# Evidence-Aware Inferential Text Generation with Vector Quantised Variational AutoEncoder

**Daya Guo**[1], Duyu Tang[2], Nan Duan[2], Jian Yin[1], Daxin Jiang[2] and Ming Zhou[2]

[1]Sun Yat-sen University [2]Microsoft Research

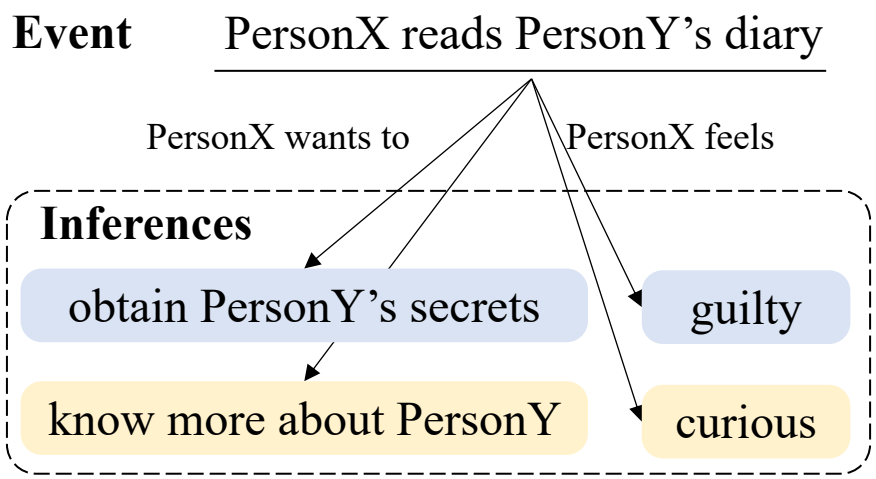

Microsoft Research
微软亚洲研究院

# Task Definition

## Inferential Text Generation

(1) Take an event as the input

(2) Generate multiple inferences (e.g. intent of participants).



# Related Works

(1) Sequence-to-Sequence approaches. [Rashkin et al.,2018; Sap et al.,2019]

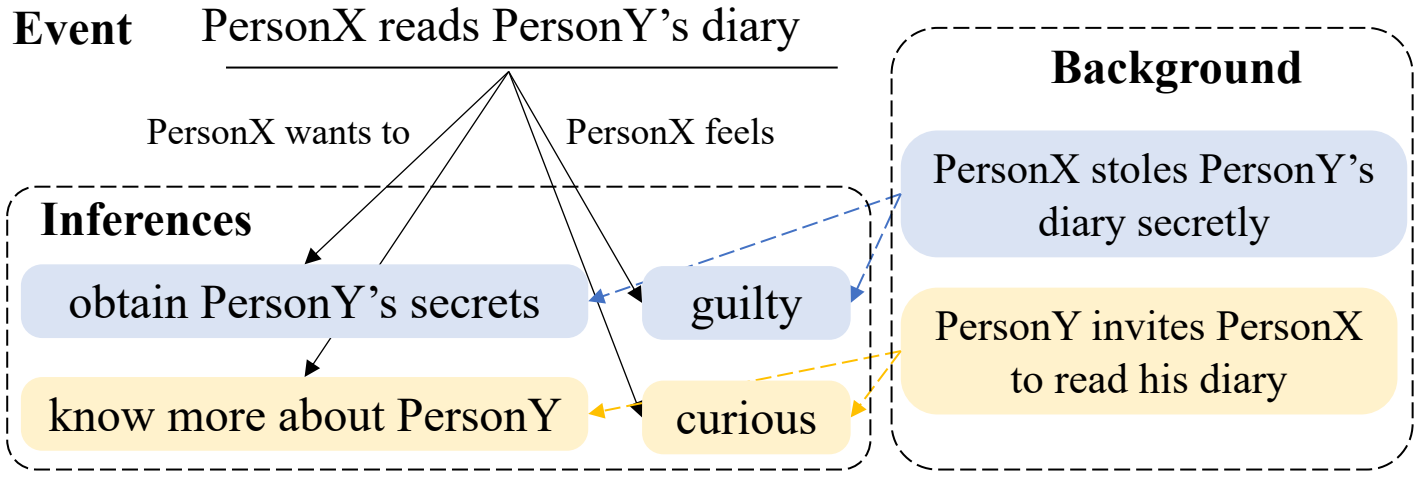
(2) Pre-trained language models such as GPT-2. [Bosselut et al., 2019]

(3) Introduce variational autoencoder to generate diversified inferences. [Du et al., 2019]

