel-Softmax and ***softly*** inject the information into them

Encoding into Primitives and Back-Again



0%

Semantic knowledge should be there. **None** are known during learning.

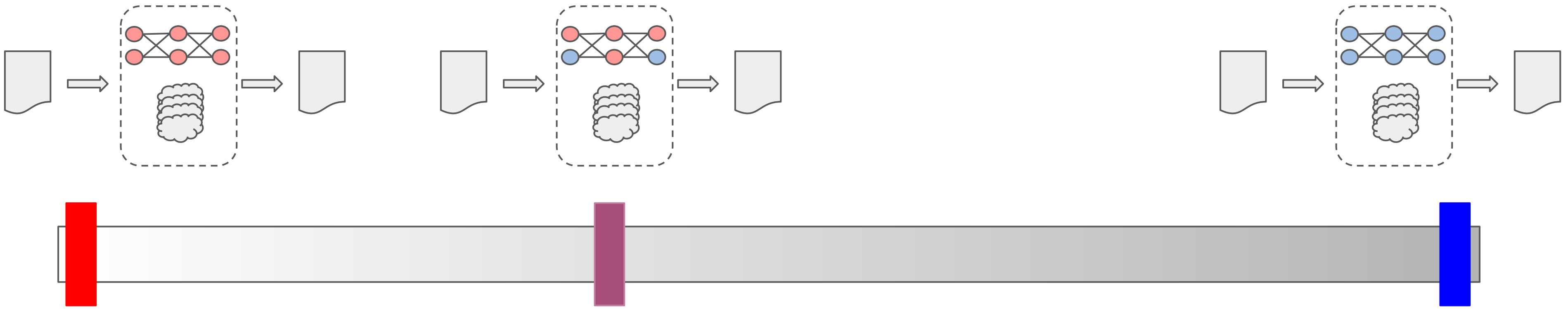
40%

Semantic knowledge should be there. **Some** are known during learning.

A neural generative model that uses partially-observed, semantic (ontology-based) knowledge to explain and predict events

Represent latent variables with Gumbel-Softmax and ***softly*** inject the information into them

Encoding into Primitives and Back-Again



0% | 40% | 100%

Semantic knowledge should be there. **None** are known during learning.

Semantic knowledge should be there. **Some** are known during learning.

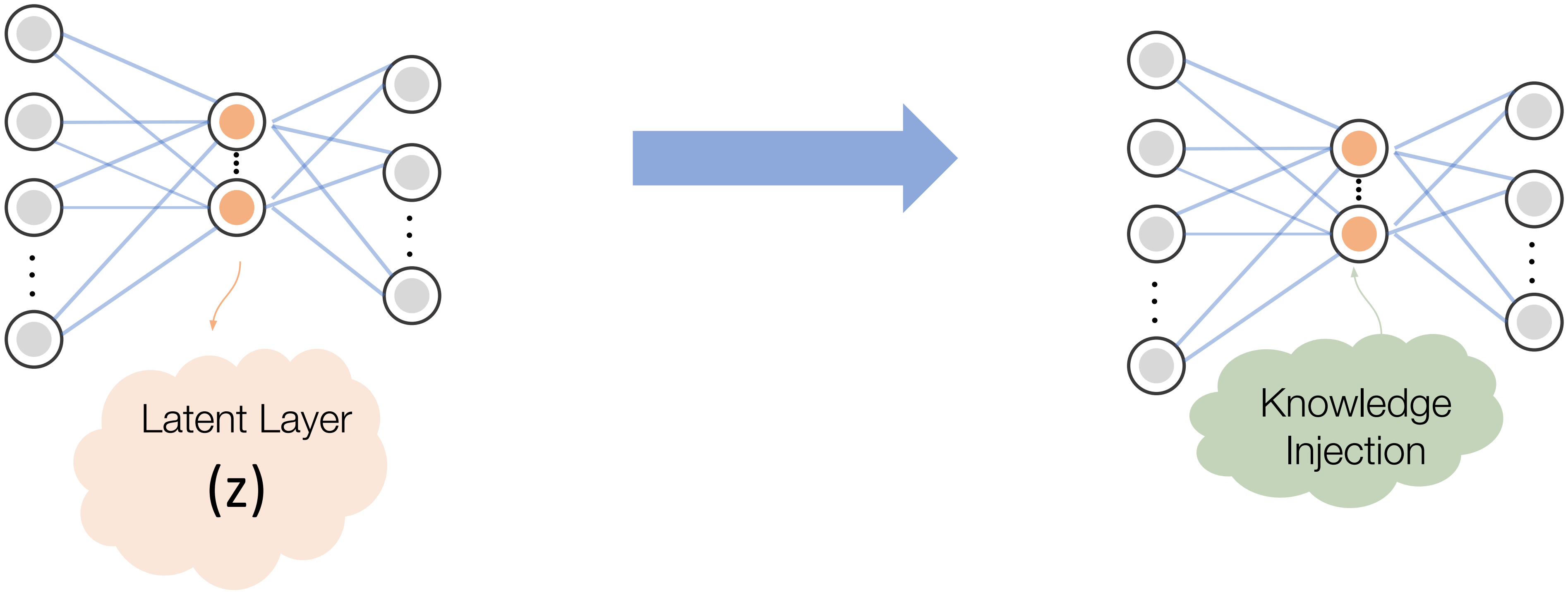
Semantic knowledge is there. **(Almost) All** are known during learning.

