CONTEXTUALIZED SCENE IMAGINATION FOR GENERATIVE COMMONSENSE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans use natural language to compose common concepts from their environment into plausible, day-to-day scene descriptions. However, such generative commonsense reasoning (GCSR) skills are lacking in state-of-the-art text generation methods. Descriptive sentences about arbitrary concepts generated by neural text generation models (e.g., pre-trained text-to-text Transformers) are often grammatically fluent but may not correspond to human common sense, largely due to their lack of mechanisms to capture concept relations, to identify implicit concepts, and to perform generalizable reasoning about unseen concept compositions. In this paper, we propose an Imagine-and-Verbalize (I&V) method, which learns to imagine a relational scene knowledge graph (SKG) with relations between the input concepts, and leverage the SKG as a constraint when generating a plausible scene description. We collect and harmonize a set of knowledge resources from different domains and modalities, providing a rich auxiliary supervision signal for I&V. The experiments demonstrate the effectiveness of I&V in improving language models on both concept-to-sentence and concept-to-story generation tasks, while enabling the model to learn well from fewer task examples and generate SKGs that make common sense to human annotators [1].

1 INTRODUCTION

Humans describe everyday scenes in natural language based on their understanding of common concepts encountered in their environment (Tincoff & Jusczyk, 1999). Analogously, the task of *generative commonsense reasoning* (GCSR) asks machines to generate a description of everyday situations based on a set of concepts and an initial context (Liu et al., 2020; Li et al., 2021). For example, given concept words {*dog, frisbee, catch, throw*}, a machine is expected to generate a plausible description, e.g., "*A man throws a frisbee and his dog catches it in the air*". Machines with GCSR skills would communicate fluidly with humans, e.g., when summarizing a document by preserving its key details (Sha, 2020), composing a creative story according to a set of clues (Yao et al., 2019), and generating a conversation reply that includes specified keywords (Mou et al., 2016).

GCSR poses three unique challenges for automatic text generation methods. To depict plausible scenes when composing sentences, machines require commonsense knowledge to reason about the relations between concepts and the affordances of objects (e.g., "*dog*" performs the action "*catch*" but not the action "*throw*"). Moreover, machines require a compositional generalization ability (Keysers et al., 2019), i.e., the ability to judge the plausibility of a new concept composition that has not been observed during training, and to identify concepts related to the scene that are not explicitly provided (e.g., "*person*" to perform "*throw*" in the above example).

GCSR can be directly attempted by fine-tuning pre-trained text-to-text language models (LMs) (Raffel et al., 2019; Radford et al., 2019). While pre-trained LMs capture certain encyclopedic knowledge mentioned in text corpora (e.g., Wikipedia) (Petroni et al., 2019) and can combine concepts in novel ways, they may generate grammatically fluent but implausible sentences that conflict with human common sense (Lin et al., 2020). This is because LMs have no intrinsic mechanism to reason over high-level relations between concepts Zhou et al. (2020). To close the knowledge gap, recent work augment LM input with knowledge graph triples (e.g., (*dog, CapableOf, catch*)) retrieved

[1] Code and data used in our experiments can be found at `https://anonymous.4open.science/r/ImagineAndVerbalize-EE98/`.