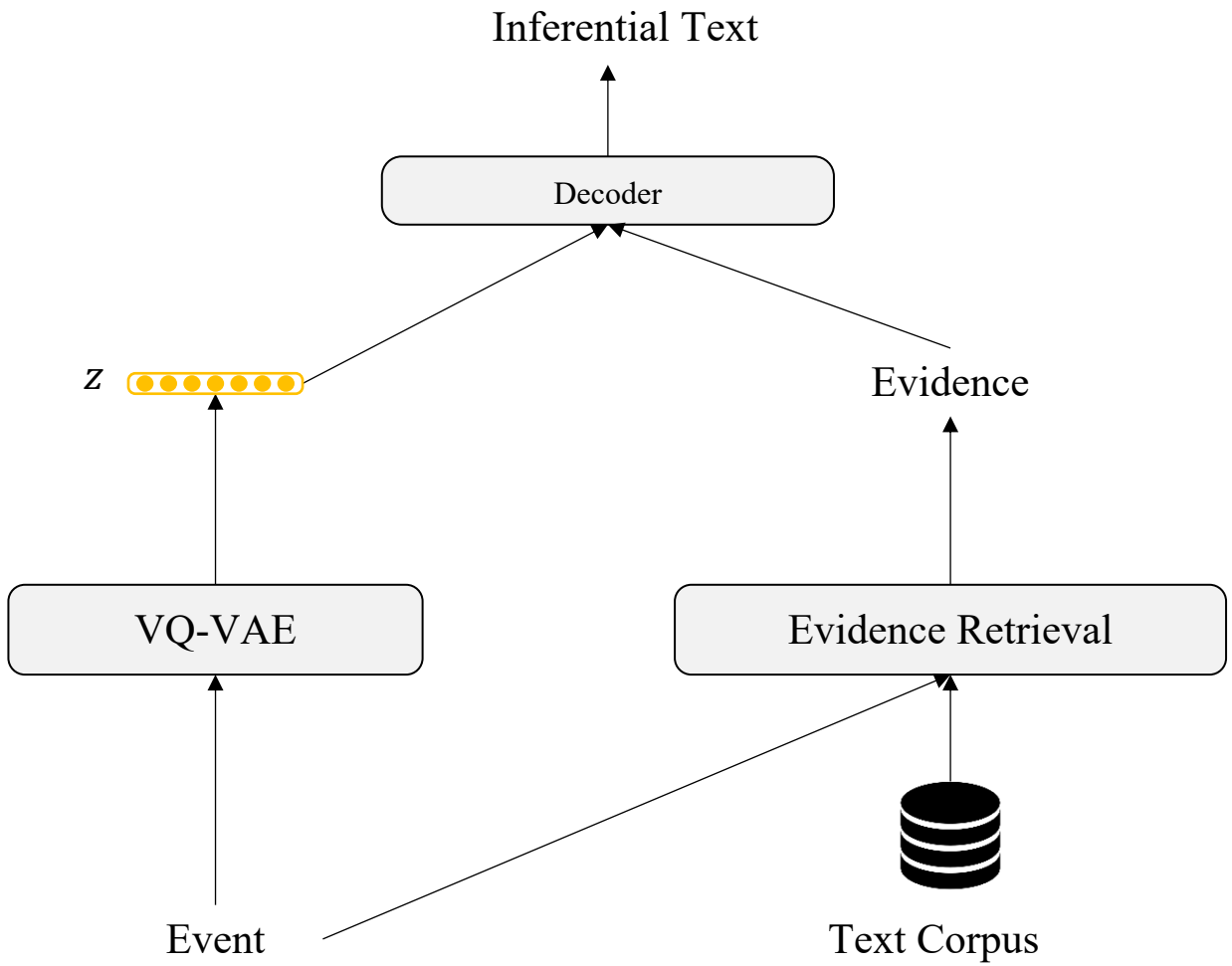
# Motivation

(1) Background knowledge usually provides crucial evidence to generate reasonable inferences.

(2) Different background knowledge could help generate inferences in different perspective.

