

# Knowledge Graph-Augmented Language Models for Knowledge-Grounded Dialogue Generation

Anonymous authors  
Paper under double-blind review

## Abstract

Language models have achieved impressive performances on dialogue generation tasks. However, when generating responses for a conversation that requires factual knowledge, they are far from perfect, due to an absence of mechanisms to retrieve, encode, and reflect the knowledge in the generated responses. Some knowledge-grounded dialogue generation methods tackle this problem by leveraging facts from Knowledge Graphs (KGs); however, they do not guarantee that the model utilizes a relevant piece of knowledge from the KG. To overcome this limitation, we propose **SUB**graph **R**etrieval-augmented **G**eneration (**SURGE**), a framework for generating context-relevant and knowledge-grounded dialogues with the KG. Specifically, our SURGE framework first retrieves the relevant subgraph from the KG, and then enforces consistency across facts by perturbing their word embeddings conditioned by the retrieved subgraph. Then, we utilize contrastive learning to ensure that the generated texts have high similarity to the retrieved subgraphs. We validate our SURGE framework on OpendialKG and KOMODIS datasets, showing that it generates high-quality dialogues that faithfully reflect the knowledge from KG.

## 1 Introduction

Dialogue systems aim to engage in ongoing conversations with humans by generating human-like responses. While pre-trained language models (PLMs) (Radford et al., 2019; Raffel et al., 2020) are capable of generating fluent responses, they often generate factually incorrect responses due to a lack of explicit knowledge (Shuster et al., 2021). To overcome such limitations, recent methods access external knowledge sources such as Wikipedia (Dinan et al., 2019), Web (Komeili et al., 2021), Freebase (Bollacker et al., 2008) and Wikidata (Vrandečić & Krötzsch, 2014) to retrieve the relevant knowledge for the dialogue context. In this work, we focus on the Knowledge Graphs (KGs)-based dialogue generation as existing works (Tuan et al., 2019; Wu et al., 2020; Zhang et al., 2020a; Cui et al., 2021; Zhou et al., 2021; Galetzka et al., 2021; Li et al., 2022). KGs represent facts in the most compact and effective symbolic structured form (See the leftmost of Figure 1), consisting of entities as nodes and their relation as an edge. Each of them is formed with a triplet, which can help generate knowledge-grounded responses.

Most of the dialogue generation models with KGs (Galetzka et al., 2021; Li et al., 2022) utilize all the triplets associated with entities in the dialogue context. However, through observation, we found not all of the facts are actually relevant to the conversation (e.g., Jane Austen was born in Steventon in Figure 1), which could mislead the models to generate factually incorrect responses. 87% of facts from 1-hop KG are irrelevant to the context in the OpendialKG dataset (Moon et al., 2019). Moreover, encoding all the facts including the unnecessary ones is computationally inefficient (Galetzka et al., 2021; Rony et al., 2022). On the other hand, even after correctly retrieving the relevant facts, it is not straightforward to combine two heterogeneous modalities: the dialogue context is represented as a *text*; the knowledge is represented as a *graph*. In other words, since PLMs already have tons of pre-trained parameters trained on the unstructured texts, properly conditioning the structured graph to PLMs is highly challenging. Otherwise, PLMs may generate inconsistent responses disregarding the knowledge from the retrieved subgraph, which is a phenomenon known as hallucination (Rohrbach et al., 2018), where they generate responses with their own memorized yet incorrect knowledge.

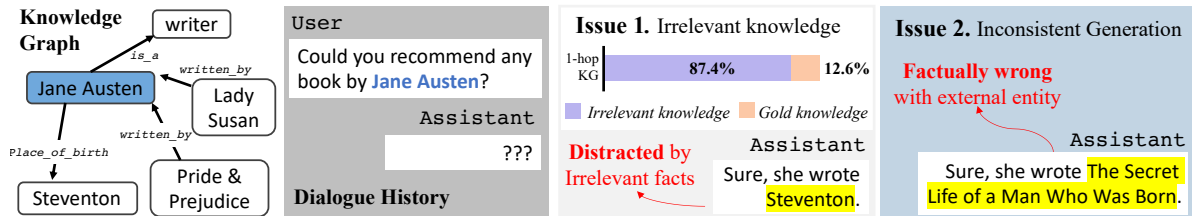


Figure 1: **Motivation.** Existing knowledge-grounded dialogue generation models with a KG often utilize the multi-hop subgraph associated with entities in the dialogue context (e.g., **Jane Austen**). However, they suffer from a couple of the following problems: (1) **irrelevant knowledge** where only 12.6% of facts from the 1-hop KG are useful to generate the target responses given a dialogue context, and (2) **inconsistent generation** where generated texts include the factually incorrect statements.

In this work, we tackle such challenging and fundamental issues of knowledge-grounded dialogue generation with the KGs. We propose an end-to-end dialogue generation framework that considers all aspects from knowledge retrieval, encoding, and reflection along the generation process. As a first step, we propose a context-relevant subgraph retriever that retrieves only the relevant triplets from KGs to prevent the model from generating context-irrelevant responses. Notably, our subgraph retrieval method embeds the KG considering its relational structure with the Graph Neural Network (GNN) (Kipf & Welling, 2017) instead of using PLMs as in previous work (Li et al., 2022). Furthermore, without any labels for pairs of dialogue contexts and their relevant subgraphs, our method is end-to-end trainable jointly with the generation objective by marginalizing the likelihood of the generated sentences over the latent retrieved subgraph (Guu et al., 2020; Lewis et al., 2020b). Then, to encode the retrieved subgraph along with the input text sequence, we propose a graph encoding that is permutation and relation invariant yet efficient. Specifically, we devise the graph encoding method that reflects the graph structure onto the representation space of PLMs, instead of prepending them in front of the text sequence to avoid the computational burden. Furthermore, to ensure that the model does make use of the encoded knowledge when generating responses, we propose a multi-modal contrastive learning objective between two different graph-text modalities to enforce consistency across the retrieved facts and the generated texts. We call our framework **SUB**graph **R**etrieval-augmented **G**eneration (**SURGE**).

We validate our SURGE framework on the OpendialKG (Moon et al., 2019) and KOMODIS (Galetzka et al., 2020) datasets. Note that, when evaluating the generated responses from dialogue models, conventional metrics (e.g., BLEU (Papineni et al., 2002); Rouge (Lin, 2004)) can not measure how faithfully the generated responses reflect the related knowledge in KGs. Thus, in evaluation, we further introduce an additional performance metric, referred to as Knowledge-verifying Question Answering (KQA), which evaluates whether the generated responses contain the correct knowledge with an additional extractive question answering scheme. The experimental results show that SURGE generates responses that not only agree with the gold knowledge but are also consistent with the retrieved knowledge from KGs. Our contributions are summarized as follows:

- We propose a GNN-based context-relevant subgraph retriever that extracts the context-relevant piece of knowledge from KGs, for generating appropriate responses to the ongoing conversation.
- We propose an invariant yet efficient graph encoder and a graph-text contrastive learning objective to ensure that the generated responses faithfully reflect the retrieved knowledge.
- We validate our SURGE framework against relevant baselines, demonstrating its efficacy in generating responses that are more informative by retrieving and reflecting the relevant knowledge.

## 2 Related Work

**Language Models** Pre-trained Language Models (PLMs) (Radford et al., 2019; Lewis et al., 2020a; Raffel et al., 2020) that use a Transformers-based (Vaswani et al., 2017) encoder-decoder architecture have achieved great successes on language generation tasks. As they can accurately contextualize the given context and then generate human-like sentences, they are recently used as the base architecture for neural dialogue

systems (Zhang et al., 2020b; Hosseini-Asl et al., 2020). Moreover, when PLMs become larger, dialogue models are capable of generating high-quality responses (Adiwardana et al., 2020), suggesting that pre-trained parameters do contain certain knowledge (Petroni et al., 2019). However, despite the fluency of such dialogue methods, they often generate factually incorrect responses that are unfaithful to the context but look plausible – widely known as hallucination (Maynez et al., 2020). To tackle this challenge, recent studies (Roller et al., 2021; Shuster et al., 2021) propose to retrieve knowledge from external sources, and then augment it to dialogue models.

**Dialogue Generation with KGs** Regarding dialogue generation tasks with KGs that we target, Moon et al. (2019) introduce a knowledge-grounded dialogue dataset, where each dialogue turn comes with facts from the large-scale KG. Several following works (Tuan et al., 2019; Wu et al., 2020; Zhang et al., 2020a; Cui et al., 2021; Zhou et al., 2021) suggest sequence-to-sequence models trained from scratch, which focus on generating dialogue by conditioning the output word distribution with entities from the KG. Further, Galetzka et al. (2021) propose an efficient method that encodes all facts in the  $k$ -hop neighbors of entities that appear in the dialogue history, in order to reduce the number of input tokens forwarded in PLMs. On the other hand, Rony et al. (2022) propose to mask out model weights for irrelevant facts in PLMs. However, all of these methods simply match and retrieve *all facts* for entities that appear in the dialogue context, which either may mislead models to generate out-of-context responses from irrelevant facts, or can increase the computational overheads from prepending all tokens for all facts in PLMs. Our work differs from them since we aim at retrieving only a context-relevant subgraph among all associated facts with its retriever, which is end-to-end trainable along with a generative model.

### 3 Method

In this section, we first discuss the basic ingredients: Transformer and Graph Neural Network. We then formalize the dialogue generation problem and describe key components for our **SUB**graph **R**etrieval-augmented **GE**neration (**SURGE**) framework: context-relevant subgraph retrieval, invariant graph encoding, and graph-text contrastive learning. Figure 2 illustrates the overview of our framework.

#### 3.1 Preliminaries

As we use two different modalities, namely text and graph, we first define them, and then describe the neural networks to encode them. In particular, a text is defined as a sequence of tokens  $\mathbf{x} = [x_1, \dots, x_N], \forall x_i \in \mathcal{V}$ , where  $x_i$  is a token and  $\mathcal{V}$  is a pre-defined vocabulary formed with specific tokenization algorithms (Sennrich et al., 2016). On the other hand, a knowledge graph (KG) is a type of multi-relational graphs  $\mathcal{G} = \{(\mathbf{e}_h, \mathbf{r}, \mathbf{e}_t)\} \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ , where  $\mathbf{e}_h$  and  $\mathbf{e}_t$  are head and tail entities (nodes) along with their relation (edge)  $\mathbf{r}$ ; and  $\mathcal{E}$  and  $\mathcal{R}$  are sets of entities and relations, respectively.

To easily access different modalities in the same framework, we define the tokenization (mapping) function that maps entities and relations to tokens used in Pre-trained Language Models (PLMs), represented as follows:  $q: \mathcal{E} \cup \mathcal{R} \rightarrow \mathcal{V}^l$  where  $l$  is an arbitrary length varying across different entities and relations. In other words, any entity  $\mathbf{e} \in \mathcal{E}$  and relation  $\mathbf{r} \in \mathcal{R}$  is tokenized to a sequence of  $l$  tokens  $\mathbf{x} \in \mathcal{V}^l$ :  $q(\mathbf{e}) = \mathbf{x}_e$  and  $q(\mathbf{r}) = \mathbf{x}_r$ . For instance, an entity *New York* (i.e.,  $\mathbf{e}$ ), is tokenized into two tokens ‘New’ and ‘York’, i.e.,  $\mathbf{x}_e = [\text{‘New’}, \text{‘York’}]$ .

**Transformer** A Transformer (Vaswani et al., 2017) is the most basic building block of recent PLMs (Devlin et al., 2019; Radford et al., 2019). Given a sequence  $\mathbf{x} = [x_1, \dots, x_N], \forall x_i \in \mathcal{V}$ , generative transformers generate a sequence  $\mathbf{y}_{1:t-1} = [y_1, \dots, y_{t-1}], \forall y_i \in \mathcal{V}$ , with encoder **Enc**, decoder **Dec**, and token embedding  $f$ . A hidden state at time  $t$  for generating  $y_t$  is  $\mathbf{h}_t = \text{Dec}(\text{Enc}(\mathbf{X}), \mathbf{Y}_{1:t-1})$ , where  $\mathbf{X} = f(\mathbf{x}) = [f(x_1), \dots, f(x_N)]$  and  $\mathbf{Y}_{1:t-1} = f(\mathbf{y}_{1:t-1}) = [f(y_1), \dots, f(y_{t-1})]$ . Both **Enc** and **Dec** functions are **permutation sensitive** with positional embedding.