A neural generative model that uses partially-observed, semantic (ontology-based) knowledge to explain and predict events

Represent latent variables with Gumbel-Softmax and ***softly*** inject the information into them

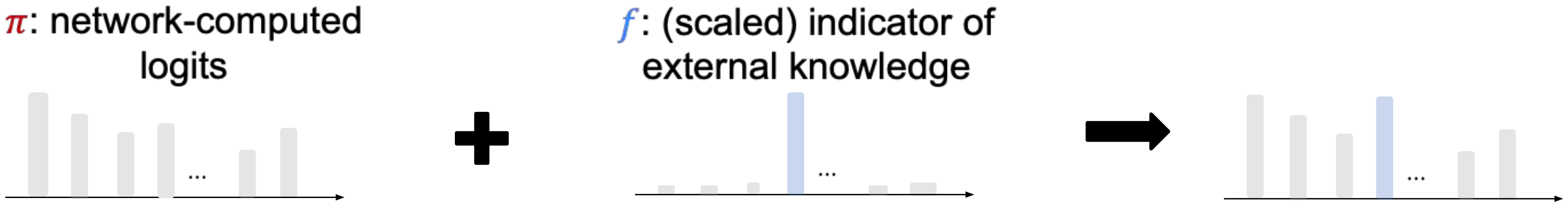
Previous Solutions

- HAQAE (Weber et al., 2018): latent tree-based model
- SSDVAE (Rezaee and Ferraro, 2021): latent variable method, injecting semantic information to the discrete latent layer parameters.



Previous Solutions

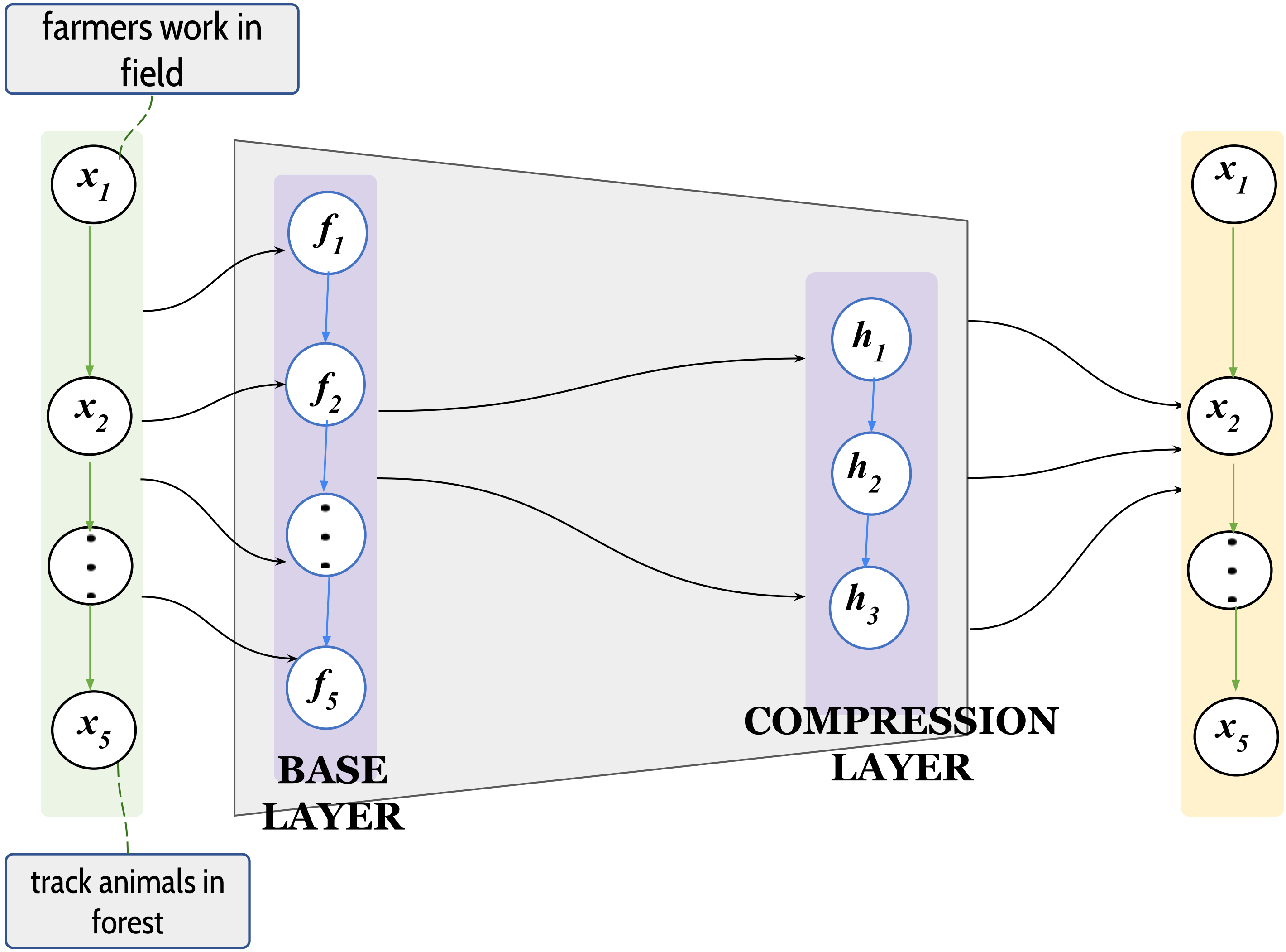
- HAQAE (Weber et al., 2018): latent tree-based model
- SSDVAE (Rezaee and Ferraro, 2021): latent variable method, injecting semantic information to the discrete latent layer parameters.