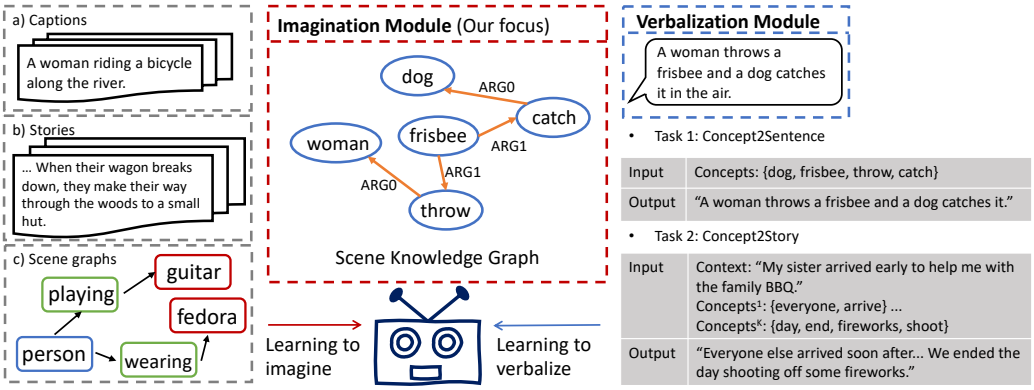

Figure 1: Overview of the proposed I&V method: (1) We leverage SKGs for unifying scene knowledge from different resources. (2) We pre-train a contextualized imagination module to construct an SKG for a set of concepts, based on the collected SKG instances. (3) At inference time, our verbalization module realizes the generated SKG into natural language.

from ConceptNet (Liu et al., 2020; Li et al., 2020), or prototype sentences that cover input concepts retrieved from external text corpora (Fan et al., 2020; Wang et al., 2021). However, despite the input augmentation, GCSR skills are implicitly learned based on the concept-text pairs in the training data, without explicit supervision. While some recent work propose content planning in story generation in the form of plots or scripts (Yao et al., 2019; Fan et al., 2019), only the narrative order of concepts are planned in those methods instead of their plausible roles and relations. Given the complexity of the GCSR task, machines need a direct mechanism to create a high-level relational representation of the provided concepts, which would allow them to judge the plausibility of their combination.

In this paper, we propose to model an explicit *scene imagination* step which constructs a structured representation of a plausible scene based on input concepts and initial context. The scene imagination module formalizes the background knowledge required for the reasoning through a contextualized relational graph, called *scene knowledge graph* (SKG). An SKG allows us to collect and harmonize diverse commonsense knowledge across resources and modalities into a comprehensive SKG distribution (see Figure 1 for an illustration). We develop an imagine-and-verbalize framework: an imagination module learns to construct a contextualized SKG from input concepts and context by pretraining over a large amount of external SKGs; a verbalization module learns to faithfully realize the imagined SKG into natural language by training over downstream datasets. By learning from a large number of diverse SKGs, our method is able to capture plausible relations between concepts. By integrating these SKGs with LMs, the imagination module is able to compose new objects in novel ways, and identify implicit concepts for a scene. Imagine-and-verbalize decomposes the challenging scene description task into two realistic tasks for which a wealth of training data can be collected, simulaneously enabling for effective and explainable GCSR.

We experiment with two GCSR tasks and three scene graph resources, observing consistently better or competitive performance to SotA baselines. We find that (1) SKGs extracted from visual captions and story datasets are more helpful than other resources; (2) our model can learn faster (with less training data) with the help of scene imagination; and (3) the imagination module with a larger backbone LM demonstrates larger capacity in encoding commonsense knowledge. Our human evaluation study on the generated imagination indicates that these SKGs capture common sense and that the verbalization module generates the text by following the guidance of the imagination.

2 METHOD

Formally, in GCSR, we consider a list of *concepts sets* $\{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^K\}$ and a textual *context* $\mathbf{c} \in \mathcal{C}$ as input. Each concept set $\mathbf{x}^i$ is unordered and consists of multiple concept words $\{x_j\}$. A concept word $x_j \in \mathcal{X}$ (or *concept* for brevity) is a commonly seen object (nouns such as "*dog*" or "*frisbee*") or commonly performed action (verbs such as "*throw*" or "*catch*"). The goal of GCSR is to generate K sentences $\{\mathbf{y}^1, \mathbf{y}^2, ..., \mathbf{y}^K\}$, each describing a plausible situation following human common sense for a concept set $\mathbf{x}^i$. The i-th sentence $\mathbf{y}^i \subset \mathcal{Y}$ should be generated using all concepts in $\mathbf{x}^i$.