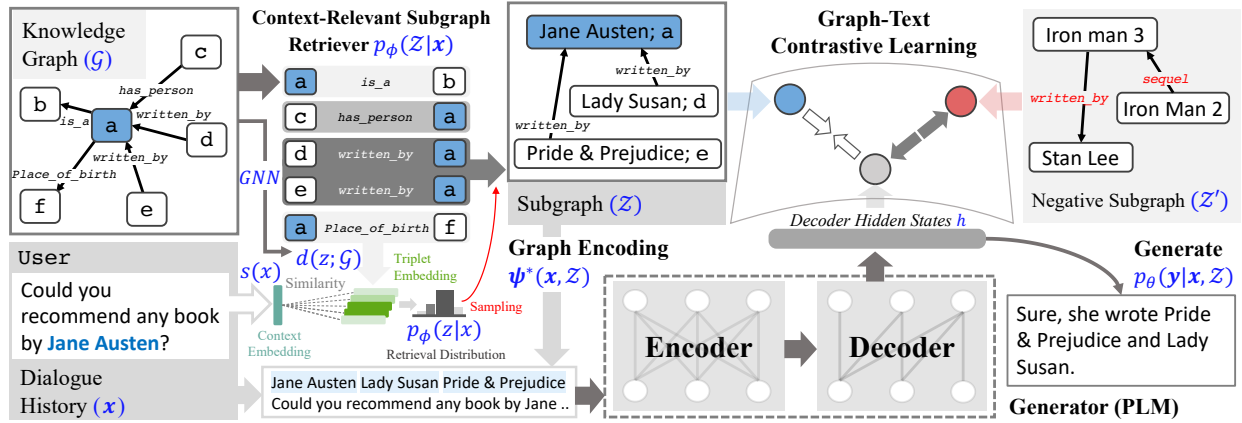


Figure 2: **Framework Overview.** Our framework, SURGE, consists of three parts. First, a context-relevant subgraph retriever  $p_\phi(\mathcal{Z}|\mathbf{x})$  retrieves the subgraph  $\mathcal{Z}$  relevant to the given dialogue history  $\mathbf{x}$  from a knowledge graph  $\mathcal{G}$  (e.g., 1-hop KG from entity *Jane Austen*; **a**). Specifically, we measure the similarity of a context and triplet embedding to compose the retrieval distribution  $p_\phi(\mathcal{Z}|\mathbf{x})$  (§ 3.3). Then, we encode the retrieved subgraph  $\mathcal{Z}$  using the graph encoding  $\psi(\mathbf{x}, \mathcal{Z})$  (§ 3.4). Finally, we use contrastive learning to enforce the model to generate a knowledge-grounded response with the retrieved subgraph (§ 3.5).

**Graph Neural Network** A Graph Neural Network (GNN) represents a node with its neighboring nodes over graphs (Hamilton, 2020), as follows:

$$\text{GNN}(e_t; \mathcal{G}) = \text{UPD}(e_t, \text{AGG}(\{e_h \mid \forall e_h \in \mathcal{N}(e_t; \mathcal{G})\})),$$

where  $\mathcal{N}(e_t; \mathcal{G}) = \{e_h \mid (e_h, \mathbf{r}, e_t) \in \mathcal{G}\}$  is a set of neighboring entities of  $e_t$ ;  $e_t$  and  $e_h$  are embeddings of entities (nodes)  $e_t$  and  $e_h$ ; **AGG** is a function that aggregates embeddings of neighboring entities; and **UPD** is a function that updates  $e_t$  with the aggregated messages from **AGG**.

### 3.2 Problem Statement

Given a dialogue history  $\mathbf{x} = [x_1, \dots, x_N]$ , a model with generative PLMs models a conditional distribution  $p(\mathbf{y}|\mathbf{x})$  to generate an output response  $\mathbf{y} = [y_1, \dots, y_T]$ . To generate knowledge-grounded dialogue, this problem requires a piece of specific knowledge for an ongoing conversation.

To that end, given a dialogue history  $\mathbf{x}$ , we aim at retrieving a subgraph  $\mathcal{Z} \subseteq \mathcal{G}$  consisting of a set of triplets  $z \in \mathcal{Z}$  where  $z = (e_h, \mathbf{r}, e_t)$ , which encodes relevant knowledge for ongoing conversation. Thus, the distribution of the context-relevant facts  $\mathcal{Z}$  is  $p(\mathcal{Z}|\mathbf{x})$ , and our final likelihood of generating responses then becomes  $p(\mathbf{y}|\mathbf{x}, \mathcal{Z})$ . To jointly optimize the objective of graph retrieval with response generation, we treat  $\mathcal{Z}$  as a latent variable and then marginalize the likelihood of the generative model over all possible latent variables for retrieved subgraphs  $\mathcal{Z}$ , formalized as follows:

$$p(\mathbf{y}|\mathbf{x}) = \sum_{\mathcal{Z} \subseteq \mathcal{G}} p_\phi(\mathcal{Z}|\mathbf{x}) p_\theta(\mathbf{y}|\mathbf{x}, \mathcal{Z}) = \sum_{\mathcal{Z} \subseteq \mathcal{G}} p_\phi(\mathcal{Z}|\mathbf{x}) \prod_{t=1}^T p_\theta(y_t|\mathbf{x}, \mathcal{Z}, \mathbf{y}_{0:t-1}), \quad (1)$$

where  $y_0$  is the start token for the generation,  $p_\phi(\mathcal{Z}|\mathbf{x})$  is an output distribution of the context-relevant subgraph retriever, and  $p_\theta(\mathbf{y}|\mathbf{x}, \mathcal{Z})$  is the target distribution of the knowledge-augmented generator, parameterized as  $\phi$  and  $\theta$ , respectively, which we specify in next few subsections.

### 3.3 Context-Relevant Subgraph Retriever

We now provide a concrete description of our context-relevant subgraph retriever, i.e.,  $p(\mathcal{Z}|\mathbf{x})$ , formalized in Equation 1. Given the dialogue history  $\mathbf{x}$ , we assume that retrieval of each triplet in  $\mathcal{Z} = \{z_1, \dots, z_n\}$  is independent. Then, for simplicity, we decompose the retrieval of a set of triplets  $p(\mathcal{Z}|\mathbf{x})$  into the product of

individual triplet retrieval, represented as follows:  $p(\mathcal{Z}|\mathbf{x}) = p(z_1|\mathbf{x})p(z_2|\mathbf{x}) \cdots p(z_n|\mathbf{x})$ , for  $n$  retrieved triples. From this decomposition, it is now sufficient to focus on a single triplet retrieval. We define the score for the single triplet with the inner product between embeddings of the dialogue history  $\mathbf{x}$  and the candidate triplet  $z$  (Guu et al., 2020), as follows:

$$p_\phi(z|\mathbf{x}) \propto \exp(d(z)^\top s(\mathbf{x})), \quad (2)$$

where  $d$  is a triplet embedding function and  $s$  is a dialogue context embedding function. For implementing  $s$ , we can use any off-the-shelf PLMs, but for  $d$ , we need another effective approach that captures the property of the graph. Therefore, we utilize the Graph Neural Networks (GNNs) for the triplet embedding function  $d$  to consider the relational structure between entities in the KG.

More specifically, we consider a set of triplets associated to the entities that appear in the given dialogue context:  $\{(\mathbf{e}, \mathbf{r}, \mathbf{e}_t) \text{ or } (\mathbf{e}_h, \mathbf{r}, \mathbf{e}) \mid q(\mathbf{e}) \subseteq \mathbf{x}\}$ , as the retrieval candidates. Then, to effectively represent triplets consisting of entities and their relations as items, we use GNNs described in Section 3.1. In our triplet retriever, utilizing both nodes and edges, which are equally essential components for the multi-relational graph, is worthwhile to represent an entire triplet. We adopt the existing node (Kipf & Welling, 2017) and edge message passing frameworks (Jo et al., 2021). Formally, our triplet embedding function is denoted as follows:

$$\begin{aligned} d(z; \mathcal{G}) &= \text{MLP}([e_h \parallel \mathbf{r} \parallel e_t]), \quad e_h = \text{GNN}(e_h^0; \mathcal{G}), \\ \mathbf{r} &= \text{GNN}(\mathbf{r}^0; \mathcal{G}^*), \quad e_t = \text{GNN}(e_t^0; \mathcal{G}), \end{aligned} \quad (3)$$

where  $z = (\mathbf{e}_h, \mathbf{r}, \mathbf{e}_t)$ ,  $q(\mathbf{e}) = [\mathbf{x}_i, \dots, \mathbf{x}_j]$ ,  $\mathbf{0}$  is a zero vector, and  $\parallel$  is the concatenation operator. For node embedding  $e^0$ , we reuse the word embedding from Enc, if it exists in  $\mathbf{x}$ . For relation embedding  $\mathbf{r}^0$ , we use the trainable relation embedding matrix. Please refer to Appendix E.1 for more details.

### 3.4 Invariant Graph Encoding

We now specify graph encoding, which aims to condition the structural graph  $\mathcal{Z}$  along with the text sequence  $\mathbf{x}$  over PLMs to generate  $\mathbf{y}$ . Let  $\psi(\mathbf{x}, \mathcal{Z})$  be a graph encoding function. Then, the simplest way to encode graphs into PLMs is to prepend the tokens of entities and relations to the input  $\mathbf{x}$  (Li et al., 2022; Ma et al., 2022). Formally, given a text  $\mathbf{x} = [x_1, \dots, x_N]$  and a graph  $\mathcal{Z} = \{(\mathbf{a}, \mathbf{r}_1, \mathbf{b}), (\mathbf{b}, \mathbf{r}_2, \mathbf{a}), (\mathbf{a}, \mathbf{r}_1, \mathbf{c})\}$ , a naïve graph encoding is defined as follows:  $\psi(\mathbf{x}, \mathcal{Z}) = f([a, r_1, b, b, r_2, a, a, r_1, c, x_1, \dots, x_N])$  where  $a = q(\mathbf{a})$ ,  $r_1 = q(\mathbf{r}_1)$ , and so on. Here  $f$  is a token embedding and  $q$  is a mapping function defined in Section 3.1. However, it violates two important properties for consistent encoding of a multi-relational graph into PLMs: permutation invariance (Zaheer et al., 2017) and relation-inversion invariance, formalized in Definition 3.1 and 3.2 as follows:

**Definition 3.1. (Permutation Invariance)** For any permutation  $\pi \in S_n$ ,  $\psi(\mathbf{x}, \mathcal{Z}) = \psi(\mathbf{x}, \pi \cdot \mathcal{Z})$ .

**Definition 3.2. (Relation Inversion Invariance)** Let  $\neg\mathbf{r}$  be an inverse relation to  $\mathbf{r}$ , if  $(\mathbf{a}, \mathbf{r}, \mathbf{b}) = (\mathbf{b}, \neg\mathbf{r}, \mathbf{a}) \forall \mathbf{a}, \mathbf{b} \in \mathcal{E}$ . Then,  $\psi(\mathbf{x}, \mathcal{Z} \cup \{(\mathbf{a}, \mathbf{r}, \mathbf{b})\}) = \psi(\mathbf{x}, \mathcal{Z} \cup \{(\mathbf{b}, \neg\mathbf{r}, \mathbf{a})\})$  for any subgraph  $\mathcal{Z}$ .

**Invariant Graph Encoding** To satisfy both properties, we consider two operations on a set of triplets up to the naïve encoding. We first define a SORT operator that returns the same output regardless of the order of input set elements, as follows:

$$\text{SORT}(\pi \cdot \mathcal{Z}) = \text{SORT}(\pi' \cdot \mathcal{Z}), \quad \forall \pi, \pi' \in S_n, \quad (4)$$

where  $S_n$  is a set of all possible permutations for  $n$  elements. We then define a INV operator that adds the inverse triplet of each triplet in  $\mathcal{Z}$ , as follows:

$$\text{INV}(\mathcal{Z}) = \mathcal{Z} \cup \{(\mathbf{e}_t, \neg\mathbf{r}, \mathbf{e}_h) \mid (\mathbf{e}_h, \mathbf{r}, \mathbf{e}_t) \in \mathcal{Z}\}. \quad (5)$$

Based on them, our graph encoding function,  $\psi(\mathbf{x}, \text{SORT}(\text{INV}(\mathcal{Z})))$ , satisfies both invariances.

**Invariant and Efficient Graph Encoding** However, the above encoding is not efficient since it requires  $\mathcal{O}(n)$  space complexity with  $n$  triplets. Thus, we newly define  $\tilde{\psi}$  that encodes the sorted sequence of only the unique entities, as follows:

$$\tilde{\psi}(\mathbf{x}, \text{SORT}(\text{ENT}(\mathcal{Z}))) = f([a, b, c, x_1, \dots, x_N]),$$

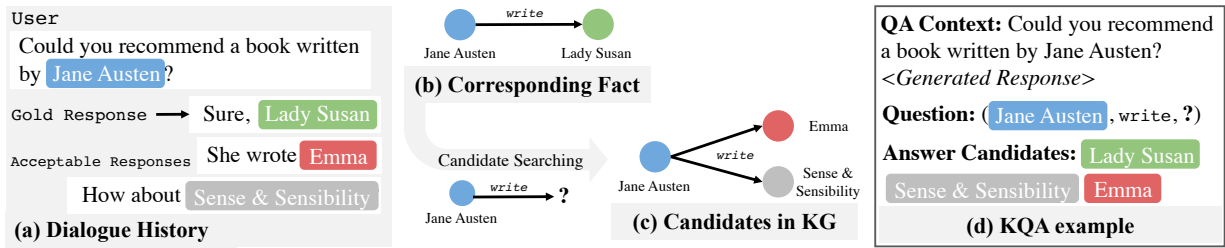


Figure 3: **KQA**. (Left) An example where multiple responses are acceptable. (Middle) In KG, we search for the fact that reflects the relation between entities within the user input and gold response (b), and then search candidate facts from the KG (c). (Right) Corresponding KQA example. If a generated response contains the one of answer candidates, the KQA can predict it (success).

where  $\text{ENT}(\mathcal{Z})$  returns the set of unique entities in  $\mathcal{Z}$ . This encoding meets both invariance properties but also efficient since it only costs  $\mathcal{O}(k)$ , for the  $k$  entity where  $k < n$ . However, as it does not consider the relational information in  $\mathcal{Z}$ , we further perturb the token embeddings of each entity  $f(\cdot)$  in PLMs with respect to their graph representations in  $\mathcal{Z}$ . Specifically, for each entity  $\mathbf{a} \in \text{ENT}(\mathcal{Z})$ , we apply a learnable affine transformation (Perez et al., 2018) on the token embedding of  $\mathbf{a}$  as follows:

$$\begin{aligned} \beta(f(\mathbf{a}), \mathcal{Z}) &= (1 + \gamma) * f(\mathbf{a}) + \delta, \\ \gamma, \delta &= \text{MLP}(\eta), \quad \eta = \text{R-GNN}(f(\mathbf{a}); \mathcal{Z}), \end{aligned} \quad (6)$$

where MLP is a Multi-Layer Perceptron,  $\beta: \mathbb{R}^d \rightarrow \mathbb{R}^d$  perturbs the embedding according to  $\mathcal{Z}$ , R-GNN is the relation-aware GNN (Vashishth et al., 2020). In sum, we denote a relation-aware and invariant yet efficient encoding  $\psi^*$ , defined as follows:

$$\psi^*(\mathbf{x}, \mathcal{Z}) = \beta(\tilde{\psi}(\mathbf{x}, \text{SORT}(\text{ENT}(\mathcal{Z}))), \text{INV}(\mathcal{Z})).$$

We conclude that our graph encoding satisfies both properties. For further details on the proof and comprehensive illustration, please refer to Appendix D.