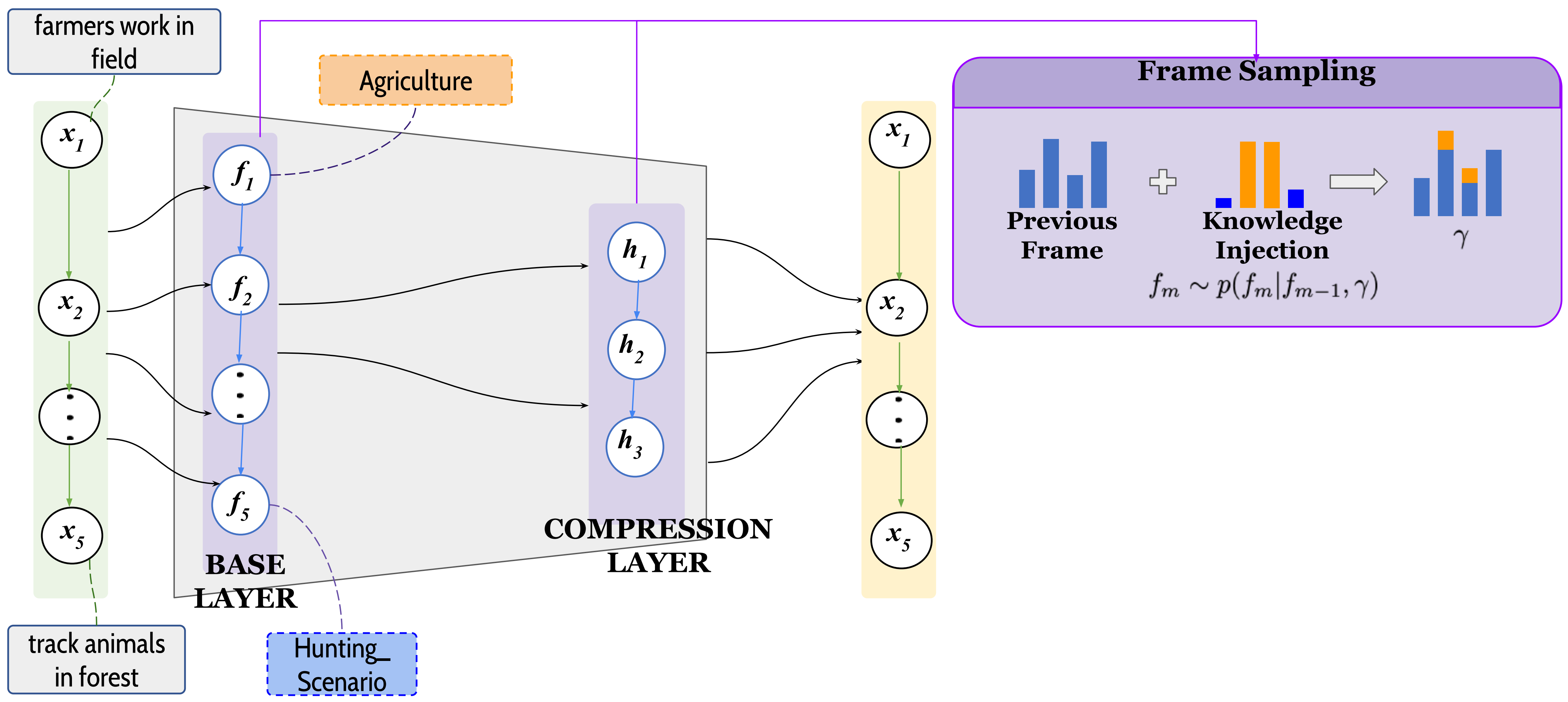
Our Model Description

- Two Layers
  - Base Layer
    - Token Encoder
    - Latent Variable Classification
    - Token Decoder
  - Compression Layer
    - Semantic Knowledge Extraction
    - Sequence Encoder
    - Latent Variable Compression
    - Token Decoder