We consider two variants of GCSR: 1) *concepts-to-sentence* generation (Lin et al., 2020), where no context is given (i.e., $\mathbf{c}$ is empty) and only one concept set is provided ($K = 1$); and 2) *concepts-to-story* generation task, where $\mathbf{c}$ is the leading sentence of a multi-sentence story and more than one concept sets are provided, each corresponding to one sentence to be generated ($K > 1$). Both tasks are evaluated by comparing the machine-generated text with human-generated (gold) references.

2.1 The Imagine-and-Verbalize Approach

Pre-trainined LMs struggle with learning a generalizable mapping from concepts to plausible sentences solely based on the training data. Augmenting concepts with external knowledge to form the input $\mathcal{X}^{'}$ and fine-tuning a pretrained LM to model $P(\mathcal{Y}|\mathcal{C}, \mathcal{X}^{'})$ (Liu et al., 2020; Fan et al., 2020; Li et al., 2021) alleviates this issue partially, while still learning a direct mapping of $\{\mathcal{C}, \mathcal{X}^{'}\} \rightarrow \mathcal{Y}$. In this work (Figure 1), we decompose the GCSR task into two sub-tasks: contextualized scene generation (*imagination*) and scene-aware text generation (*verbalization*).

$$P(\mathcal{Y}|\mathcal{C}, \mathcal{X}) = \sum_{\mathcal{Z}} P(\mathcal{Y}|\mathcal{C}, \mathcal{X}, \mathcal{Z}) P(\mathcal{Z}|\mathcal{C}, \mathcal{X}), \tag{1}$$

where $\mathcal{Z}$ denotes the scene representation for the given concepts and context.

The contextualized scene imagination module $P(\mathcal{Z}|\mathcal{C}, \mathcal{X})$ aims to construct a multi-relational graph representation $\mathcal{Z}$ (scene knowledge graph, or SKG) that describes a plausible scene that involves all input concepts and corresponds to the provided context. To learn this module, we collect a diverse set of SKG instances from different resources and modalities to form a comprehensive distribution of scenes (§2.2). The imagination module is pre-trained over the collected scene instances and learns to generate SKGs depicting plausible day-to-day situation. The imagination module is based on a neural architecture, which enables it to generate concept compositions that might not have been observed during training (§2.3).[2] We leverage the contextualized SKG for text generation with a verbalization module $P(\mathcal{Y}|\mathcal{C}, \mathcal{X}, \mathcal{Z})$ which takes the context, concepts, and the generated SKG as input, and composes a grammatical and plausible scene description in natural language (§2.4).

To perform GCSR, where one or multiple concept sets are given, we apply the imagination module to sample $\mathbf{z^i}$. Since the marginalization over $\mathcal{Z}$ is generally intractable due to the complex structure of the SKGs, we only sample the most probable scene representation $\mathbf{z}^{*\mathbf{i}}$ that maximizes $P(\mathbf{z^i}|\mathbf{c}', \mathbf{x}^i)$, where $\mathbf{c}'$ includes the given context $\mathbf{c}$ and the previously generated $\mathbf{y}^j, (j < i)$. We then apply the verbalization module to generate one sentence at a time by sampling from $P(\mathbf{y}^i|\mathbf{c}', \mathbf{x}^i, \mathbf{z}^{i*})$. Multiple sentences are generated by iteratively applying the imagination and verbalization modules.

2.2 Imagination via Generating SKG

Imagination through SKGs We adopt the term "scene graph" from the computer vision community, and we generalize it to a novel relational schema that represents knowledge from multiple modalities. Our SKG is defined as a relational graph $\mathcal{G} = (\mathcal{E}, \mathcal{R})$ that organizes a set of concepts in a coherent scene that follows common sense. The node set $\mathcal{E}$ of the graph includes both given and implicit concepts, while each relation (edge type) $r \in \mathcal{R}$ denotes how two concepts should be related. We follow the Abstract Meaning Representation (AMR) (Banarescu et al., 2013) schema to consider the core relations between two concepts, which corresponds to the commonsense knowledge required by GCSR. Table 7 in the appendix illustrates a few representative relations and their examples.