### 3.5 Knowledge-Grounded Generation with Graph-Text Contrastive Learning

Our framework now can retrieve and encode the context-relevant subgraph given the user input. Then, reflecting the subgraph into the model is important when generating a knowledge-grounded response. The generative model should be able to generate different sequences when providing different subgraphs, for the same dialogue history.

However, we only access the single ground-truth response regardless of the retrieved knowledge, while the generative model is trained with a teacher forcing. Thus, this setting can raise the problem of *exposure bias* (Ranzato et al., 2016): the model is never exposed to other generated tokens during training. To overcome such limitations, we introduce a graph-text contrastive learning method. Formally, for a single pair of a graph and text, the contrastive learning objective is defined as follows:

$$\mathcal{L}_{cont} = \frac{1}{2} \log \frac{\exp(\text{sim}(\zeta(\mathbf{z}), \xi(\mathbf{h}))/\tau)}{\sum_{\mathbf{h}'} \exp(\text{sim}(\zeta(\mathbf{z}), \xi(\mathbf{h}'))/\tau)} + \frac{1}{2} \log \frac{\exp(\text{sim}(\zeta(\mathbf{z}'), \xi(\mathbf{h}))/\tau)}{\sum_{\mathbf{z}'} \exp(\text{sim}(\zeta(\mathbf{z}'), \xi(\mathbf{h}))/\tau)}, \quad (7)$$

where  $\mathbf{z} = \frac{1}{m} \sum_{i=1}^m \tilde{\mathbf{z}}_i$  is the average representations of the graph from  $\text{Enc}(\psi^*(\mathbf{x}, \mathcal{Z})) = [\tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_m, \mathbf{z}_1, \dots, \mathbf{z}_N]$ ,  $\mathbf{h} = \frac{1}{T} \sum_{t=1}^T \mathbf{h}_t$  is the mean of decoder representations,  $\text{sim}$  is the cosine similarity,  $\zeta$  and  $\xi$  are learnable linear projection layers, and  $\tau$  is a learnable temperature parameter. Furthermore,  $\sum_{\mathbf{h}'}$  and  $\sum_{\mathbf{z}'}$  indicate the summation over negative samples, which are other texts or graphs within a same mini-batch.

### 3.6 Training

We train the entire model by maximizing the log-likelihood  $\log p(\mathbf{y}|\mathbf{x})$  defined in Equation 1 with respect to parameters of both the retriever  $\phi$  and the generator  $\theta$ . Since computing the marginal probability over entire

Table 1: Experimental results on the OpendiaKG dataset with the T5-small model. † indicates the model under the oracle setting, which uses the gold facts even in the test time. The best results are emphasized in bold.

Method	KQA		BLEU				ROUGE			Unigram	
	EM	F1	B-1	B-2	B-3	B-4	R-1	R-2	R-L	F1	
<i>Baselines</i>	No Knowledge	12.25	20.69	15.79	9.19	5.61	3.43	19.67	7.13	19.02	22.21
	All Knowledge	43.58	50.60	15.95	9.98	6.72	4.65	20.96	8.50	20.21	24.34
	Space Efficient ( <i>series</i> )	36.60	42.64	16.15	10.03	6.66	4.50	21.15	8.56	20.44	24.55
	Space Efficient ( <i>parallel</i> )	38.54	44.34	16.33	10.22	6.81	4.64	21.42	8.85	20.68	24.87
	EARL	32.47	35.88	11.49	6.34	4.06	2.75	15.36	4.37	14.61	16.88
	DiffKG	12.25	20.99	15.68	9.13	5.60	3.46	19.50	7.07	18.84	22.26
<i>Retrieval variants</i>	Random Retrieval	31.72	38.95	15.70	9.52	6.12	3.99	20.21	7.88	19.55	23.28
	Sparse Retrieval (BM25)	29.50	36.96	15.63	9.44	6.05	3.96	20.05	7.67	19.37	23.10
	Dense Retrieval (Bi-encoder)	46.17	52.52	16.67	10.44	7.05	4.91	20.41	8.38	19.66	23.85
	Dense Retrieval (Poly-encoder)	46.05	52.57	17.56	11.01	7.45	5.18	20.66	8.46	19.87	24.24
<i>Ours</i>	SURGE ( <i>unsupervised</i> )	48.49	55.77	<b>17.77</b>	<b>11.30</b>	<b>7.69</b>	<b>5.36</b>	<b>21.64</b>	<b>9.14</b>	<b>20.75</b>	<b>25.24</b>
	SURGE ( <i>semi-supervised</i> )	<b>51.00</b>	57.63	17.70	11.21	7.61	5.28	21.43	8.85	20.57	25.07
	SURGE ( <i>contrastive</i> )	50.45	<b>57.70</b>	17.29	11.04	7.54	5.28	21.35	8.98	20.48	25.10
<i>Oracle</i>	Gold Knowledge†	63.32	67.90	18.47	12.79	9.32	6.92	24.93	11.97	24.03	28.82
	Gold Response	93.30	95.21	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 2: Experimental results on the KOMODIS dataset with the T5-small model. For full experimental results, please see Supplementary Files.

Method	KQA		BLEU		
	EM	F1	B-1	B-2	F1
Random	12.41	14.17	7.74	4.02	16.29
SE ( <i>series</i> )	12.41	14.70	8.34	5.13	17.37
SE ( <i>parallel</i> )	16.46	18.70	9.33	5.66	17.72
SURGE (Ours)	<b>17.30</b>	<b>19.50</b>	<b>11.51</b>	<b>7.18</b>	<b>19.51</b>

Table 3: Knowledge-grounded generation results by using the modified gold subgraphs instead of the retrieved ones, to evaluate the efficacy of contrastive learning, with F1 and KF1 as metrics.

Method	F1	KF1
SURGE (unsupervised)	27.78	24.09
SURGE (semi-supervised)	<b>28.30</b>	26.38
SURGE (contrastive)	28.17	<b>27.58</b>

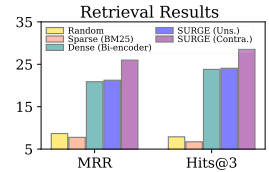


Figure 4: Fact retrieval results on the OpendiaKG dataset, with the metrics of MRR and Hits@3.

subgraphs is infeasible, we approximate it by summing over  $k$  sampled subgraphs (Gua et al., 2020; Lewis et al., 2020b). Our end-to-end training objective for retrieval-augmented generation is then defined as follows:

$$\mathcal{L}_{ret} = \log \sum_{\mathcal{Z} \subseteq \Pi} p_{\phi}(\mathcal{Z}|\mathbf{x}) p_{\theta}(\mathbf{y}|\mathbf{x}, \mathcal{Z}), \quad (8)$$

where  $\Pi = \text{samplek}(p_{\phi}(\cdot|\mathbf{x}))$  denotes sampling  $k$  subgraphs over the subgraph distribution and each subgraph sampling is decomposed into sampling  $n$  triplets from  $p_{\phi}(z_i|\mathbf{x}) \forall i \in [1, n]$  as in Section 3.3. We further assume that the gold subgraph  $\mathcal{Z}^*$  is partially available in training. In this case, we utilize the following supervised loss to train the retriever:  $\mathcal{L}_{sup} = \log p_{\phi}(\mathcal{Z}^*|\mathbf{x})$ . By combining all objectives in Equation 7, 8, and  $\mathcal{L}_{sup}$ , our training objective is defined as  $\mathcal{L} = \mathcal{L}_{ret} + \mathcal{L}_{sup} + \mathcal{L}_{cont}$ .

## 4 KQA Metric: Knowledge-verifying QA

Existing automatic evaluation metrics, namely BLEU and ROUGE (Papineni et al., 2002; Lin, 2004), are limited in that they only consider the lexical overlaps of words without measuring the factual correctness. For instance, as shown in Figure 3 (a), there could be multiple correct responses, but existing metrics score them lower due to the lexical mismatch with the gold response. To solve this issue, we propose Knowledge-verifying Question Answering (KQA) which measures whether generated responses contain factually correct knowledge given the dialogue history. To realize this, we formulate extractive QA task (Rajpurkar et al., 2016) by automatically deriving QA pairs from the dialogue and the large-scale KG in each dataset (See Figure 3). Then, we fine-tune BERT (Devlin et al., 2019) on synthetic KQA pairs to build a QA model. To evaluate generated responses, we concatenate the dialogue history and the generated response then forward it into the trained QA model. If the QA model yields the correct answer, we regard this case as the generated response containing accurate knowledge. See Appendix E.1 for details.

## 5 Experiment

### 5.1 Experimental Setup

We conduct experiments on **OpendialKG** (Moon et al., 2019), which consists of 15K dialogues with 91K utterances associated with a large-scale KG, namely Freebase (Bollacker et al., 2008) with 100k entities and 1M facts. We note that, among them, 49% of the utterances come with gold knowledge, whereas others are not. We randomly split the dataset into training (70%), validation (15%), and test sets (15%). We also use **KOMODIS** (Galetzka et al., 2020), which contains 7.5K dialogues associated with the KG having 88k facts. As retrieval candidates, we use 1-hop KG associated with entities in the given dialogue for OpendialKG and 2-hop KG for KOMODIS. Except Table 2, most of the experiments are conducted on OpendialKG. We use **T5-small** (Raffel et al., 2020) for all experiments. For more details, see Appendix E.

### 5.2 Baselines and Our Models

**No Knowledge.** The model only with the dialogue history. **All Knowledge.** The model with entire facts within a  $k$ -hop subgraph. **Gold Knowledge.** The model with the exact gold knowledge if it exists. **Space Efficient Encoding.** The model with all facts from the  $k$ -hop subgraph. We use two variants from (Galetzka et al., 2021), namely *series* and *parallel*. **EARL.** The latest RNN-based model, where the entities are conditioned in response generation (Zhou et al., 2021). **DiffKG.** PLM-based model with differentiable path traversal (Tuan et al., 2022). **Random/Sparse Retrieval.** The model with selected facts via random sampling or sparse retrieval (Robertson & Zaragoza, 2009). **Dense Retrieval.** A variant of our framework where T5 encoder (Raffel et al., 2020) is used for  $d$  in Equation 3 instead of GNNs similar to Bi- and Poly-encoder (Humeau et al., 2020). **SURGE (unsupervised).** Ours with retrieved context-relevant facts from  $k$ -hop subgraph, where the retrieval is trained without any supervision. **SURGE (semi-supervised).** Ours but the retriever is trained with  $\mathcal{L}_{sup}$  if the gold exists. **SURGE (contrastive).** Ours with  $\mathcal{L}_{ret}$ ,  $\mathcal{L}_{sup}$ , and  $\mathcal{L}_{cont}$  in Section 3.6.

### 5.3 Evaluation Metrics

We use BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and F1 score as metrics. We also use our new metric, KQA (§ 4), which measures whether the generated responses contain proper knowledge. In Table 3, we use Knowledge F1 (KF1) (Shuster et al., 2021) to measure unigram overlaps between the retrieved knowledge and generated response.

### 5.4 Experimental Results and Analysis

In Table 1, we report the knowledge-grounded response generation performances of baselines and our SURGE on OpendialKG. As shown in Table 1, our models significantly outperform all the baseline models, excluding oracles, in all evaluation metrics. The high BLEU, ROUGE, and F1 refer that ours sufficiently learns the syntactic and semantic structure of the responses. Our models also achieve high F1 and EM scores in KQA. The high KQA scores indicate that the generated responses are formed with the correct facts, which are relevant to the dialog context. Even the baseline models such as *All Knowledge*, *Space Efficient Encoding* (Galetzka et al., 2021), *EARL* (Zhou et al., 2021), and *DiffKG* (Tuan et al., 2022), which are provided with all of  $k$ -hop facts, underperform than ours. The result demonstrates that selecting relevant knowledge is critical in knowledge-augmented response generation. In Table 2, we additionally report the experimental results on KOMODIS to show the applicability of our method to other datasets. Our SURGE (*contrastive*) also outperforms other baselines in KOMODIS. For results with all metrics, please see Appendix F.3.