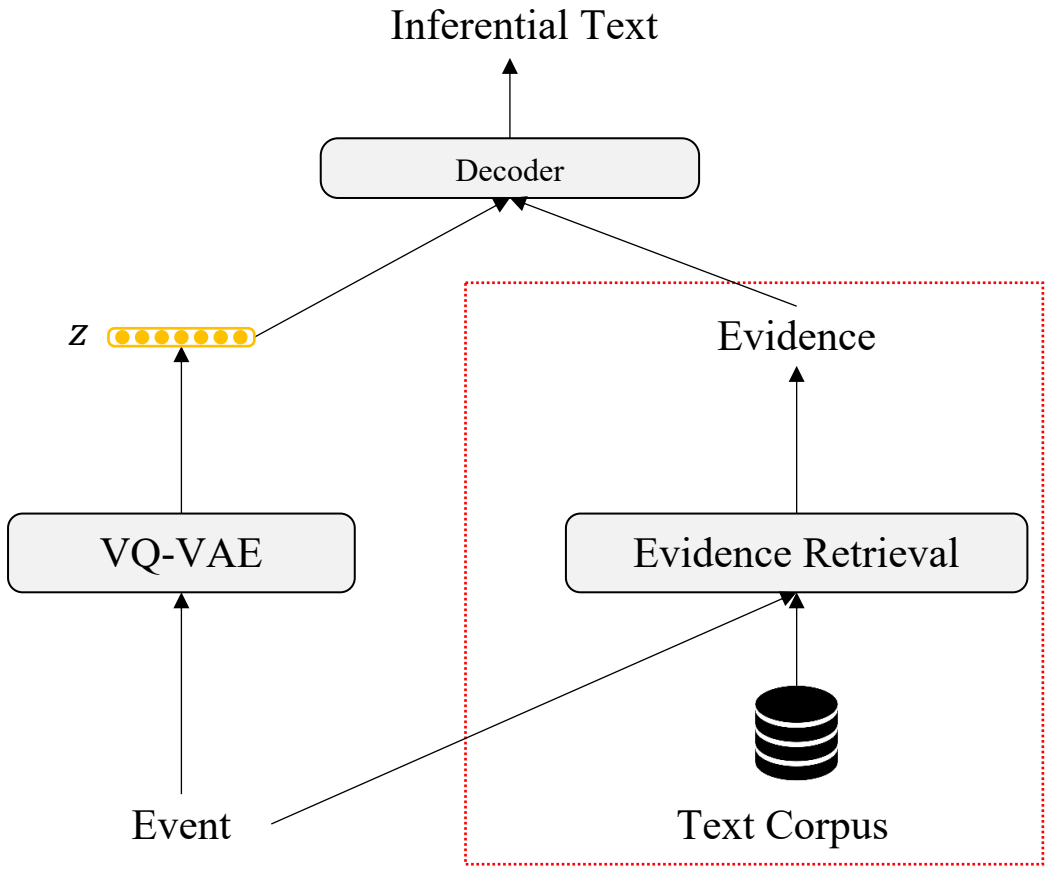


# Overview of Our Approach



# Overview

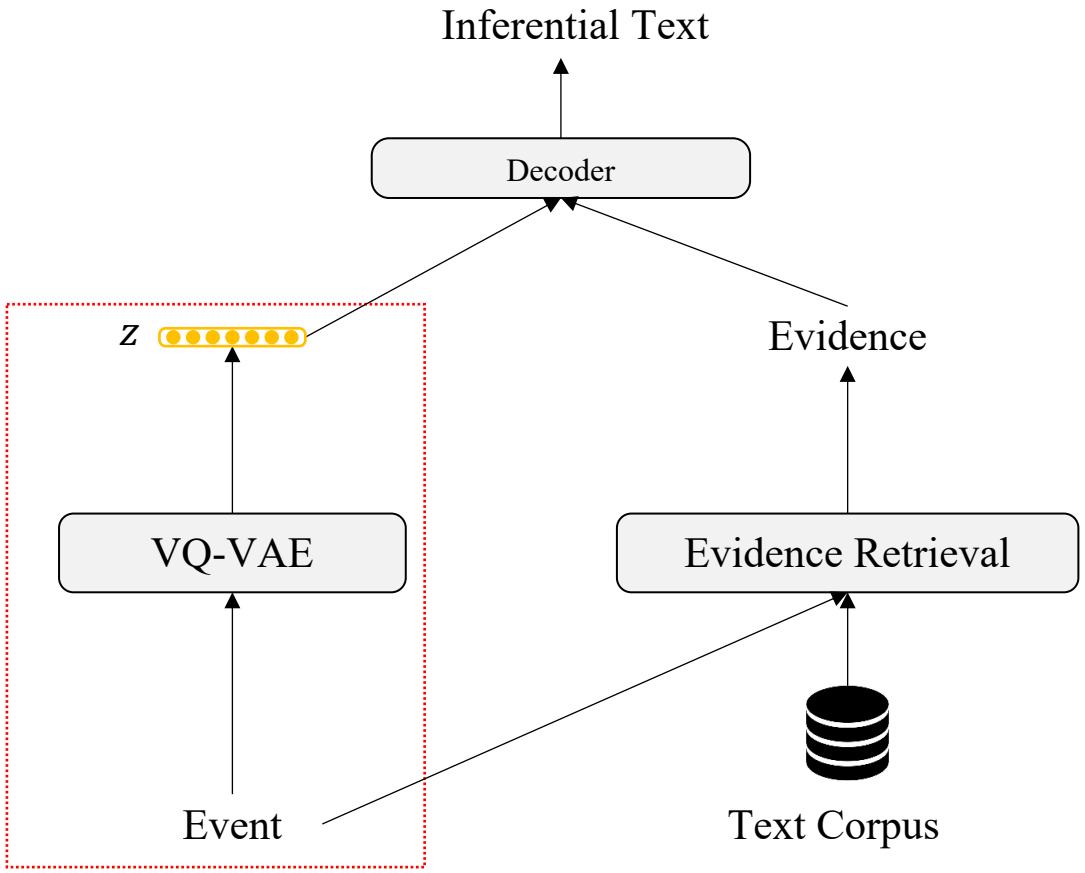
(1) First , the search engine retrieves top K evidence from a large text corpus.



# Overview

(1) First , the search engine retrieves top K evidence from a large text corpus.

(2) Second, VQ-VAE takes an event as the input, and outputs a discrete latent variable $z$.