Collecting Diverse SKGs We consider two complementary modalities, text and vision, as some concepts and relationships are more likely to occur in one modality versus another. (1) *Textual Modality:* According to pragmatic principles of human language, people generally leave out expected details about common scenes (Grice, 1975). For this reason, we extract SKGs from visual captions and narrative stories, in which human annotators are asked to explicitly describe scenes that may happen using

Table 1: Statistics of the SKG instances collected from different resources.

Knowledge source	# SKGs	# Concepts
Caption-AMR	584,252	22,961
Story-AMR	927,163	41,272
VG-SceneGraph	292,596	41,629
All	1,792,941	84,835

[2]The imagination module can be further fine-tuned over the downstream datasets.

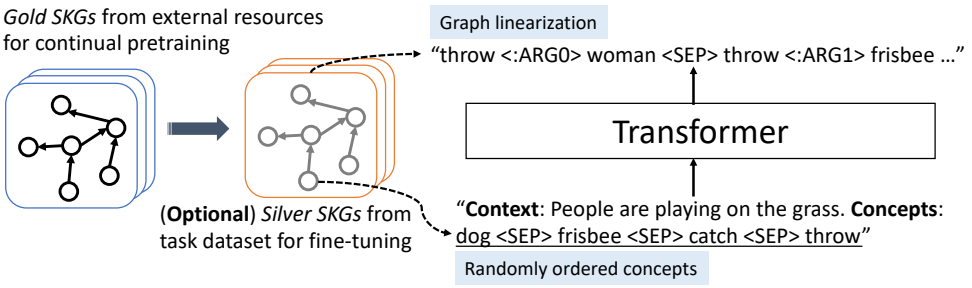

Figure 2: Continual pretraining and fine-tuning of the imagination module to output a linearized SKG based on a sequential input (context and concepts).

descriptive language as shown in Figure 1(a,b). To extract an SKG out of these textual signals, we adopt the AMR parsing tool to transform each sentence into an AMR graph. This process yields a single SKG per sentence. For the story SKGs, we also keep the sentences (up to 256 tokens) that precede the sentence that corresponds to the SKG, as context $\mathbf{c}$. (2) *Visual Modality:* Image captions focus on salient information and may not capture all useful visual signals. Thus, we also capture the scene structures directly from images, by using VisualGenome (Krishna et al., 2016), a large-scale scene graph dataset annotated by humans. To adopt a unified SKG schema, we manually map the relations in scene graphs from VisualGenome to the ones used in textual SKGs. A full set of mapping rules can be found in the appendix (A.1). The statistics of the SKGs collected from each resource/modality are summarized in Table 1. We note that visual scene graphs may be biased towards knowledge about spatial relationships and object affordance, which further motivates our decision to extract SKGs from multiple modalities.

2.3 Learning the Scene Imagination Module

We describe how we pre-train the scene imagination model using multimodal SKG examples collected from diverse sources, and how we fine-tune the imagination module to downstream datasets.

A straightforward way to construct a SKG is to retrieve ones that contains all the given concepts from the collected SKGs. However, performance of such method is limited by the coverage of the SKG collection and will fail when encountering novel concept composition. We propose to model $P(\mathcal{Z}|\mathcal{C}, \mathcal{X})$ with a neural graph generator. Inspired by previous work on (conditional) graph generation (You et al., 2018), we formulate SKG construction as an auto-regressive sequence generation task, where a linearized SKG is generated sequentially conditioned on the context, input concepts, and the graph sequence generated so far. Sequence generation formulation is advantageous, as it can be natively tackled by pre-trained auto-regressive LMs (e.g., GPT-2). Thus, we adopt these LMs as the backbone of our imagination module (Bosselut et al., 2019; Wang et al., 2020).