**Knowledge Retrieval** Figure 4 shows performances of retrievers, for which we measure the performance on 45% of test dialogues containing the gold knowledge, with Mean Reciprocal Rank (MRR) and Hits@k as metrics. Our models outperform all baselines. Further, our model with contrastive learning and semi-supervised retriever training outperforms an unsupervised version. See Appendix H for examples.

Table 4: Examples of responses from the baseline (Space Efficient, parallel) and responses from SURGE.

Context	Gold response	Baseline response	SURGE response
(a) I loved Moby Dick. Can you recommend something similar?	It was written by Herman Melville in 1851. It's sometimes called The Whale.	Moby Dick is a sailor. Do you like her work?	Moby Dick was written by Herman Melville. He also wrote The Whale.
(b) Do you know anything the actor Adam Brown?	Yes, he was in the movie The Hobbit: An Unexpected Journey.	Adam Brown starred in King Kong. Have you seen it?	Adam Brown starred in The Hobbit: The Desolation of Smaug and The Hobbit: The Battle of the Five Armies.
(a) Retrieved Subgraph from SURGE (Moby Dick; The Whale, written_by, Herman Melville) (Moby Dick, written_by, Norman Corwin) (Moby Dick, written_by, Ray Bradbury))		(b) Retrieved Subgraph from SURGE (The Hobbit: The Battle of the Five Armies, starred_actors, Adam Brown) (The Hobbit: An Unexpected Journey, starred_actors, Adam Brown) (The Hobbit: The Desolation of Smaug, starred_actors, Adam Brown)	

Table 5: Experimental results on OpendialKG with additional three metrics other than KQA for measuring whether the generated responses contain appropriate knowledge.

Method	KQA		Knowledge F1	Entity F1	String Matching
	EM	F1			
All Knowledge	43.58	50.60	18.91	21.10	44.25
Space Efficient (Parallel)	38.54	44.34	17.43	18.93	40.56
Dense Retrieval (Poly-encoder)	46.05	52.57	19.72	21.46	48.41
SURGE (ours, semi-supervised)	<b>51.00</b>	<b>57.63</b>	<b>21.87</b>	<b>23.03</b>	<b>55.79</b>
Gold Response (oracle)	93.30	95.21	28.62	29.06	85.75

Table 6: Performance comparisons of variants of graph encodings, described in Section 3.4. (Inv.= Invariant)

Method	KQA F1	Knowledge Length
Naïve Encoding	55.23	68.21
Invariant	56.78	121.82
Inv. & Efficient (entity only)	49.99	15.42
Inv. & Efficient (w/ Equation 6)	<b>57.63</b>	<b>15.42</b>

Table 7: Human evaluation on Consistency, Informativeness, and Fluency with bold marks  $p < 0.05$ .

Method	Consis.	Info.	Fluency
All Knowledge	<b>2.52</b>	1.99	2.62
Space Efficient	2.47	1.75	2.46
SURGE (ours)	<b>2.71</b>	<b>2.39</b>	<b>2.92</b>

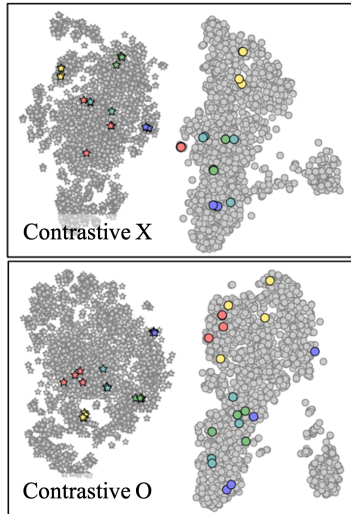


Figure 5: Embedding visualization of graph (star) - text (circle) contrastive learning.

**Knowledge-Grounded Generation** We conduct an ablation study on our models to validate the knowledge consistency performance of the response generation by computing the Knowledge F1 (KF1) score (Shuster et al., 2021). We use the gold knowledge rather than the retrieved one to focus solely on the case where a given knowledge is consistently reflected in the generated responses. We randomly modify the tail entity of each gold knowledge to ensure that responses are generated from the given knowledge rather than the trained knowledge. Table 3 shows that our model with a contrastive learning term outperforms all others in the KF1, implying that the generated responses accurately reflect the encoded knowledge.

**Retrieval and Generation Examples** Table 4 shows the examples of generated responses along with the retrieved knowledge. We compare our SURGE against *Space Efficient (parallel)* baseline. In example (a), the baseline response contains an incorrect fact distracted by the contextually irrelevant entity ‘sailor’. Contrarily, SURGE successfully retrieves relevant facts from the KG and then generates the factually correct response. In (b), the baseline generates the response with a wrong fact, meanwhile, SURGE retrieves context-relevant facts and generates an informative response.

**Automatic Evaluations on Knowledge Groundedness** In Table 5, we measure Knowledge F1 (KF1 in Table 3), string matching (check whether at least one of answer entities exists the generated response), and entity F1 (measuring F1 score with each entity in answer candidates) for representative baselines and our SURGE (semi-supervised) in OpendialKG. For KF1, we measure the F1 score regarding the concatenation of

the question (head entity and relation) and all answer candidates (available tail entities) in KQA as the gold response. The results show that all metrics show the same tendency with KQA and our proposed method still outperforms other baselines by generating responses with more proper knowledge. See Appendix F.6 for more details.

**Sensitive Analysis on Graph Encoding** We conduct an analysis on graph encoding variants introduced in Section 3.4. The knowledge length in Table 6 indicates the average token length used for graph encoding. Our encoding  $\psi^*$  with Equation 6 performs the best against other variants while using the lesser space at the graph encoding phase.

**Human Evaluation** We sample 30 responses of SURGE, *All Knowledge*, and *Space Efficient* on the test set of OpendialKG, then conduct a human study of them. We recruit 46 annotators and ask them to evaluate the quality of the generated responses with consistency, informativeness, and fluency criteria using a 3-point Likert-like scale. As shown in Table 7, ours obtains significantly higher scores than others in all criteria, which is another evidence that our framework generates consistent, informative, and fluent responses. We observe that the informativeness score and KQA F1 score have a 0.42 Pearson correlation coefficient. This allows us to confirm that our KQA metric positively correlates with the human evaluation results.

**Embedding Space Visualization** We further visualize the latent space of graph and text learned from Equation 7 in Figure 5. The visualization shows that, for the same dialogue with different subgraphs, our SURGE with graph-text contrastive learning (right) generates distinct response embeddings pertaining to different subgraphs, unlike the one without contrastive learning which shows less variety over responses for the same dialogue (left). We include zoomed Figure 5 in Appendix H.

## 6 Conclusion

In this work, we proposed a novel end-to-end framework for knowledge-grounded dialogue generation which retrieves context-relevant subgraph, encodes a subgraph with the text, and generates natural and informative responses based on the retrieved subgraph, called as **SUB**graph **R**etrieval-augmented **G**eneration (**SURGE**). Our results demonstrate the effectiveness of our framework in both quantitative and qualitative experiments in knowledge retrieval and response generation tasks. The analysis shows the contribution of each proposed component: retrieval, encoding, and graph-text representation learning. Our work suggests a new direction to generate informative responses for knowledge graph-based dialogue task by empirically showing the importance of retrieving the more relevant subgraph knowledge rather than using all the relevant knowledge graphs when generating knowledge-grounded responses.

## Acknowledgments

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST)) and Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

## References

- Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977, 2020.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International*

- Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pp. 1247–1250, 2008.
- Fuwei Cui, Hui Di, Hongjie Ren, Kazushige Ouchi, Ze Liu, and Jinan Xu. Syntactically diverse adversarial network for knowledge-grounded conversation generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pp. 4620–4630. Association for Computational Linguistics, 2021. URL <https://doi.org/10.18653/v1/2021.findings-emnlp.394>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. URL <https://doi.org/10.18653/v1/n19-1423>.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
- Fabian Galetzka, Chukwuemeka Uchenna Eneh, and David Schlangen. A corpus of controlled opinionated and knowledgeable movie discussions for training neural conversation models. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pp. 565–573. European Language Resources Association, 2020. URL <https://aclanthology.org/2020.lrec-1.71/>.
- Fabian Galetzka, Jewgeni Rose, David Schlangen, and Jens Lehmann. Space efficient context encoding for non-task-oriented dialogue generation with graph attention transformer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 7028–7041. Association for Computational Linguistics, 2021.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020.
- William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159, 2020.
- Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
- Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL <https://openreview.net/forum?id=SkxggnNFvH>.
- Jaehyeong Jo, Jinheon Baek, Seul Lee, Dongki Kim, Minki Kang, and Sung Ju Hwang. Edge representation learning with hypergraphs. *CoRR*, abs/2106.15845, 2021.
- Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

- Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. *CoRR*, abs/2107.07566, 2021.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 7871–7880. Association for Computational Linguistics, 2020a.
- Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 110–119. The Association for Computational Linguistics, 2016. URL <https://doi.org/10.18653/v1/n16-1014>.
- Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. Knowledge-grounded dialogue generation with a unified knowledge representation. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 206–218. Association for Computational Linguistics, 2022. URL <https://aclanthology.org/2022.naacl-main.15>.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013>.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Kaixin Ma, Hao Cheng, Xiaodong Liu, Eric Nyberg, and Jianfeng Gao. Open domain question answering with A unified knowledge interface. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 1605–1620. Association for Computational Linguistics, 2022. URL <https://doi.org/10.18653/v1/2022.acl-long.113>.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 1906–1919. Association for Computational Linguistics, 2020.
- Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 845–854. Association for Computational Linguistics, 2019.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pp. 311–318. ACL, 2002.

- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, 2019.
- Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 3942–3951. AAAI Press, 2018. URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16528>.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 2463–2473. Association for Computational Linguistics, 2019. URL <https://doi.org/10.18653/v1/D19-1250>.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 2383–2392. The Association for Computational Linguistics, 2016. URL <https://doi.org/10.18653/v1/d16-1264>.
- Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.
- Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 4035–4045. Association for Computational Linguistics, 2018. URL <https://doi.org/10.18653/v1/d18-1437>.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 300–325. Association for Computational Linguistics, 2021.
- Md. Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. Dialogk: Knowledge-structure aware task-oriented dialogue generation. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 2557–2571. Association for Computational Linguistics, 2022. URL <https://doi.org/10.18653/v1/2022.findings-naacl.195>.

- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pp. 3784–3803. Association for Computational Linguistics, 2021.
- Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. Dykchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 1855–1865. Association for Computational Linguistics, 2019.
- Yi-Lin Tuan, Sajjad Beygi, Maryam Fazel-Zarandi, Qiaozi Gao, Alessandra Cervone, and William Yang Wang. Towards large-scale interpretable knowledge graph reasoning for dialogue systems. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 383–395. Association for Computational Linguistics, 2022. URL <https://doi.org/10.18653/v1/2022.findings-acl.33>.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha P. Talukdar. Composition-based multi-relational graph convolutional networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.
- Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.
- Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *EMNLP 2020 - Demos, Online, November 16-20, 2020*, pp. 38–45, 2020. URL <https://doi.org/10.18653/v1/2020.emnlp-demos.6>.
- Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 5811–5820. Association for Computational Linguistics, 2020. URL <https://doi.org/10.18653/v1/2020.acl-main.515>.
- Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob

Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 3391–3401, 2017.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 2031–2043. Association for Computational Linguistics, 2020a. URL <https://doi.org/10.18653/v1/2020.acl-main.184>.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*, 2020b.

Hao Zhou, Minlie Huang, Yong Liu, Wei Chen, and Xiaoyan Zhu. EARL: informative knowledge-grounded conversation generation with entity-agnostic representation learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 2383–2395. Association for Computational Linguistics, 2021.

## Appendix

### A Limitations

As discussed in [Appendix H](#), our work is limited in a variety of dimensions primarily in terms of the lack of a well-formulated public dataset, retrieval accuracy, and generation quality. First, the public benchmark dataset for knowledge-consistent dialogue generation is highly limited. Despite the fact that there are several public Knowledge Graphs (KGs) available ([Vrandecic & Krötzsch, 2014](#); [Bollacker et al., 2008](#)), only one dataset ([Moon et al., 2019](#)) provides the diverse set of dialogue and its corresponding large-scale KG. This circumstance may limit the rigorous evaluation of our framework’s adaptability in various settings. Future work may study applying our approach for a wider range of dialogue datasets based on Wikipedia ([Dinan et al., 2019](#)) by leveraging existing public large-scale KG such as Wikidata ([Vrandecic & Krötzsch, 2014](#)). Second, the search space for retrieving context-relevant subgraphs can be expanded. Our SURGE framework now runs on a  $k$ -hop KG that is rooted in entities in the given dialogue history. Finding the entity within the text, on the other hand, necessitates precise named entity extraction and entity linking. Therefore, future work may investigate extending our approach to a framework that can retrieve the context-relevant subgraph among the entire KG instead of  $k$ -hop KG. Third, there is still room for improvement in generation quality since we generate knowledge-enhanced responses with a small-scale Pre-trained Language Model (PLM) for efficiency. Such PLMs occasionally fail to generate high-quality natural sentences ([Raffel et al., 2020](#)). Future work could improve generation quality based on a larger PLM.