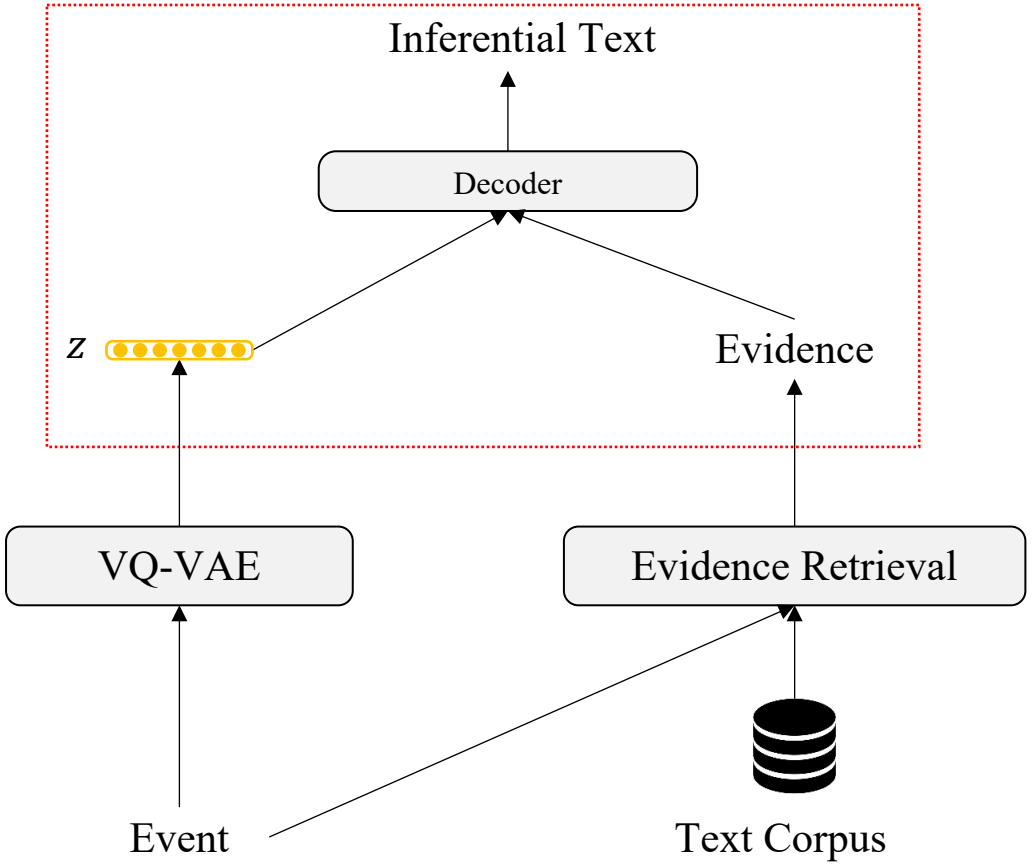


# Overview

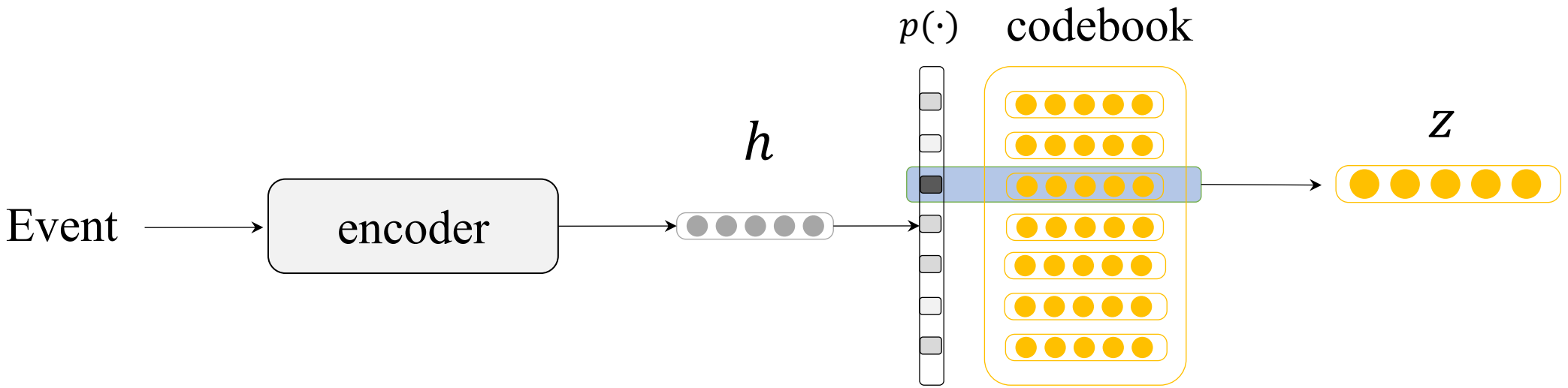
(1) First , the search engine retrieves top K evidence from a large text corpus.

(2) Second, VQ-VAE takes an event as the input, and outputs a discrete latent variable $z$.

(3) Lastly, the decoder takes the latent variable and retrieved evidence as the input and generates the inferential text.