Linearized SKG Generation To form training instances for the imagination module, we treat the nodes in an SKG instance as input concepts and the linearized SKG as the target output (Figure 2). The input concepts are concatenated into a sequence $\mathbf{x} = [x_1, x_2, ..., x_n]$, preceded by the context $\mathbf{c}' \in \mathcal{C}$. When $\mathbf{c}'$ is not given, we prepend the word "*none*" to the concept sequence. To linearize an AMR-based SKG into a sequence $\mathbf{z} = [z_1, z_2, ..., z_m]$, we adopt the PENMAN serialization format (Goodman, 2020) which converts AMR into a spanning tree over the graph. This format is shown to be more suitable than other linearization strategies like depth-first-search (DFS) in enabling LMs to learn the graph structure (Mager et al., 2020). We conduct DFS and follow PENMAN format to prioritize nodes associated with core relations (e.g., ARG0).

During training, we randomize the order of the concepts at every training step such that the graph generator learns to be invariant to concept order (Zhang et al., 2019). For each training instance, we randomly discard a small subset of the SKG nodes (concepts) in each training epoch. This simulates the scenario in which a subset of the concepts that constitute a scene will be given, thus teaching the model to infer implicit concepts for completing a plausible scene.

Continual-Pretraining and Fine-tuning With both the input concepts (plus context) and the output graph linearized as sequences based on the collected SKG instances, we continually pretrain an auto-

regressive LM to generate $\mathbf{z} = \text{Transformer}(\mathbf{c}', \mathbf{x})$. The training objective is to maximize $P(\mathcal{Z}|\mathcal{C}, \mathcal{X})$ by minimizing the log-likelihood loss:

$$\mathcal{L}_{imagine} = -\sum_{t=1}^{t=m} \log P(z_t|z_{<t}, \mathbf{c}', \mathbf{x}). \quad (2)$$

Our pre-trained imagination module generates an SKG on the fly, and it can be further fine-tuned on downstream datasets, when their distributions of context and concepts are different from the pretraining data (see Figure 2 for illustration). Since downstream datasets cannot be expected to have ground-truth SKGs paired with each training example, we apply the AMR parsing tool described in §2.2 on the training sentences to obtain silver-standard SKGs. We then follow the same training procedure to continually pretrain the module into a customized imagination module for a specific downstream dataset.

2.4 Scene-aware Verbalization

Iterative Imagine-and-Verbalize At model inference time, we apply the trained imagination module iteratively to generate the most plausible SKG for each given concept set $\mathbf{x}^i$, i.e., $\mathbf{z}^{i*} = \arg\max_{\mathbf{z}^i} P(\mathbf{z}^i|\mathbf{c}', \mathbf{x}^i)$, where the context $\mathbf{c}'$ includes both the given context $\mathbf{c}$ and the previously generated sentences $\{\mathbf{y}^j\}$ $(j < i)$. The generated SKG is used by the scene-aware verbalization module to model $P(\mathcal{Y}|\mathcal{C}, \mathcal{X}, \mathcal{Z})$. The verbalization module generates the i-th sentence by sampling from $P(\mathbf{y}^i|\mathbf{c}', \mathbf{x}^i, \mathbf{z}^{i*})$. Multiple sentences are generated iteratively by alternating between the scene imagination (to construct SKG) and verbalization (to produce the next sentence). See Figure 3 for an illustration of this iterative inference process.

Model Training Since both the linearized SKG (generated by the imagination module) and the target sentences are sequences by nature, we design $P(\mathcal{Y}|\mathcal{C}, \mathcal{X}, \mathcal{Z})$ as a sequence-to-sequence generative model and learn this verbalization module by fine-tuning another pre-trained auto-regressive LM, i.e., $\mathbf{y}^i = \text{Transformer}(\mathbf{c}', \mathbf{x}^i, \mathbf{z}^i)$. To form the input for generating the sentence $\mathbf{y}^i$ we concatenate the context $\mathbf{c}'$, the concept set sequence $\mathbf{x}^i$ and $\mathbf{z}^i$into one sequence as illustrated in Figure 3. We then train the model to maximize $P(\mathcal{Y}|\mathcal{C}, \mathcal{X}, \mathcal{Z})$ by optimizing the cross-entropy loss:

$$\mathcal{L}_{verbalize} = -\sum_{t=1}^{t=l} \log P(y_t^i|y_{<t}^i, \mathbf{c}', \mathbf{x}^i, \mathbf{z}^i). \quad (3)$$