### B Broader Impacts

Our proposed knowledge-grounded dialogue generation model is essential for designing user-friendly real-world AI systems. Among various types of dialogue generation models, knowledge-grounded dialogue models are trained to interact with users and convey factual information to users in natural languages. Their conversational features can be adapted to any user interface that connects the bilateral interaction between humans and computers. We believe that conversational interfaces can enhance the users’ experiences and reduce the users’ efforts in learning how to use the systems. However, knowledge-grounded dialogue models can become vulnerable to generating offensive and harmful content or responses with misinformation depending on the users or data. When deploying the models in the real world, in addition to generating realistic responses, they also need to be robust to adversarial feedback from malicious users and biases inherited in pre-training or training corpus, or else they could malfunction. Therefore, along with the quantitative and qualitative evaluations of generated responses, it would be worthwhile to examine the robustness of the dialogue models.

### C Notations

We organize the notations we used for formally describing our method in [Table 8](#).

### D Intuitions & Proofs for Graph Encoding

#### D.1 Intuitions

In [Section 3.4](#), we focus to introduce the novel graph encoding which meets both permutation and relation inversion invariances. However, one may draw the question of why such invariances are important for graph encoding with the pre-trained language models (PLMs) and need more detailed explanations on this.

First of all, we want to recapitulate why the permutation invariance is important in encoding multi-relational graphs into the PLMs along with the text sequence. As noted in [Section 3.1](#), PLMs are permutation sensitive since the meaning of the sentence can vary when we change the order of words in the sentence (e.g., "A is born in C"  $\neq$  "C is born in A"). However, the multi-relational graphs are permutation invariant since they are represented as a set of triplets. For instance, given the multi-relational graphs with two triplets,  $\{(a, \text{born-in}, c), (b, \text{born-in}, d)\}$ , the order of elements (triplets) does not affect the entire semantic of the graph. (e.g.,  $\{(a, \text{born-in}, c), (b, \text{born-in}, d)\} = \{(b, \text{born-in}, d), (a, \text{born-in}, c)\}$ ).

Table 8: A list of notations that we used for defining our method.

---

$\mathcal{V}$	pre-defined vocabulary of tokens for pre-trained language models (text)
$\mathcal{E}$	pre-defined vocabulary of entities (symbol)
$\mathcal{R}$	pre-defined vocabulary of relations (symbol)
$\mathbf{a}, \dots, \mathbf{z}$	knowledge graph symbols written in typewrite font
$\mathbf{x}$	input sequence (vector)
$x_1, \dots, x_N$	input tokens (scalar)
$\mathbf{y} = [y_1, \dots, y_T]$	output sequence and tokens
$\mathcal{G}$	multi-relational graph, such as knowledge graph
$\mathcal{Z}$	retrieved subgraph: $\mathcal{Z} \subset \mathcal{G}$
$z$	triplet (edge): $z \in \mathcal{Z}$
$q$	tokenization (mapping) function of KG symbol to the sequence of tokens
$s(\cdot)$	text representation function for retrieval
$d(\cdot)$	triplet representation function for retrieval
<b>Enc</b>	Transformer Encoder
<b>Dec</b>	Transformer Decoder
$f$	token (word) embedding function
$\theta$	generator parameter
$\phi$	retriever parameter
$\psi$	set encoding function
$\beta$	perturbation function
$\pi$	set permutation
$n$	the number of triplets in a retrieved subgraph $\mathcal{Z}$
$k$	the number of samples in a marginalization term
$z$	encoder hidden state (single token)
$\mathbf{Z}$	encoder hidden states (sequence of tokens)
$h$	decoder hidden state (single token)
$\mathbf{H}$	decoder hidden states (sequence of tokens)
$\mathbf{X}$	input embeddings after token embedding function (sequence)
$\mathbf{Y}$	output embeddings after token embedding function (sequence)

---

With a naïve encoding, the PLM yields different representations for different orders of triplets in the subgraph. Therefore, if the PLM is only fine-tuned with the input [A, born-in, C, B, born-in, D, where was A born?], there is no guarantee that the PLM will output the exact same response given the input with a permuted subgraph [B, born-in, D, A, born-in, C, where was A born?] in the inference since the PLM is order-sensitive due to its positional encoding. In order to prevent the aforementioned scenarios, we decide to design the permutation-invariant graph encoding which yields stable results regardless of the order of triplets in the graph. Similarly, the inversion of the triplet yields the same semantic (e.g.,  $\{(a, \text{born-in}, c)\} = \{(c, \neg\text{born-in}, a)\}$ ), but the graph encoding without considerations for the inverse relation results in different representations from PLM given the triplet and its inversed one.

## D.2 Proofs

In this section, we first show that a naïve encoding function  $\psi$  in Section 3.4 is neither permutation invariant nor relation inversion invariant, formalized in Proposition D.1. After that, we prove that our invariant and efficient encoding function  $\psi^*$  with graph-conditioned token embedding perturbation is both permutation invariant and relation inversion invariant, formalized in Proposition D.2.

**Proposition D.1.** *A naïve encoding function  $\psi$  is neither permutation invariant nor relation inversion invariant.*

*Proof.* We prove this by contradiction.

Suppose  $\mathbf{x} = [x_1, \dots, x_n]$  and  $\mathcal{Z} = \{(a, r_1, b), (b, r_2, a), (a, r_1, c)\}$ . Moreover, let  $\mathcal{Z}' = \{(b, r_2, a), (a, r_1, b), (a, r_1, c)\}$  be one of permutations of  $\mathcal{Z}$  with the permutation order  $\pi = (2, 1, 3)$ .

With a naïve encoding,  $\psi(\mathbf{x}, \mathcal{Z}) = [a, r_1, b, b, r_2, a, a, r_2, c, x_1, \dots, x_n]$  and  $\psi(\mathbf{x}, \mathcal{Z}') = [b, r_2, a, a, r_1, b, a, r_1, c, x_1, \dots, x_n]$ . Therefore, it is easy to notice that  $\psi(\mathbf{x}, \mathcal{Z}) \neq \psi(\mathbf{x}, \mathcal{Z}')$ , thus the naïve encoding is not permutation invariant.

We then show that a naïve encoding is not relation inversion invariant. Suppose  $\mathcal{Z}'' = \{(a, r_1, b), (b, r_2, a), (c, \neg r_1, a)\}$ , where  $(a, r_1, c) \in \mathcal{Z}$  is changed to its inverse relation  $(c, \neg r_1, a)$ . Then,  $\psi(\mathbf{x}, \mathcal{Z}'') = [a, d, b, b, e, a, c, \neg d, a, x_1, \dots, x_n]$  that is different against  $\psi(\mathbf{x}, \mathcal{Z})$ :  $\psi(\mathbf{x}, \mathcal{Z}) \neq \psi(\mathbf{x}, \mathcal{Z}'')$ . Therefore, the naïve encoding function is not relation inversion invariant.

In conclusion, from the above two counterexamples, we prove that a naïve encoding function  $\psi$  is neither permutation invariant nor relation inversion invariant.  $\square$

We now provide proof of the permutation invariance and the relation inversion invariance of our invariant and effective graph encoding  $\psi^*$ , described in Section 3.4. Before starting the proof, we first revisit the permutation invariant property of graph neural networks that sum, mean and max operators are permutation invariant for the input set of AGGR. Thus, if we use sum, mean, or max for AGGR, then the token embedding perturbation function  $\beta$  naturally satisfies the permutation invariance property. In other words,  $\beta(\mathbf{X}, \mathcal{Z}) = \beta(\mathbf{X}, \pi \cdot \mathcal{Z})$ , where  $\mathbf{X} = \tilde{\psi}(\mathbf{x}, \text{SORT}(\text{ENT}(\mathcal{Z})))$  for any permutation  $\pi$ .

**Proposition D.2.** *Invariant and efficient encoding  $\psi^*$  is both permutation invariant and relation inversion invariant.*

*Proof.* Suppose  $\mathbf{x} = [x_1, \dots, x_n]$  and  $\mathcal{Z} = \{(a, r_1, b), (b, r_2, a), (a, r_1, c)\}$ . We first consider the permutation invariance for any permuted set  $\mathcal{Z}' = \pi \cdot \mathcal{Z}$ . While  $\mathcal{Z}$  and  $\mathcal{Z}'$  can have different orders of elements thus the outputs of  $\text{ENT}(\mathcal{Z})$  and  $\text{ENT}(\mathcal{Z}')$  could be different, we always obtain the same output with the usage of the SORT operator for encoding. In other words,  $\text{SORT}(\text{ENT}(\mathcal{Z})) = \text{SORT}(\text{ENT}(\mathcal{Z}'))$  holds due to the definition of the SORT operation in Eq. 5 of the main paper. Therefore,  $\tilde{\psi}(\mathbf{x}, \text{SORT}(\text{ENT}(\mathcal{Z}))) = \tilde{\psi}(\mathbf{x}, \text{SORT}(\text{ENT}(\mathcal{Z}')))$  holds.

Further, since the token embedding perturbation function  $\beta(\cdot, \mathcal{Z})$  along with sum, max, or mean in AGGR is also permutation invariant with regards to any permutation on  $\mathcal{Z}$ , we conclude our invariant and efficient encoding  $\psi^*$  is permutation invariant.

We finally prove the relation inversion invariance property of  $\psi^*$ . Suppose  $\mathcal{Z}'' = (\mathcal{Z} \cup t') \setminus t$  where  $t \in \mathcal{Z}$  is any triplet in a set and  $t'$  is inverse of  $t$ . Then,  $\text{ENT}(\mathcal{Z}) = \text{ENT}(\mathcal{Z}'')$  that is trivial as  $\text{ENT}(\mathcal{Z})$  returns the set of only unique nodes in  $\mathcal{Z}$ . Therefore,  $\tilde{\psi}(\mathbf{x}, \text{SORT}(\text{ENT}(\mathcal{Z}))) = \tilde{\psi}(\mathbf{x}, \text{SORT}(\text{ENT}(\mathcal{Z}''))) \text{ correspondingly holds.}$

The remaining step to conclude the proof is to show the following equality:  $\beta(\cdot, \text{INV}(\mathcal{Z})) = \beta(\cdot, \text{INV}(\mathcal{Z}''))$ , to conclude that  $\psi^*(\mathbf{x}, \mathcal{Z}) = \psi^*(\mathbf{x}, \mathcal{Z}'')$  from  $\beta(\tilde{\psi}(\mathbf{x}, \text{SORT}(\text{ENT}(\mathcal{Z}))), \text{INV}(\mathcal{Z})) = \beta(\tilde{\psi}(\mathbf{x}, \text{SORT}(\text{ENT}(\mathcal{Z}''))), \text{INV}(\mathcal{Z}''))$ . We note that  $\text{INV}(\mathcal{Z}) = \text{INV}(\mathcal{Z}'')$ , as  $\text{INV}$  makes any graph as bidirectional one by the definition in Eq. 6 of the main paper. Therefore,  $\beta(\cdot, \text{INV}(\mathcal{Z})) = \beta(\cdot, \text{INV}(\mathcal{Z}''))$  holds, and the relation inversion invariance property of  $\psi^*$  holds. □

## E Experimental Setup

In this section, we introduce the detailed experimental setups for our models and baselines. Specifically, we describe the details on implementation, dataset, training and model in the following subsections of [E.1](#), [E.2](#), [E.3](#) and [E.4](#), one by one.

### E.1 Implementation Details

We use the T5-small ([Raffel et al., 2020](#)) as the base Pre-trained Language Model (PLM) for all experiments. For the pre-trained checkpoint, we use the version that the authors released. For all implementations, we use Pytorch ([Paszke et al., 2019](#)). To easily implement the language model, we use the huggingface transformers library ([Wolf et al., 2020](#)).

**Retriever Details** In this paragraph, we describe the implementation details of our context-relevant subgraph retriever, including the triplet embedding and dialogue context embedding for the retriever.

For the dialogue history embedding function  $q$ , we use the existing pre-trained language model (PLM). Specifically, we use the encoder part of the T5-small model ([Raffel et al., 2020](#)) and freeze the parameters of it not to be trained. We then instead add a Multi-Layer Perceptron (MLP) on top of it, to give a point-wise attention ([Bahdanau et al., 2015](#)) to each token, whereby all tokens are not equally considered in the sentence encoding. Formally,

$$q(\mathbf{x}) = \sum_{i=1}^n \alpha_i * \mathbf{z}_i, \quad \mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n] = \text{Enc}(\mathbf{X}),$$

$$\alpha_i = \frac{\exp(\text{MLP}(\mathbf{z}_i))}{\sum_{j=1}^n \exp(\text{MLP}(\mathbf{z}_j))} \quad \forall i$$

where  $\alpha_i$  is a scalar, and MLP is a Multi-Layer Perceptron consisting of two linear layers and ReLU nonlinearity.

For obtaining triplet representations, we need to embed the entity (node) and relation (edge) into the latent space. Similar to the token embedding matrix used in PLMs, we can introduce the entity and relation embedding matrices. However, since the number of entities used in Freebase of OpendialKG ([Moon et al., 2019](#)) is too large compared to the number of tokens in T5 (100,814 vs 32,000) ([Raffel et al., 2020](#)), it is inefficient to introduce the trainable entity embedding matrix for the retriever. Furthermore, the use of standalone entity embedding matrix might be sub-optimal in terms of generalization since there is no evidence that all entities in a large-scale KG emerge in training dataset.