# VQ-VAE: Vector Quantised-Variational AutoEncoder

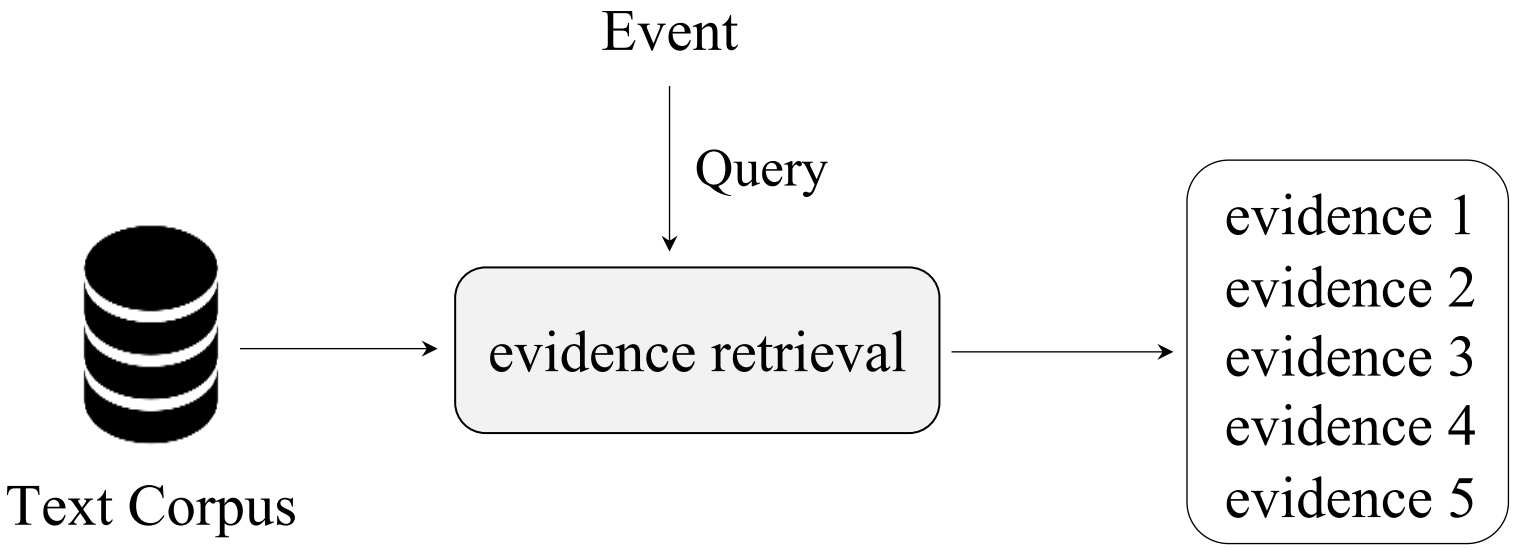



## Advantages

(1) Avoiding the problem of posterior collapse.

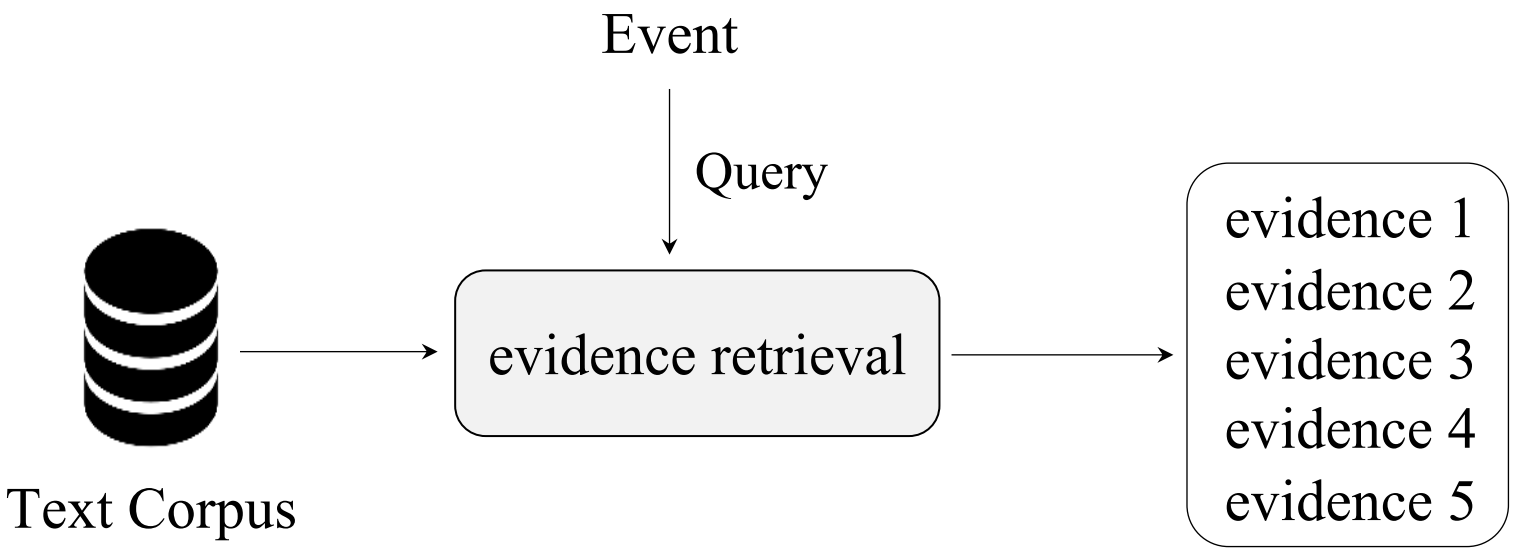
(2) Convenient visualization.

# Evidence Retrieval and Selection




(1) Text Corpus: BookCorpus [Zhu et al., 2015] of 11,038 story books.