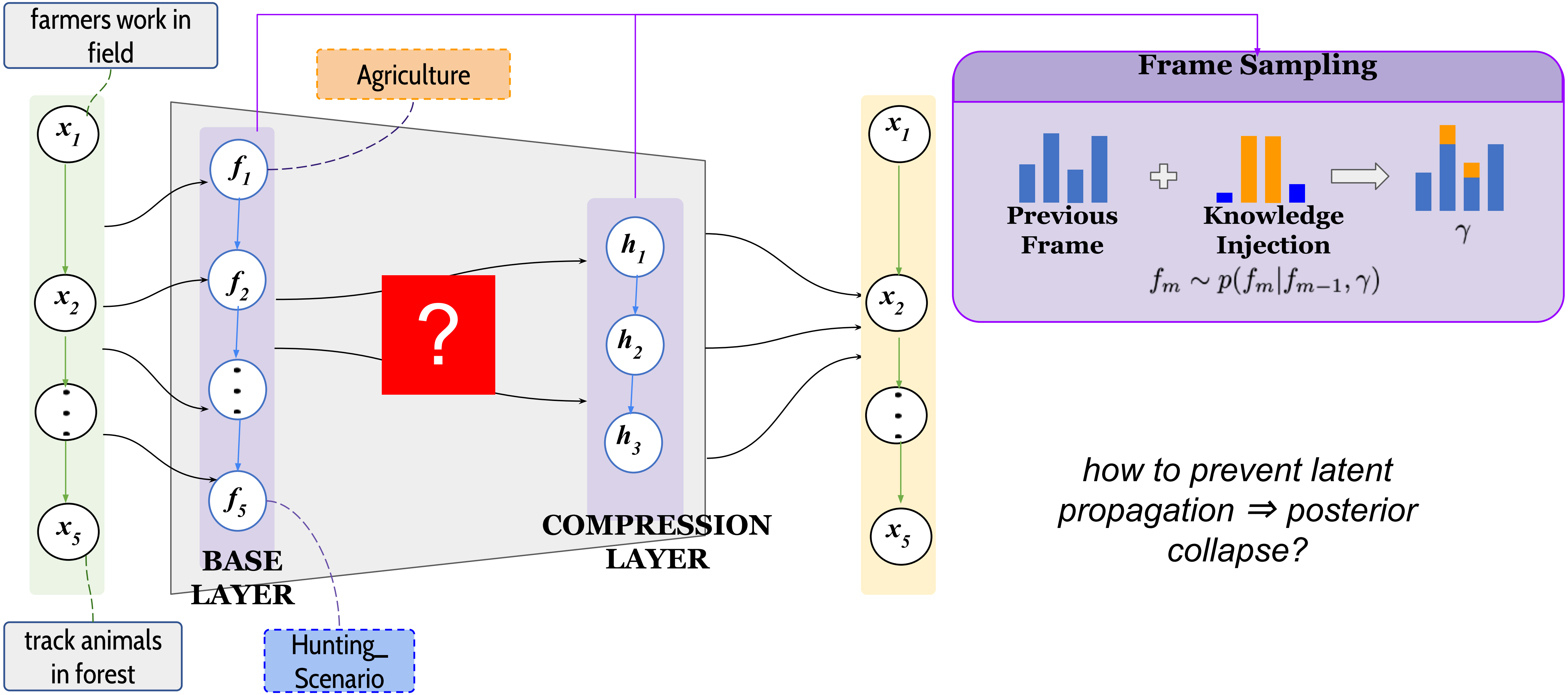
Capturing Joint Hierarchy: Multi-layer Encoder-Decoder