Thus, we instead reuse the contextualized representation from the PLM encoder, to embed each node if the corresponding entity exists in the dialogue context. Formally, suppose that there is a triplet  $\{(\mathbf{e}_h, \mathbf{r}, \mathbf{e}_t)\}$  in the 1-hop subgraph  $\mathcal{G}$ , which satisfies the following condition:  $q(\mathbf{e}_h) \subseteq \mathbf{x}$  or  $q(\mathbf{e}_t) \subseteq \mathbf{x}$ . If so, we can know the position of the mapped entity within the dialogue history:  $[x_{start}, \dots, x_{end}] = q(\mathbf{e}_h)$  from  $q(\mathbf{e}_h) \subseteq \mathbf{x}$ . Therefore, the node embedding for the entity  $\mathbf{e}_h$  is obtained by  $\text{EntEmb}(\mathbf{e}_h) = \frac{1}{|q(\mathbf{e}_h)|} \sum_{i=start}^{end} \text{Enc}(\mathbf{X})_i$  iff  $q(\mathbf{e}_h) \subseteq \mathbf{x}$ . If the entity mention does not exist in the dialogue history, we use the zero vector as the node embedding. For

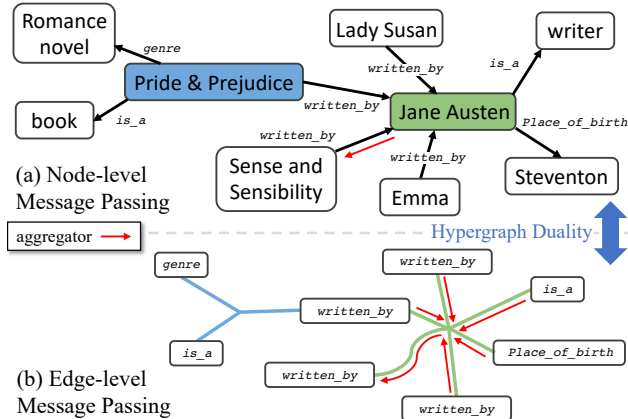


Figure 6: **GNN-based Triplet Representation for Retrieval.** To represent each triplet with regards to its graph structure, we use the message passing on both nodes and edges. (a) Node-level Message Passing. To represent the entity *Sense and Sensibility*, the message from its neighbors – the entity *Jane Austen* – is aggregated. (b) Edge-level Message Passing. To represent the relation `written_by`, the messages from relations associated to a green hyperedge are aggregated. We do not draw self-loops and inverse edges for simplicity.

edge embedding, we use the trainable relation embedding matrix  $\mathbf{R} \in \mathbb{R}^{|\mathcal{R}| \times 128}$  to represent the edge, since the number of relations is relatively small (1,357).

With our node and edge representations, we now focus on representing the triplet in Eq. 4 of the main paper for its retrieval. In particular, we use the Graph Neural Networks (GNNs) for encoding triplets, where we obtain the node representations from the Graph Convolutional Network (GCN) (Kipf & Welling, 2017) that is a widely used architecture for representing the nodes with respect to their graph structures. However, for representing the edges, we use the Edge Hypergraph Graph Neural Network (EHGNN) used in Jo et al. (2021), due to its simplicity but effectiveness for edge representations. We summarize our triplet representation in Figure 6.

**Graph Encoding Details** In this paragraph, we describe the implementation details of the token embedding perturbation function  $\beta$  used in our *Invariant and Efficient* graph encoding introduced in Section 3.4. To be aware of the relation of the graph over GNNs, we use the simplified version of CompGCN (Vashishth et al., 2020). For architectural details, instead of using the different linear layers to distinguish the inverse relation from its opposite relation, we use the same linear layer. Also, we use subtraction as the specific composition operator for reflecting relations in CompGCN.

Then, we form the learnable affine transformation based on the aggregated representation from GNN layers, to perturb the token embeddings with respect to their graph information as in Equation 6 of the main paper. In particular,

$$\begin{aligned} \boldsymbol{\eta} &= \mathbf{R}\text{-GNN}(f(a); \mathcal{Z}) = \text{UPD}(f(a), \text{AGGR}(\{f(b), \mathbf{r} \mid \forall \mathbf{b} \in \mathcal{N}(\mathbf{a}; \mathcal{Z})\})), \\ \boldsymbol{\gamma} &= \text{MLP}_1(\boldsymbol{\eta}), \quad \boldsymbol{\delta} = \text{MLP}_2(\boldsymbol{\eta}), \quad \beta(f(a), \mathcal{Z}) = (\mathbf{1} + \boldsymbol{\gamma}) * f(a) + \boldsymbol{\delta}, \end{aligned}$$

where  $\text{MLP}_1$  and  $\text{MLP}_2$  are learnable MLPs consisting of two linear layers with ReLU nonlinearity. In Figure 7, we illustrate comprehensive diagram of Equation 6, which enables our *Invariant and Efficient* graph encoding to understand the structure of the retrieved subgraph  $\mathcal{Z}$ .

**Contrastive Learning Details** For contrastive learning, we initialize  $\tau$  in Equation 7 as 0.01.

**KQA Details** In this paragraph, we describe the implementation details for our Knowledge-verifying Question Answering (KQA) introduced in Section 4. For building the QA dataset, we first gather the dialogue sessions where the gold response contains the entity from the whole OpendialKG dataset. Then, we extract the triplet from the given whole KG where the head entity is placed within the dialogue history and the tail entity is placed within the gold response. We build a QA training dataset based on the extracted triplets and a corresponding dialogue session. To diversify the training data, we replace the tail entity of each triplet with

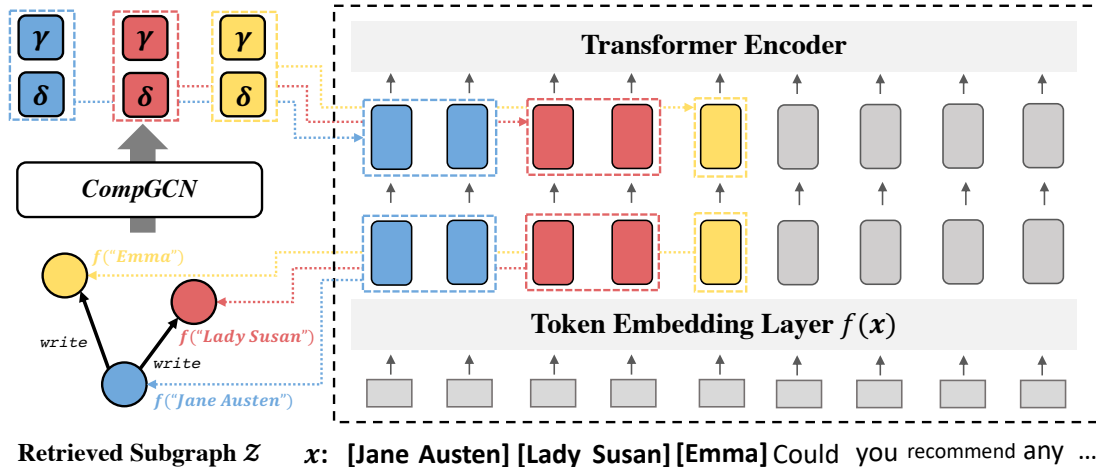


Figure 7: **Comprehensive diagram for Invariant and Efficient graph encoding.** Our proposed graph encoding first concatenates the sorted list of entities in front of the dialogue history. Then, we form the learnable affine transformation  $\gamma, \delta$  for each entity using relation-aware GNN such as CompGCN (Vashishth et al., 2020).

plausible candidate entities within KG and change the entity in the response following the changed entity on the triplet. As a result, we obtain the QA dataset size of 200k. We train the BERT-base (Devlin et al., 2019) with the constructed QA dataset. We hold out 10% of data for validation and obtain the fine-tuned BERT model with 88.89 F1 score on the hold-out validation set. When we apply the fine-tuned QA model on the evaluation of the generated responses, we rebuild the QA evaluation set with the generated response instead of a gold response as illustrated in Figure 3 of the main paper.

## E.2 Dataset Details

We mainly conduct experiments on **OpendialKG** (Moon et al., 2019), which provides the parallel dialogue corpus corresponding to the existing large-scale Knowledge Graph (KG) named Freebase (Bollacker et al., 2008). The provided large-scale KG consists of total 1,190,658 fact triplets over 100,813 entities and 1,358 relations. This dataset is collected from 15K human-to-human role-playing dialogues, having multi-turns, from which we pre-process that each assistance response is the label and its corresponding dialogue history is the input. Although some of the data contain the gold knowledge that is useful for generating the response on the ongoing conversation, we found that 51% of data has no gold knowledge. To overcome this limitation, we additionally find entities from the dialogue history using the Named Entity Recognition module in spaCy<sup>1</sup>, and then include the extracted entities’ corresponding triplets in the KG to the dataset. For entity linking, we use the exact match. Since the dataset does not provide the pre-defined data split, we randomly split sessions into train (70%), validation (15%), and test sets (15%). We also conduct experiments on **KOMOIDS** (Galetzka et al., 2020) dataset and follows the same preprocessing as in OpendialKG dataset.

## E.3 Training Details

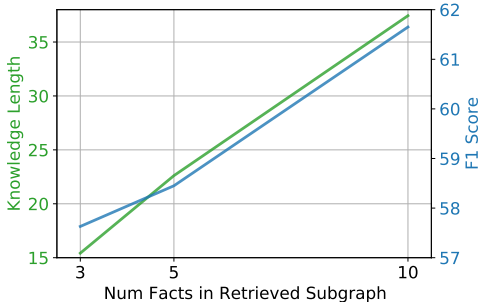
All experiments are constrained to be done with a single 48GB Quadro 8000 GPU. SURGE training needs 12 GPU hours. For all experiments, we select the best checkpoint on the validation set. We fine-tune the SURGE for 30 epochs on the training set, where we set the learning rate as 1e-4, weight decay as 0.01, learning rate decay warmup rate as 0.06, maximum sequence length for dialogue history as 256, maximum sequence length for knowledge as 128, and batch size as 24. For retrieval, we use the subgraph size  $n$  as 3, and sample size  $k$  for marginalization as 4. We use the AdamW (Loshchilov & Hutter, 2019) optimizer for training. For fair evaluation, we apply the same training setting to all baselines if applicable. All experimental results are reported with a single run.

<sup>1</sup><https://spacy.io/>

## E.4 Model & Baselines Details

In this subsection, we describe the details of baselines and our models used in our experiments, as follows:

1. **No Knowledge**: This model is provided with only the dialog history. No knowledge is used to generate responses.
2. **Gold Knowledge**: This model is provided with the dialogue history along with its exact gold knowledge for the gold response. Thus, since this model uses such gold knowledge, we expect the results of it as the upper bound of the task.
3. **Space Efficient (series)**: This model is provided with all the knowledge which are related to the entities that appeared in the dialogue history (Galetzka et al., 2021), by matching the entities in the dialogue history and the entities in the KG. In particular, this model encodes the entities and their relations explicitly in the words in the encoder part.
4. **Space Efficient (parallel)**: This model is mostly the same as the above model – space Efficient (series) – except the knowledge encoding part. Specifically, it encodes the entities in the words like the above, whereas, encoding the relation between entities in the segmentation block of the entities Galetzka et al. (2021).
5. **EARL**: This model uses the RNN-based encoder-decoder architecture with the entity-agnostic representation learning (Zhou et al., 2021), with all the provided knowledge associated with the entities in the dialogue history. Specifically, this model first calculates the probability of words obtained by encoding the entities in the KG, and then uses such probabilities to generate a word in the decoding phase.
6. **DiffKG**: This model (Tuan et al., 2022) uses a differentiable path reasoning, which is jointly trainable along with the dialogue generation. After the path reasoning, the entities in the reasoning path are naively appended in front of the dialogue history, then concatenated input is forwarded to the pre-trained language model.
7. **Random Retrieval**: This model is provided with entire facts from k-hop subgraphs of entities that appeared in the dialogue history. However, instead of encoding all the knowledge in one-hop subgraph as in Space Efficient, this model randomly samples them, which are then used for generating responses.
8. **Sparse Retrieval (BM25)**: This model is also provided with entire facts from k-hop subgraphs of entities. To sample relevant facts to the dialogue history among the entire facts, this model uses BM25 (Robertson & Zaragoza, 2009) that is a sparse retrieval model. To be specific, let assume we have a dialogue history and its corresponding facts from k-hop subgraphs of matched entities. Then, to run BM25, we first concatenate components of each fact consisting of two entities and one relation, and tokenize the dialogue history and the facts for obtaining corpus and queries, respectively, for BM25. After that, BM25 calculates the lexical overlapping score between the dialogue context (corpus) and the one-hop fact (query), from which we use the relevant facts having top- $k$  scores by BM25.
9. **Dense Retrieval (Bi-encoder, Poly-encoder)**: This model uses a pre-trained language model for the triplet embedding of the retriever instead of using GNN. Specifically, we consider each triplet as a single sentence (e.g, (Jane Austen, write, Susan)  $\rightarrow$  “Jane Austen write Susan”) and embed them with the pre-trained language model. For scoring, we use both bi-encoder and poly-encoder architectures (Humeau et al., 2020).
10. **SURGE (unsupervised)**: Our basic subgraph retrieval-augmented generation framework that is provided with entire facts from k-hop subgraphs of entities. In particular, this model trains the structure-aware subgraph retriever without any guidance of the gold knowledge (i.e., ground truth knowledge for the dialogue history is not given). In other words, for the given dialogue context, this model implicitly learns to retrieve the context-relevant knowledge, and then generates the response with the retrieved knowledge.
11. **SURGE (semi-supervised)**: Our subgraph retrieval-augmented generation framework with semi-supervised learning of graph retrieval, with provided entire facts from k-hop subgraphs of entities. Unlike the unsupervised version of SURGE, this model trains the retriever to select the gold knowledge if the dialogue context has such knowledge during training.
12. **SURGE (contrastive)**: Our full subgraph retrieval-augmented generation framework with the contrastive learning of graph-text modalities as well as the semi-supervised learning of graph retrieval, with provided entire facts from k-hop subgraphs of entities. Unlike aforementioned frameworks of ours, this additionally



Method	MRR	Hits@1	Hits@3	Hits@5	Hits@10	Hits@100
Random Retrieval	8.67	3.11	7.89	10.86	17.84	66.22
Sparse Retrieval (BM25)	7.76	2.61	6.72	9.89	16.67	61.21
Dense Retrieval (Bi-encoder)	20.89	11.99	23.79	30.34	39.26	67.26
Dense Retrieval (Poly-encoder)	21.47	12.13	24.12	31.51	41.43	68.03
SURGE (unsupervised)	21.24	10.46	24.06	31.97	44.17	74.74
SURGE (semi-supervised)	22.53	13.40	24.79	31.84	42.37	69.33
SURGE (contrastive)	<b>25.98</b>	<b>16.67</b>	<b>28.50</b>	<b>35.72</b>	<b>46.11</b>	<b>74.31</b>

Figure 8: (Left:) Performances of our SURGE by varying the number of facts for retrieving the subgraph (i.e., varying the number of triplets in the subgraph) from three, to five, to ten, with the length of sequence for knowledge (knowledge length) and F1 scores of KQA as evaluation metrics. (Right:) We additionally report the knowledge retrieval performances, with MRR and Hits@K as evaluation metrics.