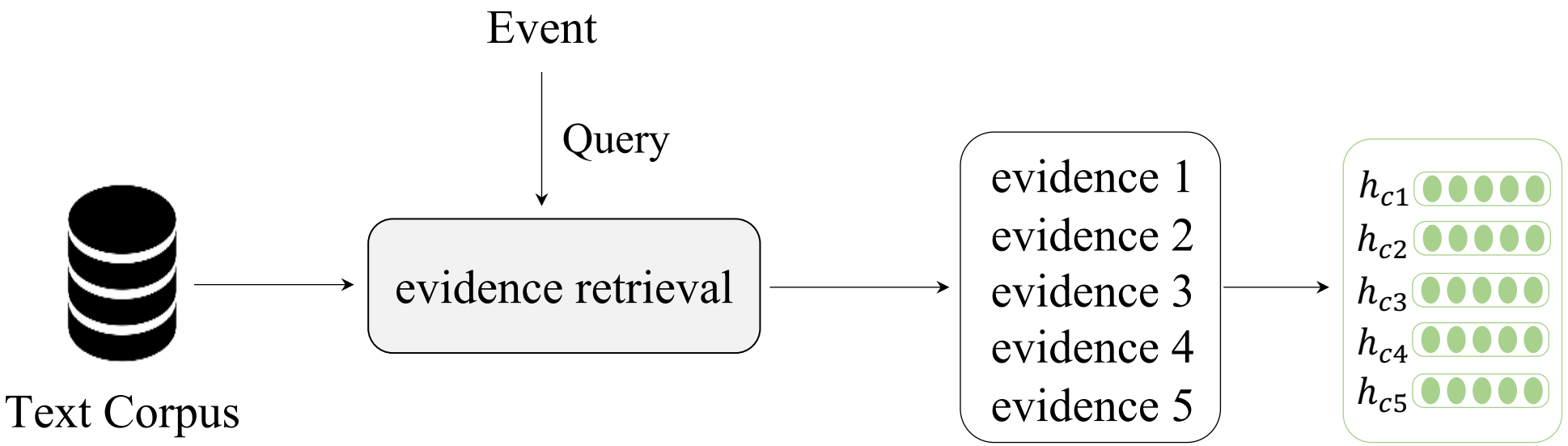
# Evidence Retrieval and Selection




(1) Text Corpus: BookCorpus [Zhu et al., 2015] of 11,038 story books.

(2) Retrieve evidence from corpus by Elastic Search engine.

# Evidence Retrieval and Selection



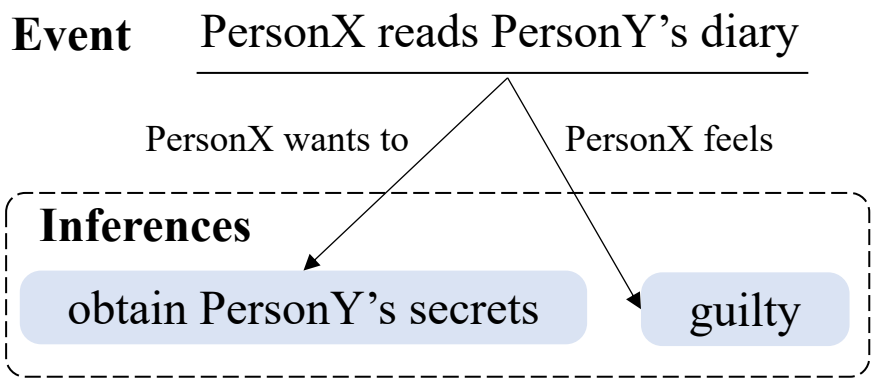

(1) Text Corpus: BookCorpus [Zhu et al., 2015] of 11,038 story books.

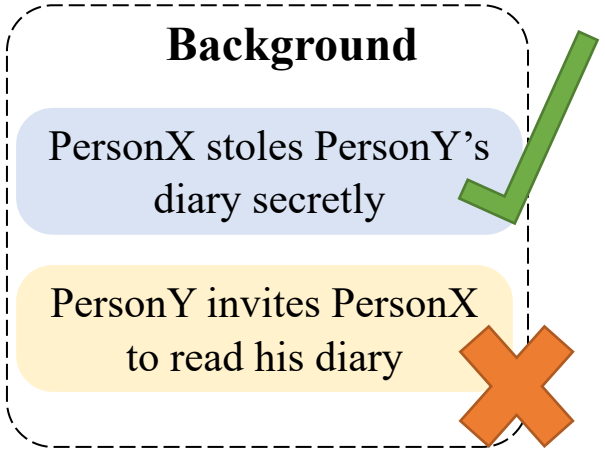
(2) Retrieve evidence from corpus by Elastic Search engine.

(3) Transformer with two layers is used to encode retrieved evidence.

# Evidence Retrieval and Selection

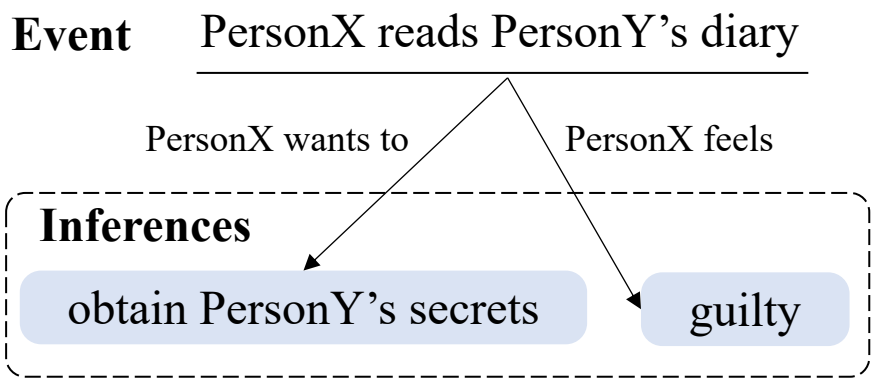
(1) The relevance of evidence is different depending on the semantic of inference.

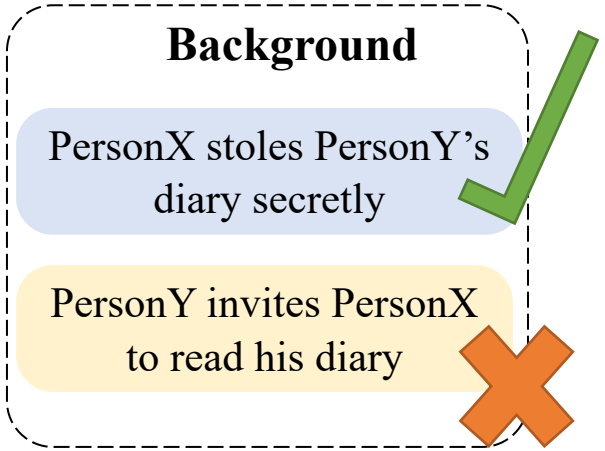


# Evidence Retrieval and Selection

(1) The relevance of evidence is different depending on the semantic of inference.