Capturing Joint Hierarchy: Inject Partially Observable Knowledge

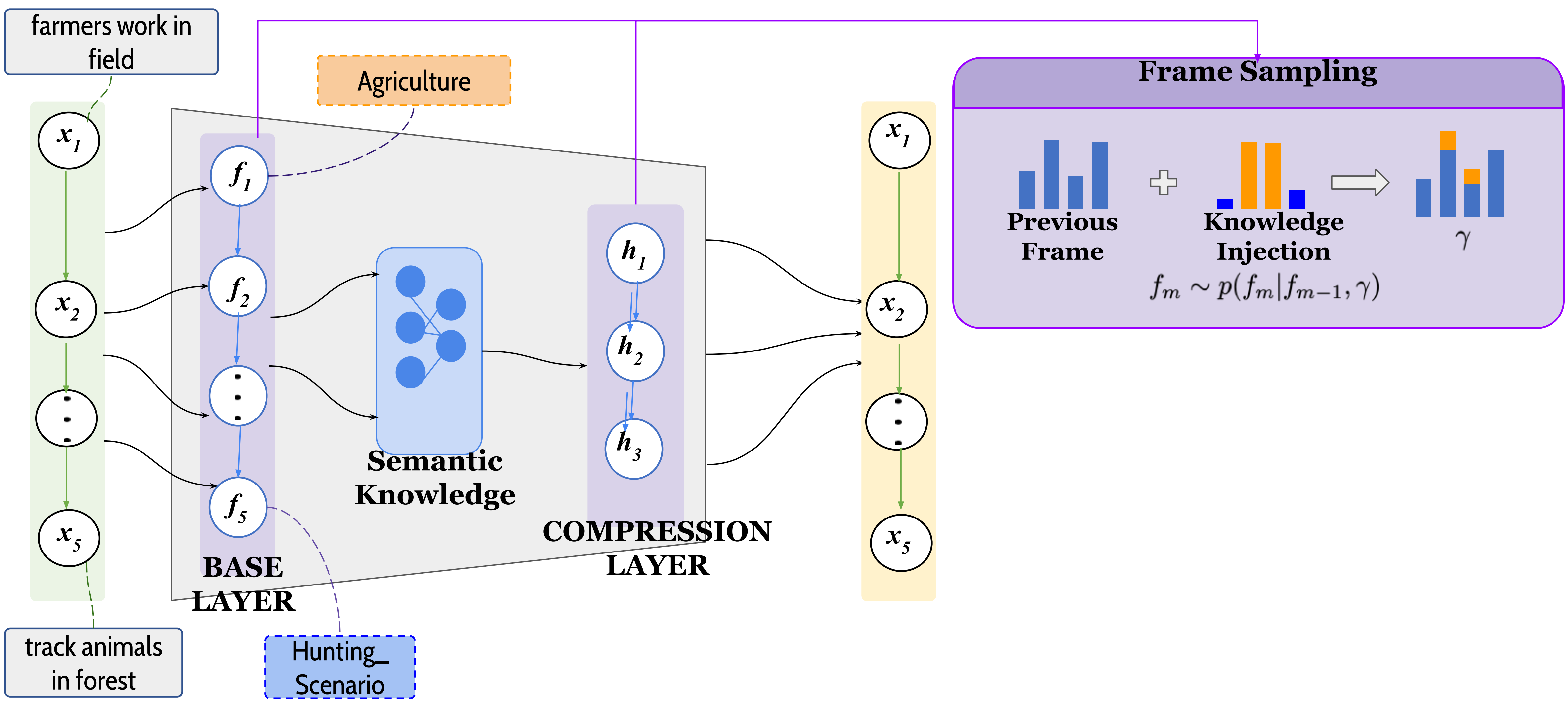


Capturing Joint Hierarchy: Inject Partially Observable Knowledge

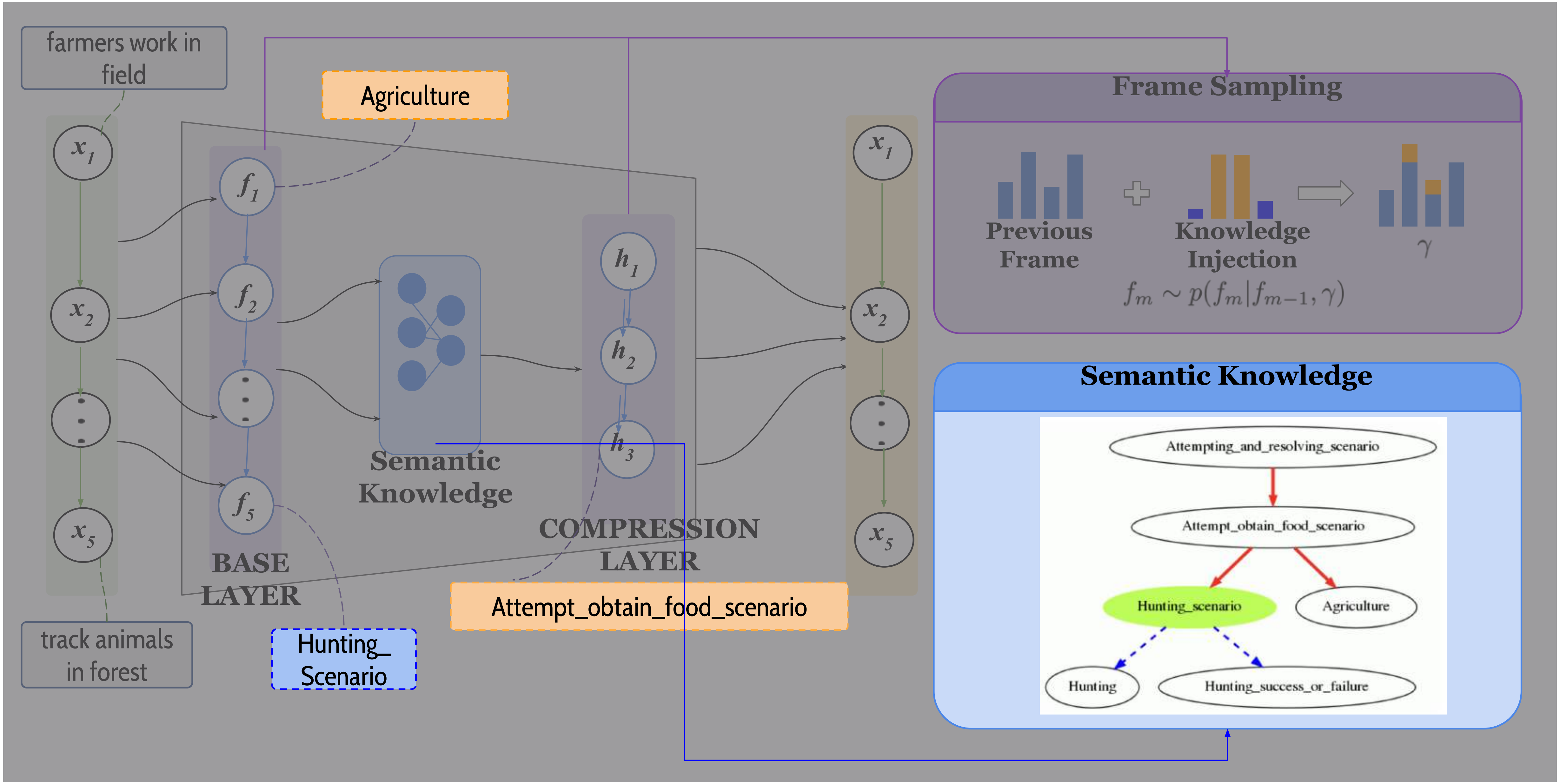


how to prevent latent propagation ⇒ posterior collapse?

Capturing Joint Hierarchy: Project over Semantic Resource



Capturing Joint Hierarchy: Project over Semantic Resource



Generalized Loss Function

We optimize a weighted variant of the **ELBO** to train the model:

$\mathcal{L} =$ **Reconstruct** from sampled *base* layer frames $+$ **Reconstruct** from sampled *compression* layer frames $+$ Predict the **observed base layer frames** (if any) $+$ **KL** (Regularizer, each layer)

Research Questions

1. Is frame inheritance sufficient?

2. How effective are other frame relations?

3. Can our model generate missing events?

4. Can our model generate better event embeddings?

Datasets (Train)

Wikipedia Dataset

Partition	*#Docs*
Train	457k
Dev	16k

- Our Wikipedia dataset is a partial dump of English Wikipedia.
- The FrameNet annotations are automatically extracted.