Table 9: Experimental results on OpendialKG dataset with **BART-base**.

Method	KQA		BLEU				ROUGE			Unigram
	EM	F1	B-1	B-2	B-3	B-4	R-1	R-2	R-L	F1
No Knowledge ( <i>BART-base</i> )	31.17	37.54	17.38	10.79	7.16	4.81	20.64	8.22	19.92	24.36
Space Efficient ( <i>BART-base, Series</i> )	48.49	53.83	18.56	11.85	8.01	5.56	22.36	9.43	21.48	26.38
Space Efficient ( <i>BART-base, Parallel</i> )	49.80	55.06	<b>18.90</b>	<b>12.19</b>	<b>8.35</b>	<b>5.81</b>	<b>22.63</b>	<b>9.79</b>	<b>21.76</b>	<b>26.79</b>
SURGE ( <i>BART-base, semi-supervised, n = 10</i> )	50.84	57.35	17.80	11.12	7.48	5.18	18.64	7.27	17.77	22.07
SURGE ( <i>T5-small, semi-supervised, n = 3</i> )	51.32	58.45	17.63	11.28	7.41	5.39	21.74	9.18	20.85	25.57
SURGE ( <i>T5-small, semi-supervised, n = 10</i> )	<b>54.50</b>	<b>61.65</b>	17.70	11.37	7.81	5.50	21.55	9.09	20.65	25.44

enforces the model to faithfully reflect the retrieved knowledge in the input, to the generated response with contrastive learning.

## F Additional Experiments

### F.1 Varying the Number of Facts in Subgraphs

We experiment our SURGE framework with varying the number of facts in retrieval, which are then used in our graph encoding function to condition the encoded graph information for response generation. Specifically, in Figure 8, we report the length of sequence for knowledge (knowledge length) and F1 scores measured by our KQA for our SURGE framework, with different numbers of facts within a retrieved subgraph:  $n = [3, 5, 10]$ . Note that, in this experiment, we only use the semi-supervised model without the contrastive loss. We expect that the performance of our SURGE will increase as we increase the number of facts within the retrieved subgraph, since the model can leverage more numbers of knowledge for response generation. As shown in Figure 8, we observe the significant performance improvements on using ten facts against using three and five facts, while the performance difference between the three and five is marginal. We suggest that this result should be interpreted with the retrieval results on the right side of Figure 8, where about 40% of retrieved subgraphs including the ten different facts contain at least one necessary knowledge, thus the generation performance is boosted according to the improvement in retrieval.

### F.2 Discussions on Using Larger PLMs

Notably, we observe that the use of larger Pre-trained Language Models (PLMs) – three times more number of parameters compared to T5-small that we use – does not result in better performance for the knowledge-grounded dialogue task. Specifically, in Table 9, we report the experimental results of selected baselines and our SURGE semi-supervised model with BART-base (Lewis et al., 2020a) as the base PLM. We want to clarify that the BART-base model has 220M parameters, which is about **three times larger** than the number of parameters of the T5-small model (60M).

Table 10: Experimental results on KOMODIS dataset with T5-small.

Method	KQA		BLEU				ROUGE			Unigram
	EM	F1	B-1	B-2	B-3	B-4	R-1	R-2	R-L	F1
<b>Random</b>	12.41	14.17	7.74	4.02	2.46	1.68	21.79	4.00	21.44	16.29
<b>Space Efficient</b> ( <i>Series</i> )	12.41	14.70	8.34	5.13	3.77	3.05	22.36	4.85	22.06	17.37
<b>Space Efficient</b> ( <i>Parallel</i> )	16.46	18.70	9.33	5.66	4.06	3.20	22.80	4.12	22.47	17.72
<b>SURGE</b> ( <i>unsupervised</i> )	16.18	18.51	11.46	7.10	5.15	4.07	23.49	5.77	23.09	18.70
<b>SURGE</b> ( <i>semi-supervised</i> )	16.62	19.48	11.28	6.98	5.05	3.98	23.58	5.79	23.21	18.68
<b>SURGE</b> ( <i>contrastive</i> )	<b>17.30</b>	<b>19.50</b>	<b>11.51</b>	<b>7.18</b>	<b>5.20</b>	<b>4.10</b>	<b>24.13</b>	<b>6.17</b>	<b>23.74</b>	<b>19.51</b>

We first observe that BART-base shows decent performance without any knowledge (No Knowledge) compared to the no-knowledge case of T5-small, verifying that the larger PLM generally contains more factual knowledge within its pre-trained parameters. Moreover, BART-base obtains higher scores in the simple word overlap metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), whose results further confirm that a larger PLM can generate more natural or syntactically better sentences than the smaller one, thanks to its parameter size.

On the other hand, we find that BART-base is less suffered from the irrelevant knowledge issue (i.e., conditioning irrelevant knowledge for the given context when generating responses) than T5-small, therefore, the performance of *Space Efficient Encoding* on KQA is quite high. However, the use of BART-base does not result in significant improvement on the KQA metric for our SURGE framework. Moreover, ours with T5-small shows better performance than ours with BART-base in terms of KQA scores, when the number of facts within the retrieved subgraph is 10:  $n = 10$ . This result suggests that the quality of the generated response – having relevant knowledge to the given context – might depend on the performance of the subgraph retriever whose goal is to retrieve the context-relevant knowledge, rather than the inherent performance of PLMs.

### F.3 Full Experimental Results on KOMODIS

In the main paper, we mostly focus on OpendialKG dataset (Moon et al., 2019), since it is the largest and most realistic public datasets that provides both dialogues across diverse domains and corresponding large-scale Knowledge Graph (KG) (Bollacker et al., 2008). To verify the effectiveness of our SURGE framework, the existence of the large-scale KG and the importance of relevant fact searching is important since we focus on the real-world scenario where the response generation requires the relevant fact acquirement from the large-scale KG.

However, one can raise the question regarding the versatility of our method on other datasets. To alleviate the issue, we conduct additional experiments on another dataset named KOMODIS (Galetzka et al., 2020), which is also KG-based dialogue dataset. Compared to OpendialKG, KOMODIS does not provide the corresponding large-scale KG and most of responses do not require the knowledge. Therefore, we only measure the automatic evaluation to evaluate the performance of each method on KOMODIS dataset. In Table 10, we present the experimental results on the KOMODIS dataset. Results obviously show that our SURGE framework shows superior performance against baselines on the additional dataset. Therefore, we can conclude that our method can generalize to other datasets beyond the opendialKG dataset.

### F.4 Diversity Evaluation

In the main paper, we evaluate model generation performance primarily on its quality. We measure the distinct metric (Li et al., 2016), which is one of the most popular metrics for evaluating the diversity of the generative model, to evaluate the performance of each model in more diverse aspects. In Table 11 left, we report the performance of baselines and our models in distinct metric. Our SURGE framework generates more diverse responses than all other baselines, according to the results.

Table 11: (Left:) Performance evaluation with the diversity metric named Distinct. (Right:) Ablation study results on GNN variants in our modules.

Method	Dist-1	Dist-2										
No Knowledge	6.06	15.73										
All Knowledge	9.67	24.45										
SEE (Series)	8.49	21.77										
SEE (Parallel)	8.78	22.70										
EARL	5.15	16.46										
Sparse Retrieval (BM25)	7.65	19.63										
SURGE (semi-supervised)	10.18	27.85										

Method	KQA		BLEU				ROUGE			Unigram
	EM	F1	B-1	B-2	B-3	B-4	R-1	R-2	R-L	F1
Eq 4. GCN $\rightarrow$ GAT	49.16	56.10	17.42	10.96	7.39	5.17	21.10	8.65	20.25	24.79
Eq 7. CompGCN $\rightarrow$ GCN	48.61	55.53	17.48	10.97	7.34	5.05	21.23	8.73	20.37	24.77
<b>SURGE (semi-supervised)</b>	<b>51.00</b>	<b>57.63</b>	<b>17.70</b>	<b>11.21</b>	<b>7.61</b>	<b>5.28</b>	<b>21.43</b>	<b>8.85</b>	<b>20.57</b>	<b>25.07</b>

## F.5 Ablations Studies on GNN Design Choices

We use two different types of Graph Neural Networks (GNN) in our SURGE framework. One is the Graph Convolutional Network (GCN) (Kipf & Welling, 2017), which is used to embed each node entity on the entire 1-hop subgraph in the triplet embedding function  $d$  of the main paper Equation 4. Another is Composition-Based Multi-Relational Graph Convolutional Networks (CompGCN) (Vashishth et al., 2020), which is used to embed each entity by considering the relations between entities in the token embedding perturbation function  $\beta$  of the main paper Equation 6. In this subsection, we conduct ablation studies on both GNN design choices. First of all, we replace the GCN in Equation 4 with Graph Attention Network (GAT) (Velickovic et al., 2018) to validate the effect of the GNN design choices on the node embedding in the triplet embedding function. Then, we run experiments by changing CompGCN in Equation 6 to GCN to see how important the relationships are in the graph encoding. We present the results on Table 11 right. Results indicate that the use of GAT in Equation 3 does not have any impact on the performance a lot. However, the use of relation-aware GNN is highly important in effective and efficient graph encoding, since removing the relation awareness of GNN reduces the performance of our model a lot.

## F.6 Automatic Evaluations on Knowledge Groundedness

Our KQA metric introduced in Section 4 is useful to evaluate the knowledge groundedness of the generated response. Since the novelty of the KQA metric stems from the use of the KG to resolve the issue from the missing knowledge by only considering the gold response for evaluation, we can also utilize other rule-based metrics like **string matching** (check whether at least one of the answer candidates of KQA presents in the generated response by string matching) or **Entity F1** score (measuring the F1 score against each entity in answer candidates instead of the gold response). As we all know, automatic evaluations can be imperfect when compared to human evaluations. However, we also believe that using a variety of credible automatic evaluation metrics will strengthen the validity of the experimental results. Therefore, we supplement the experimental results with three more evaluation metrics for measuring whether the generated responses contain appropriate knowledge.

In Table 12, we measure Knowledge F1 (KF1 in Table 3 in the main paper), string matching, and entity F1 for representative baselines and our SURGE (semi-supervised) in OpendialKG, as an extension of Table 1 in the main paper. For KF1, we measure the F1 score regarding the concatenation of the question (head entity and relation) and all answer candidates (available tail entities) in KQA as the gold response. The results show that all metrics show the same tendency with KQA and our proposed method still outperforms other baselines by generating responses with more proper knowledge. Although three rule-based metrics are useful for assessing the knowledge groundedness of generated responses, they do have some drawbacks. KF1 and Entity F1 are affected by the length of the generated response and answer candidates. String matching is too strict since it may miss some responses that only contain partial words of knowledge (e.g., the response only contains the first name of the author whereas the candidate answers contain the full name of the author). As a result, the use of KQA is also beneficial since the trained QA model can compensate for the shortcomings of rule-based metrics.

Table 12: Experimental results on OpendialKG with additional three metrics other than KQA for measuring whether the generated responses contain appropriate knowledge.

Method	KQA		Knowledge F1	Entity F1	String Matching
	EM	F1			
No Knowledge	12.25	20.69	13.80	9.33	13.03
Random Knowledge	31.72	38.95	16.29	16.49	32.71
All Knowledge	43.58	50.60	18.91	21.10	44.25
Space Efficient ( <i>Parallel</i> )	38.54	44.34	17.43	18.93	40.56
Dense Retrieval ( <i>Poly-encoder</i> )	46.05	52.57	19.72	21.46	48.41
DiffKG	12.25	20.99	14.44	9.37	13.23
<b>SURGE</b> ( <i>ours, semi-supervised</i> )	<b>51.00</b>	<b>57.63</b>	<b>21.87</b>	<b>23.03</b>	<b>55.79</b>
Gold Response ( <i>oracle</i> )	93.30	95.21	28.62	29.06	85.75

## Dialogue Evaluation A - (1 out of 15)

Given a dialogue context on the left (A is the user and B is the agent), we provide three responses on the right.  
Please rate each response -- scale from 1 to 3 for each criterion (consistency, informativeness, fluency).