(2) Since targets are unseen in the inference phase, we utilize the latent variable $z$ to select evidence.

$$p_s(c_k|z) = \begin{cases} 1 \;\; if \;\; k = \arg\min_j ||h_{c_j} - z||_2 \\ 0 \;\; otherwise \end{cases}$$

# Evidence-Aware Generator

Pre-trained language model GPT-2 as our generator:

(1) Given an event $x$, we first sample a latent variable $z$ from the codebook $p(z|x)$.

(2) We then select relevant evidence $c$ according to the context distribution $p(c|z)$.

(3) Finally, the generator $p(y|x,c)$ concatenate the event and selected evidence as the input and generates the inference $y$

# Dataset

(1) We conduct experiments on Event2Mind and ATOMIC datasets.

(2) Both datasets contain about 25,000 unique events.

| Event | Inference dim | Description | Target |
|---|---|---|---|
| PersonX visits friends | xIntent | because PersonX wanted to | to enjoy their time,<br>to catch up with them |
| | xNeed | before that, PersonX needed to | to go to their location,<br>to call them |
| | xAttr | PersonX is seen as | friendly,<br>sociable |
| | xEffect | has an effect on PersonX | have a nice party,<br>have good dinner |
| | xWant | as a result, PersonX wants | have fun,<br>enjoy and spend time |
| | xReact | as a result, PersonX feels | happy,<br>comfortable |
| | oReact | as a result, others feel | happy,<br>pleased |
| | oWant | as a result, others want | to wind down,<br>to clean their home |
| | oEffect | has an effect on others | make the relation stronger,<br>bring a guest into their home |

Table 7: Examples of ATOMIC dataset, including nine inference dimensions. For inference dimensions, "x" and "o" refers to PersonX and others, respectively (e.g. description of "xIntent": *Because PersonX wants*)..

# Experimental Results

## Automatic Evaluation

(1) Accuracy: average BLEU-2 score.

| Methods | xIntent | xNeed | xAttr | xEffect | xReact | xWant | oEffect | oReact | oWant | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| Single Task | | | | | | | | | | |
| S2S | 8.17 | 12.35 | 2.96 | 5.26 | 3.43 | 13.44 | 6.42 | 4.09 | 7.08 | 7.02 |
| VRNMT | 9.52 | 13.35 | 4.87 | 4.42 | 7.64 | 9.80 | 13.71 | 5.28 | 10.79 | 8.82 |
| CWVAE | 12.12 | 15.67 | 5.63 | 14.64 | 8.13 | 15.01 | 11.63 | 8.58 | 13.83 | 11.69 |
| Multi Task | | | | | | | | | | |
| S2S* | 24.53 | 23.85 | 5.06 | 9.44 | 5.38 | 24.68 | 7.93 | 5.60 | 21.30 | 14.20 |
| COMET* | 25.82 | 25.54 | 5.39 | 10.39 | 5.36 | **26.41** | 8.43 | 5.65 | 21.96 | 15.00 |
| COMET | - | - | - | - | - | - | - | - | - | 15.10 |
| EA-VQ-VAE | **26.89** | **25.95** | **5.72** | **10.96** | **5.68** | 25.94 | **8.78** | **6.10** | **22.48** | **15.40** |

Table 1: BLEU score on nine inference dimensions of the ATOMIC test dataset with different approaches. For inference dimensions, "x" and "o" refers to PersonX and others, respectively (e.g. "xAttr": attribute of PersonX, "oEffect": effect on others). The tag (*) means re-implementation.

| Methods | xIntent | xReact | oReact | Overall |
|---|---|---|---|---|
| Single Task | | | | |
| S2S | 2.75 | 2.11 | 5.18 | 3.35 |
| VRNMT | 4.81 | 3.94 | 6.61 | 4.03 |
| CWVAE | 12.98 | 5.65 | 6.97 | 8.53 |
| Multi Task | | | | |
| S2S* | 19.18 | 4.81 | 4.29 | 9.43 |
| COMET* | 21.64 | 5.10 | 4.36 | 10.37 |
| EA-VQ-VAE | **23.39** | **5.74** | **4.81** | **11.31** |

Table 2: BLEU score on three inference dimensions of the Event2Mind test dataset with different approaches. For inference dimensions, "x" and "o" refers to PersonX and others, respectively. The tag (*) means re-implementation.

# Experimental Results

## Automatic Evaluation

(1) Accuracy: average BLEU-2 score.

(2) Diversity: the number of distinct unigrams (dist-1) and bigrams (dist-2).

| Methods | Event2Mind | | ATOMIC | |
|---|---|---|---|---|
| | dist-1 | dist-2 | dist-1 | dist-2 |
| S2S* | 638 | 1,103 | 2,193 | 5,761 |
| COMET* | 1,794 | 4,461 | 3,629 | 12,826 |
| EA-VQ-VAE | **1,942** | **4,679** | **3,918** | **14,278** |

Table 3: The number of distinct n-gram (dist-1 and dist-2) overall on Event2Mind and ATOMIC test dataset with different multi-task learning based methods. The tag (*) means re-implementation.

# Experimental Results

## Automatic Evaluation

(1) Accuracy: average BLEU-2 score.

(2) Diversity: the number of distinct unigrams (dist-1) and bigrams (dist-2).

## Human Evaluation

(1) Sample 100 examples from the test set.

(2) Generate 10 candidates from different models.