Datasets (Test)

- Event Modeling
  - Wikipedia - 21k #docs
- Missing Events
  - Wikipedia - 4k #docs
- Inverse Narrative Cloze
  - Wikipedia Inverse Narrative - 2000 #docs (6 choices)
- Event Similarity Tasks
  - Hard Similarity - 230 event pairs
  - Hard Extension - 1000 event pairs
  - Transitive Sentence Similarity - 108 event pairs

1. Is frame inheritance sufficient?

Model	ε	Perplexity (⬇️)	INC Score (⬆️)
HAQAE	–		
SSDVAE	0.9		
ours			
SSDVAE	0.7		
ours			
SSDVAE	0.5		
ours			
SSDVAE	0.4		
ours			
SSDVAE	0.2		
ours			

ε: the average percent of frames observed during training

- Emulate how sufficiently accurate, extractable semantic knowledge may not always be available.
- **ε** fixed prior to training each model.
- Frames are only observed during training, and never during evaluation

1. Is frame inheritance sufficient?

Model	ε	Perplexity (⬇️)	INC Score (⬆️)
HAQAE	–	21.38	24.88
SSDVAE	0.9	19.84	35.56
ours			
SSDVAE	0.7	21.19	39.08
ours			
SSDVAE	0.5	31.11	40.18
ours			
SSDVAE	0.4	33.12	47.88
ours			
SSDVAE	0.2	33.31	44.38
ours			

1. Is frame inheritance sufficient?

Model	ε	Perplexity (⬇️)	INC Score (⬆️)
HAQAE	–	21.38	24.88
SSDVAE	0.9	19.84	35.56
ours		**19.39**	41.35
SSDVAE	0.7	21.19	39.08
ours			
SSDVAE	0.5	31.11	40.18
ours			
SSDVAE	0.4	33.12	47.88
ours			
SSDVAE	0.2	33.31	44.38
ours		30.15	**49.53**

✔ SHEM is better able to model longer event sequences

1. Is frame inheritance sufficient?

Model	ε	Perplexity (⬇️)	INC Score (⬆️)
HAQAE	–	21.38	24.88
SSDVAE	0.9	19.84	35.56
ours		**19.39**	41.35
SSDVAE	0.7	21.19	39.08
ours			
SSDVAE	0.5	31.11	40.18
ours			
SSDVAE	0.4	33.12	47.88
ours			
SSDVAE	0.2	33.31	44.38
ours		30.15	**49.53**

✔ SHEM is better able to model longer event sequences

✔ [See paper] Lexical signal vs. inferred frames: inferred frames and ontological relations are important

✔ [See paper] Hierarchical layer provides useful, less-than-full supervised feedback

1. Is frame inheritance sufficient?

Model	ε	Perplexity (⬇️)	INC Score (⬆️)
HAQAE	–	21.38	24.88
SSDVAE	0.9	19.84	35.56
ours		**19.39**	41.35
SSDVAE	0.7	21.19	39.08
ours			
SSDVAE	0.5	31.11	40.18
ours			
SSDVAE	0.4	33.12	47.88
ours			
SSDVAE	0.2	33.31	44.38
ours		30.15	**49.53**

✔ SHEM is better able to model longer event sequences

✔ [See paper] Lexical signal vs. inferred frames: inferred frames and ontological relations are important

✔ [See paper] Hierarchical layer provides useful, less-than-full supervised feedback

✔/✗ Able to leverage more observation (ε=0.9) or additional structure (ε=0.2); …

1. Is frame inheritance sufficient?

Model	ε	Perplexity (⬇️)	INC Score (⬆️)
HAQAE	–	21.38	24.88
SSDVAE	0.9	19.84	35.56
ours		19.39	41.35
SSDVAE	0.7	21.19	39.08
ours		20.26	35.86
SSDVAE	0.5	31.11	40.18
ours		22.16	37.3
SSDVAE	0.4	33.12	47.88
ours		24.02	43.25
SSDVAE	0.2	33.31	44.38
ours		30.15	49.53

✔ SHEM is better able to model longer event sequences

✔ [See paper] Lexical signal vs. inferred frames: inferred frames and ontological relations are important

✔ [See paper] Hierarchical layer provides useful, less-than-full supervised feedback

✔/✗ Able to leverage more observation (ε=0.9) or additional structure (ε=0.2); mixed signals ⇒ mixed performance

1. Is frame inheritance sufficient?

Model	ε	Perplexity (⬇️)	INC Score (⬆️)
HAQAE	–	21.38	24.88
SSDVAE	0.9	19.84	35.56
ours		**19.39**	41.35
SSDVAE	0.7	21.19	39.08
ours		20.26	35.86
SSDVAE	0.5	31.11	40.18
ours		22.16	37.3
SSDVAE	0.4	33.12	47.88
ours		24.02	43.25
SSDVAE	0.2	33.31	44.38
ours		30.15	**49.53**

✔ SHEM is better able to model longer event sequences

✔ [See paper] Lexical signal vs. inferred frames: inferred frames and ontological relations are important

✔ [See paper] Hierarchical layer provides useful, less-than-full supervised feedback

✔/✗ Able to leverage more observation (ε=0.9) or additional structure (ε=0.2); mixed signals ⇒ mixed performance

💭 Frame inheritance (e.g., IS-A type relations) helpful but not sufficient for hierarchical event modeling

2. How effective are other frame relations?

		Low Observation: ε=0.2		High Observation: ε=0.9	
Model	**Frame Relation**	**Perplexity (⬇️)**	**INC Score (⬆️)**	**Perplexity (⬇️)**	**INC Score (⬆️)**
SSDVAE	-				
SHEM (ours)	Inheritance				
	Using				
	Precedes				
	Causative_of				
	Grouping				
	scenario_only				

2. How effective are other frame relations?

		Low Observation: ε=0.2		High Observation: ε=0.9	
Model	**Frame Relation**	**Perplexity (⬇️)**	**INC Score (⬆️)**	**Perplexity (⬇️)**	**INC Score (⬆️)**
SSDVAE	-	33.31	44.38	19.84	35.56
SHEM (ours)	Inheritance	30.15	49.53	19.39	41.35
	Using	31.37	49.72	19.39	**43.23**
	Precedes	32.62	47.92	19.57	41.43
	Causative_of	31.82	**49.85**	19.42	41.38
	Grouping	**28.17**	48.88	19.44	40.76
	scenario_only	32.01	48.10	**18.81**	42.29