\* Required

---

**Please keep in mind these criteria**

When scoring, please consider the relative quality of each response, and use the neutral score sparingly.

- Consistency: Does the response make sense in the context of the conversation?  
example)  
Context: Can you recommend the song of David Guetta?  
Good Response: Yes, I would like to recommend Titanium.  
Bad Response: Yes, I like David Guetta.
- Informativeness: Does the response contain correct and enough information?  
\* We recommend you to use the internet search whether the response contains correct facts.  
example)  
Context: Do you know anything about the actor Adam Brown?  
Good Response: Adam Brown starred in the movie The Hobbit: An Unexpected Journey.  
Bad Response1 (no information): I don't know.  
Bad Response2 (wrong fact): Adam Brown starred in King Kong.
- Fluency: Is the response grammatically correct and naturally sound?  
example)  
Context: What do you think about Toni Kroos?  
Good Response: He played for Germany, right?  
Bad Response: I think he is midfielder midfielder midfielder midfielder.

Figure 9: **Human Evaluation Instructions.** To measure the qualitative performances of the generated responses, annotators are provided with the following instruction on three criteria – consistency, informativeness, and fluency.

## G Human Evaluation

In this section, we describe the details of human evaluation used in Table 7 of the main paper. We request the annotators to evaluate the responses generated from two baselines (i.e., ALL Knowledge and Space Efficient) and our SURGE framework in response to the given dialogue context, according to three criteria – consistency, informativeness, and fluency. Figure 9 is the instructions provided to each annotator. Specifically, regarding the consistency metric, we ask annotators to check whether the generated response makes sense in the context of the conversation. For informativeness, we ask annotators to check whether the response contains correct and enough information, whereby experiment participants are recommended to use the internet search, to check whether the response contains correct facts. In addition to this, we also provide the dialogue-related facts from Freebase as a reference for fact checking for annotators. For fluency, we ask annotators to check whether the response is grammatically correct and naturally sound.

## H Retrieval and Generation Examples

In this section, we provide examples for knowledge retrieval and response generation, for the given dialogue history.

**Embedding Space Visualization** In Figure 10, we present a larger version of Figure 5 in the main paper. Specifically, we embed the hidden representations before the projection layer for each graph (star) and the embedding of the generated text (circle) through the dimensionality reduction using t-SNE (van der Maaten & Hinton, 2008). As mentioned in the main paper, the visualization highlights that our SURGE framework with graph-text contrastive learning generates more distinct responses to different subgraphs, unlike the one without graph-text contrastive learning which shows less variety over responses even with different graphs.

**Retrieval Examples** We provide the retrieval examples of various models, such as random retrieval, sparse retrieval and our SURGE models. In particular, in the first (top) example of Figure 11, we are given a dialogue context in regard to books for Richard Maxwell, and baselines including random and BM25 retrievers select the facts associated to the entity Richard Maxwell, which are but irrelevant to the ongoing conversation, for example, (Richard maxwell, is-a Theatre director). Also, as shown in the second (bottom) example of Figure 11, we observe that the simple term-based matching model (i.e., BM25) cannot contextualize the current and previous dialogues, but retrieves the facts associated to frequent words, for example, song, which are less meaningful for the user’s question. In contrast to baselines, as our SURGE framework trains a retriever in an end-to-end fashion, it first contextualizes the given dialogue context, and then accurately retrieves relevant knowledge.

**Generation Examples** We provide the generation examples from our model. To be specific, we provide the dialogue context along with its corresponding retrieved subgraph and generated response obtained from our SURGE framework. In Figure 12 and Figure 13, we provide the correct examples: our model retrieves a context-relevant subgraph, but also generates a factual response from retrieved knowledge. On the other hand, in Figure 14, we provide the failure cases. In particular, as shown in the first row of Figure 14, the fact in the knowledge graph could be ambiguous or inaccurate, as it defines the release year of the book – Wicked – as both 2008 and 2014. Moreover, we further provide the failure example on retrieval in the second row of Figure 14, where the user asks about the Bourne Legacy, while the dialogue agents retrieve the irrelevant knowledge to the question. Finally, we show the common problem in PLMs in the last row of Figure 14, where the generative model repeats the meaningless words at the end, while the retriever correctly selects the relevant knowledge.

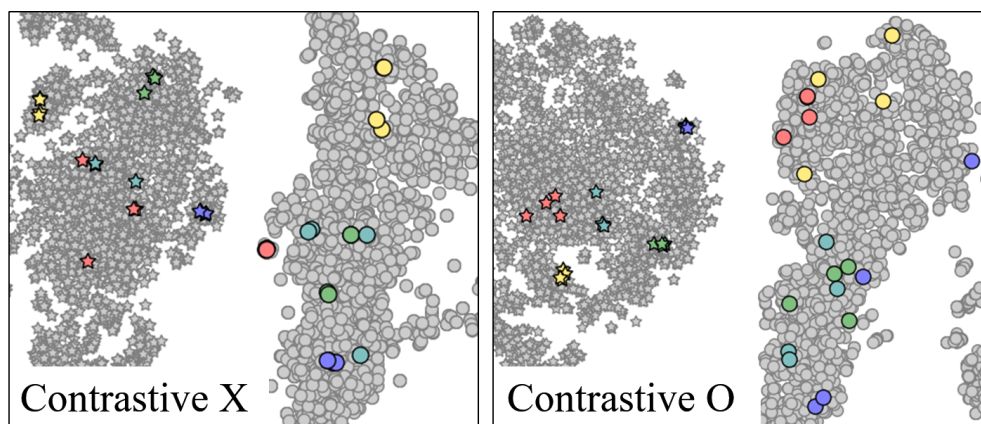


Figure 10: Large version of Figure 5 in the main paper. Stars indicate the embedding of graph and circles indicate the embedding of decoder hidden states (text), respectively.

---

**Dialogue Context**

A: Could you recommend any books written by Richard Maxwell?

---

**Gold Knowledge**

Richard maxwell, ~written\_by, a tale of two cities

---

**Random Knowledge**

Richard maxwell, sibling, jan maxwell

Screenwriter, ~is-a, Richard maxwell

Theatre director, ~is-a, Richard maxwell

---

**BM25 Knowledge**

Richard maxwell, is-a, Theatre director

Screenwriter, ~is-a, Richard maxwell

Richard maxwell, organization founded, new york city players

---

**Our Knowledge**

Richard maxwell, ~written\_by, a tale of two cities

Richard maxwell, sibling, Jan maxwell

---

---

**Dialogue Context**

A: I like Adam Levine.

B: OMG me too! I love that song Moves Like Jagger.

A: Yes, Love that too. It is really fun. Can you tell me more.

B: Did you know it's considered a power pop song?

A: No, I did'n. Do you know Love the way you Lie?

---

**Gold Knowledge**

Song, ~kind of composition, Love the way you lie

Love the way you lie, composer, Eminem

---

**Random Knowledge**

Blue monday, kind of composition, Song

The look of love, kind of composition, Song

Bad romance, kind of composition, Song

---

**BM25 Knowledge**

Song, ~kind of composition, This charming man

Behati prinsloo, ~spouse (or domestic partner), Adam levine

Song, ~kind of composition, safe & sound

---

**Our Knowledge**

g minor, ~key, Love the way you lie

Eminem, ~composer, Love the way you lie

Skylar grey, ~composer, Love the way you lie

---

Figure 11: Examples of the dialogue history with its corresponding gold knowledge as well as the retrieved knowledge from random retrieval and sparse retrieval baselines and from our SURGE framework. The retrieved fact is represented as the format of (head, relation, tail), where ~symbol in the front of relation (i.e., ~relation) in the retrieved knowledge denotes the inverse relation.

---

**Dialogue Context**

A: Do you like Rob Marshall? My friends have recommend his films.

---

**Retrieved Knowledge**

Nine, has\_genre, Drama

Pirates of the Caribbean: On Stranger Tides, directed\_by, Rob Marshall

Memoirs of a Geisha, directed\_by, Rob Marshall

---

**Generated Response**

B: Yes, he wrote Pirates of the Caribbean: On Stranger Tides, Memoirs of a Geisha Nine and many others. Have you seen either of those?

---

**Dialogue Context**

A: Are there any works by Chris Cleave, which you can recommend me?

B: Certainly, wrote The Other Hand, Incendiary, and Old. Have you read them?

A: No, I haven't read The Other Hand out of the ones you mentioned. What genre is it?

---

**Retrieved Knowledge**

The Other Hand, release\_year, 2008

2008, ~release\_year, The Other Hand

The Other Hand, has\_genre, Literary fiction

---

**Generated Response**

B: It's Literary fiction, and was released in 2008.

---

**Dialogue Context**

A: Who wrote Of Mice and Men?

B: It was written by John Steinbeck

A: Didn't he also write The Red Pony?

B: Yes He also wrote The Red Pony

A: Can you suggest other titles by him to read?

---

**Retrieved Knowledge**

Cannery Row, written\_by, John Steinbeck

The Grapes of Wrath, written\_by, John Steinbeck

Tortilla Flat, written\_by, John Steinbeck

---

**Generated Response**

B: Yes he also wrote Tortilla Flat

---

Figure 12: Examples of the dialogue history with its corresponding retrieved knowledge and generated response from our SURGE framework. The fact is represented as the format of (head, relation, tail), where ~symbol in the front of relation (i.e., ~relation) in the retrieved knowledge denotes the inverse relation. In this example, we only provide the correct cases of both retrieval and generation.

---

**Dialogue Context**

A: Could you recommend some movies by director Simon Wood?

---

**Retrieved Knowledge**

The One That Got Away, written\_by, Simon Wood  
Simon Wood, ~written\_by, The One That Got Away  
Author, ~is-a, Simon Wood

---

**Generated Response**

B: Simon Wood directed The One That Got Away. Have you seen that?

---

---

**Dialogue Context**

A: I like David McCullough. Could you recommend any books of him?  
B: Sure. He wrote The Greater Journey: Americans In Paris. Also, he wrote some documentary and Indie films.  
A: Thank you for the information. When was The Greater Journey: Americans In Paris released?

---

**Retrieved Knowledge**

The Greater Journey: Americans In Paris, release\_year, 2011  
2011, ~release\_year, The Greater Journey: Americans In Paris  
Paris, written\_by, Edward Rutherfurd

---

**Generated Response**

B: It was released in 2011. It's a great book.

---

---

**Dialogue Context**

A: I like Tony Tanner, what do you think?

---

**Retrieved Knowledge**

Sense and Sensibility, written\_by, Tony Tanner  
Tony Tanner, ~written\_by, Mansfield Park  
Mansfield Park, written\_by, Tony Tanner

---

**Generated Response**

B: I think he is a great actor. He starred in Sense and Sensibility and Mansfield Park.

---

Figure 13: Examples of the dialogue history with its corresponding retrieved knowledge and generated response from our SURGE framework. The fact is represented as the format of (head, relation, tail), where ~symbol in the front of relation (i.e., ~relation) in the retrieved knowledge denotes the inverse relation. In this example, we only provide the correct cases of both retrieval and generation.

---

**Dialogue Context**

A: Do you know what books Jennifer L. Armentrout has written?

B: She wrote Wicked, White House Kiss and Wait For You. Are you interested in her other books as well?

A: No thanks. What genre is Wicked?

B: Wicked is a Play that won the Pulitzer Prize for Drama. Would you like more information about it?

A: Sure, sounds interesting

---

**Retrieved Knowledge**

Wicked, release\_year, 2008

2008, ~release\_year, Wicked

Wicked, release\_year, 2014

---

**Generated Response**

B: It is a play written by Jennifer L. Armentrout and released in 2008.

---

---

**Dialogue Context**

A: I like the movie Armageddon.

B: Did you know Armageddon is written by Tony Gilroy?

A: Oh, didn't he also write The Bourne Legacy?

---

**Retrieved Knowledge**

Michael Clayton, starred\_actors, Tony Gilroy

Michael Clayton, written\_by, Tony Gilroy

Proof of Life, written\_by, Tony Gilroy

---

**Generated Response**

B: Yes, he did. He also wrote Michael Clayton.

---

---

**Dialogue Context**

A: I like Deadpool or any movies like Deadpool.

---

**Retrieved Knowledge**

Deadpool, starred\_actors, Ryan Reynolds

Deadpool, starred\_actors, Stefan Kapicic

Deadpool, starred\_actors, Ed Skrein

---

**Generated Response**

B: Deadpool is a great movie. Stefan Kapicic starred in it. Stefan Kapicic also starred in The Last Man and The Last Man.

---

Figure 14: Examples of the dialogue history with its corresponding retrieved knowledge and generated response from our SURGE framework. The fact is represented as the format of (head, relation, tail), where ~symbol in the front of relation (i.e., ~relation) in the retrieved knowledge denotes the inverse relation. In this example, we only provide the failure cases due to the problem on data (first row), retrieval (second row), and generation (third row).