(3) Ask five human to identify.

| Methods | Event2Mind | ATOMIC |
|---|---|---|
| S2S* | 0.3901 | 0.5174 |
| COMET* | 0.4874 | 0.6379 |
| EA-VQ-VAE | **0.5072** | **0.6528** |

Table 4: Human score (accuracy) of generations on Event2Mind and ATOMIC test dataset. The tag (*) means re-implementation.

# Ablation Study & Analysis

| Methods | xIntent | xReact | oReact | Overall |
|---|---|---|---|---|
| EA-VQ-VAE | 23.37 | 5.83 | 4.87 | 11.32 |
| - w/o evidence | 21.69 | 5.36 | 4.48 | 10.51 |
| - w/o VQ-VAE | 21.87 | 5.41 | 4.60 | 10.63 |
| - w/o SL | 21.95 | 5.54 | 4.57 | 10.69 |

Table 5: BLEU score on the Event2Mind dev dataset with different approaches. SL is short for separately learning.

# Ablation Study & Analysis

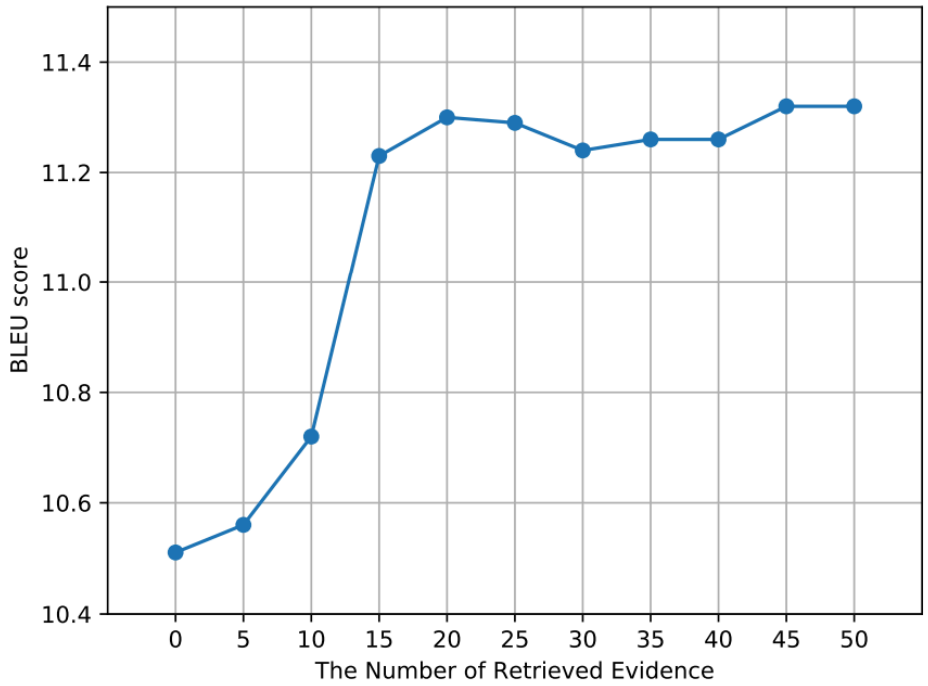



Figure 4: Overall performance with different number of retrieved evidence on Event2Mind dev dataset.

# Case Study

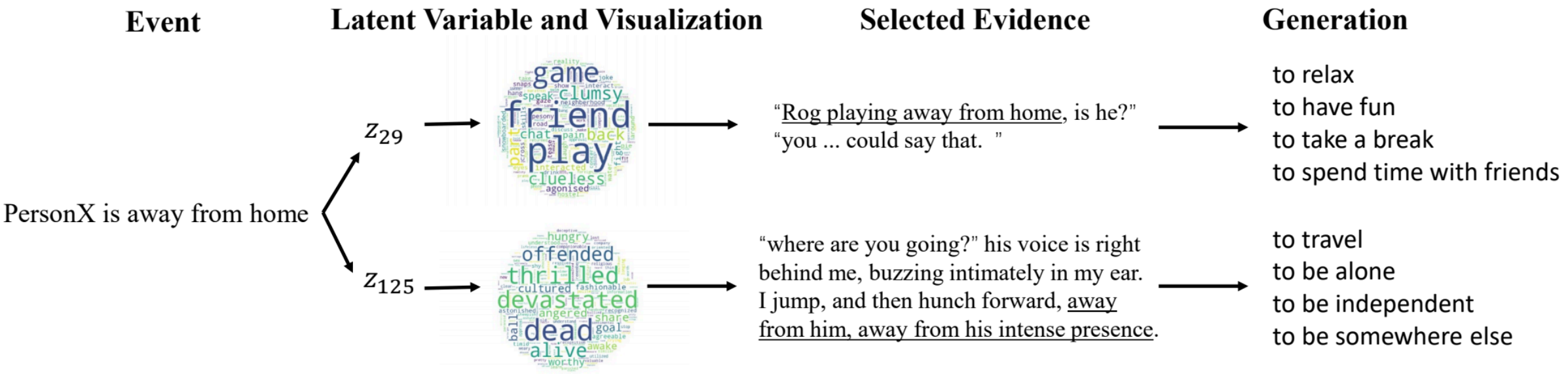



Figure 5: An examples of Event2Mind dataset on the xIntent dimension (i.e. *"PersonX wants"*).

# Conclusion

(1) We present an evidence-aware generative model based on VQ-VAE, which automatically finds evidence as background knowledge to guide the generation.

(2) In the task of inferential text generation, our approach achieves state-of-the-art performances on ATOMIC and Event2Mind datasets.

# Thanks