✔ Existence of broader associations that these relations enable are very helpful

✔ Lower/higher observation consistently better than previous

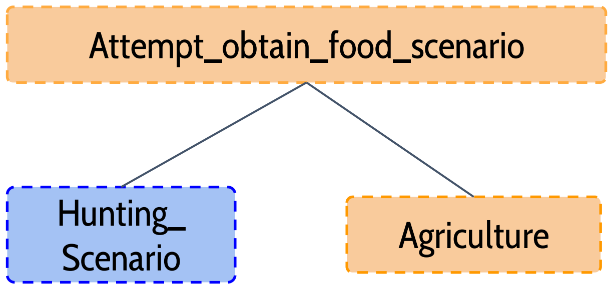
✔ [See paper] even with limited guidance, compression gives valuable feedback

Event modeling could benefit from broader semantic resource coverage: how to encode semantics of any particular relation?

3. Can our **SCENARIO** model generate missing events?

Goal: examine the robustness of our model with respect to semantically related missing events in an input sequence

1. Identify sequences where two events have different frames that are contained within the same scenario frame.
2.
3.



3. Can our **SCENARIO** model generate missing events?

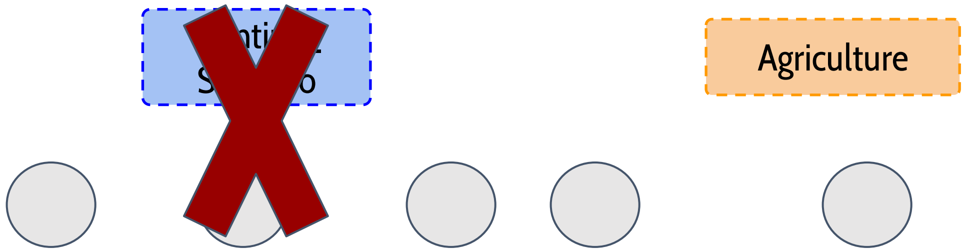
Goal: examine the robustness of our model with respect to semantically related missing events in an input sequence

1. Identify sequences where two events have different frames that are contained within the same scenario frame.
2. Train normally
3.

3. Can our **SCENARIO** model generate missing events?

Goal: examine the robustness of our model with respect to semantically related missing events in an input sequence

1. Identify sequences where two events have different frames that are contained within the same scenario frame.
2. Train normally
3. To evaluate, consider semantically impoverished input
   a. **Remove an event** associated with a scenario-connected frame in the input.
   b. Task: require the model to generate the full, unmodified sequence.



3. Are we robust to missing events?

Model	ε	Perplexity (Masked Test Data) (⬇️)
		Base Alone
SSDVAE	0.9	152.44
SHEM: grouping		**<u>61.10</u>**
SHEM: scenario		**<u>63.48</u>**
SSDVAE	0.5	182.63
SHEM: grouping		79.74
SHEM: scenario		76.01
SSDVAE	0.2	212.93
SHEM: grouping		89.73
SHEM: scenario		83.86

✔ More observation ⇒ coarser-grained grouping okay

✔ Less observation ⇒ precise groupings helpful

✔ [See paper] Suggests improved capability of SHEM to better encapsulate abstract meaning of an event sequence

4. Can our model generate better event embeddings?

- Augment our loss with an event contrastive learning loss, proposed by Gao et al. (2022: SWCC)
- Learn to *align* similar event embeddings, and repel dissimilar ones
- Form event embeddings from decoder & latent compression layer

Model	"Original" Similarity (Acc.)	"Extended" Similarity (Acc.)	Transitivity (Corr.)
SWCC (16)	78.9 ± 1.3	69.2 ± 0.9	**0.82 ± 0**
SWCC (256)	81.1 ± 0.4	72.6 ± 1.5	**0.82 ± 0**
Ours	**83.3 ± 2.3**	**78.6 ± 3.0**	0.77 ± 0.04

Contrastive Learning vs. Language Modeling

	Similarity (Orig.)	
	SWCC	SHEM
LM only	25.87	67.83
Contrastive only	78.48	67.18
Contrastive + LM	78.91	83.26

Neither contrastive nor LM/MLM loss are as strong as both together

The LM component in our approach is important to overall performance

Summary

1. We introduce **SHEM**, a novel, hierarchical, semi-supervised event learning model.

2. We show how to use FrameNet for both observable event frames and latent abstract event frames.

3. More informed signal from compression layer when performing different tasks.

4. Our model can generate better event embeddings for out-of-domain dataset.

/dipta007/SHEM

{sroydip1, ferraro}@umbc.edu

Auxiliary Slides

1- Is frame inheritance sufficient?

Model	ε	Perplexity (⬇️)	INC Score (⬆️)
HAQAE		21.38 ± 0.25	24.88 ± 1.35
SSDVAE	0.9	19.84 ± 0.52	35.56 ± 1.70
ours		**19.39** ± 0.30	41.35 ± 4.25
SSDVAE	0.7	21.19 ± 0.76	39.08 ± 1.55
ours		20.26 ± 1.36	35.86 ± 3.43
SSDVAE	0.5	31.11 ± 0.85	40.18 ± 0.90
ours		22.16 ± 1.62	37.30 ± 3.33
SSDVAE	0.4	33.12 ± 0.54	47.88 ± 3.59
ours		24.02 ± 1.28	43.25 ± 4.97
SSDVAE	0.2	33.31 ± 0.63	44.38 ± 2.10
ours		30.15 ± 2.73	**49.53** ± 1.56

2- How effective are other frame relations?

Model	Frame Relation	ε	Perplexity (⬇️)	INC Score (⬆️)
HAQAE	-	-	21.38 ± 0.25	24.88 ± 1.35
SSDVAE	-	**0.9**	19.84 ± 0.52	35.56 ± 1.70
ours	Inheritance		19.39 ± 0.53	41.35 ± 4.25
SSDVAE	Using		19.39 ± 0.51	**43.23** ± 2.51
ours	Precedes		19.57 ± 0.58	41.43 ± 3.02
SSDVAE	Causative_of		19.42 ± 0.57	41.38 ± 2.23
ours	Inchoative_of		19.28 ± 0.32	41.35 ± 3.47
SSDVAE	Prospective_on		19.76 ± 0.97	40.53 ± 2.04
ours	Subframe		18.91 ± 0.15	40.35 ± 2.91
SSDVAE	*Grouping*		19.44 ± 0.50	40.76 ± 2.86
ours	*scenario_only*		**18.81** ± 0.50	42.29 ± 2.86

2- How effective are other frame relations?

Model	Frame Relation	ε	Perplexity (⬇️)	INC Score (⬆️)
HAQAE	-	-	21.38 ± 0.25	24.88 ± 1.35
SSDVAE	-	**0.2**	33.31 ± 0.63	44.38 ± 2.10
ours	Inheritance		30.15 ± 2.73	49.53 ± 1.56
SSDVAE	Using		31.37 ± 2.08	49.72 ± 1.73
ours	Precedes		32.62 ± 1.65	47.92 ± 2.25
SSDVAE	Causative_of		31.82 ± 3.00	**49.85** ± 0.84
ours	Inchoative_of		32.65 ± 1.40	48.03 ± 3.35
SSDVAE	Prospective_on		33.20 ± 1.47	47.85 ± 3.53
ours	Subframe		32.78 ± 2.09	47.88 ± 3.31
SSDVAE	*Grouping*		**28.17** ± 2.26	48.88 ± 1.37
ours	*scenario_only*		32.01 ± 0.70	48.10 ± 2.22

3- Can our **GROUPING** model generate missing events?

Model	ε	Perplexity (Masked Test Data) (⬇️)		
		Base Alone	Compression Alone	Base + Compression
SSDVAE	0.9	152.44 ± 3.45	-	-
grouping		**61.10** ± 1.83	94.76 ± 1.96	76.08 ± 0.76
SSDVAE	0.7	163.08 ± 4.52	-	-
grouping		63.50 ± 3.49	86.23 ± 0.70	73.98 ± 2.04
SSDVAE	0.5	182.63 ± 6.11	-	-
grouping		79.74 ± 1.79	83.81 ± 0.96	81.75 ± 1.13
SSDVAE	0.4	201.55 ± 4.10	-	-
grouping		84.17 ± 4.45	81.49 ± 0.14	82.80 ± 2.13
SSDVAE	0.2	212.93 ± 2.54	-	-
grouping		89.73 ± 4.67	77.32 ± 0.72	83.28 ± 2.38

3- Can our **SCENARIO** model generate missing events?

Model	ε	Perplexity (Masked Test Data) (⬇️)		
		Base Alone	Compression Alone	Base + Compression
SSDVAE	0.9	152.44 ± 3.45	-	-
scenario		63.48 ± 4.43	80.94 ± 7.44	71.60 ± 4.12
SSDVAE	0.7	163.08 ± 4.52	-	-
scenario		<u>**60.06**</u> ± 1.68	<u>78.36</u> ± 4.52	<u>68.58</u> ± 2.30
SSDVAE	0.5	182.63 ± 6.11	-	-
scenario		76.01 ± 5.56	78.70 ± 1.63	77.33 ± 3.65
SSDVAE	0.4	201.55 ± 4.10	-	-
scenario		73.77 ± 7.87	80.00 ± 1.89	76.77 ± 4.89
SSDVAE	0.2	212.93 ± 2.54	-	-
scenario		83.86 ± 2.74	81.20 ± 1.17	82.52 ± 1.93

4- Can our model generate better event embeddings?

Model	Batch Size	Hard Similarity (Accuracy %) (⬆️)		Transitive Score Similarity (⬆️)
		Original	Extended	
SWCC	16	78.91 ± 1.31	69.20 ± 0.93	0.82 ± 0.00
SWCC	256	81.09 ± 0.43	72.55 ± 1.53	**0.82 ± 0.00**
ours	16	**83.26 ± 2.29**	**78.63 ± 2.95**	0.77 ± 0.04

4- Ablation study of ours and SWCC

Model	Training Variant	Hard Similarity (Accuracy %) (⬆️)		Transitive Score Similarity (⬆️)
		Original	Extended	
SWCC	Contrastive + LM	78.91 ± 1.31	69.20 ± 0.93	0.82 ± 0.00
	Contrastive Only	78.48 ± 0.83	67.33 ± 0.19	0.78 ± 0.05
	LM only	25.87 ± 1.31	16.78 ± 0.70	0.55 ± 0.04
ours	Contrastive + LM	83.26 ± 2.29	78.63 ± 2.95	0.77 ± 0.04
	Contrastive Only	67.18 ± 1.79	72.75 ± 2.06	0.72 ± 0.02
	LM only	67.83 ± 14.39	62.15 ± 16.52	0.56 ± 0.04