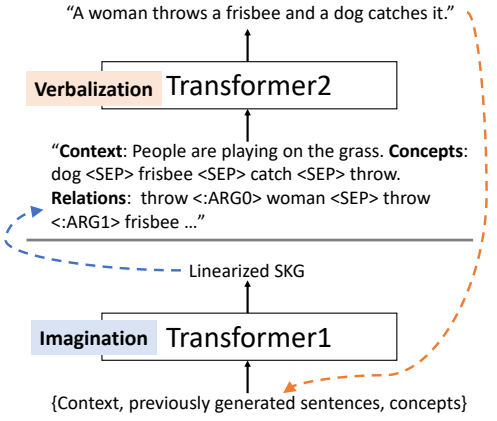

Figure 3: Our I&V method iteratively applies the imagination and the verbalization modules, by generating one sentence in each iteration.

For each training instance $(\mathbf{y}^i, \mathbf{c}', \mathbf{x}^i)$, we construct two types of SKG instances as the input $\mathbf{z}^i$: (1) We perform AMR parsing on $\mathbf{y}^i$ to obtain a silver-standard SKG; (2) We apply the trained imagination module to generate a SKG $\mathbf{z}^{i*} = \arg\max_{\mathbf{z}^i} P(\mathbf{z}^i|\mathbf{c}', \mathbf{x}^i)$, where $\mathbf{c}'$ includes the given context $\mathbf{c}$ and the ground-truth prefix sentences $\{\mathbf{y}^j\}$ $(j < i)$. We find it beneficial to train the verbalization module over these two types of SKGs in experiments. During inference, the SKG $\mathbf{z}^i$ is generated by the imagination module, while $\mathbf{c}'$ includes the given context $\mathbf{c}$ and the previous sentences $\{\mathbf{y}^j\}$ $(j < i)$ generated by the verbalization module.

3 Experimental Setup

Tasks & Datasets We consider two GCSR tasks: Concept2Sentence and Concept2Story. *(1) Concept2Sentence* is a task of generating a single sentence for a given set of concepts and no context. We evaluate concept2sentence on the CommonGen (Lin et al., 2020) benchmark. Since the labels of the official test set are not publicly available, we submit our method to the leaderboard to obtain its performance. Notably, the concept sets in CommonGen's test set are novel and do not appear in

the training set. We also create an in-house split of CommonGen to facilitate comparison between different variants of our method and the baselines. *(2) Concept2Story* is a generalization of the concept2sentence task, where the goal is to generate a coherent story with $K = 4$ sentences given a set of concepts and an initial verbal context. We construct two benchmarks based on the Visual Story Telling (VIST) (Huang et al., 2016) and ROCStories (Mostafazadeh et al., 2016) datasets. Following CommonGen, we conduct part-of-speech tagging over the sentences and further lemmatize the recognized verbs and nouns to obtain the concept sets.

Baselines (1) *Concept2Sentence:* We consider several recent submissions to the leaderboard of CommonGen that leverage auxiliary information for GCSR. KFCNet (Li et al., 2021), Re-T5 (Wang et al., 2021), and EKI-BART (Fan et al., 2020) are prototype-based models, which retrieve sentences containing as many input concepts as possible from external captions and NLI datasets, and then use these sentences as auxiliary inputs. VisCTG (Feng et al., 2021b) is an image-augmented model which retrieves images from Google by using concepts as a query, followed by an image captioning model that generates captions as auxiliary inputs. KG-BART (Liu et al., 2020) is a knowledge graph-augmented model which retrieves relations between concepts from ConceptNet as auxiliary inputs. SAPPHIRE (Feng et al., 2021a) is a keyword-based model which extracts keywords from sentences as auxiliary inputs only during training. We also compare to Node2Text, which fine-tunes a pre-trained auto-regressive LM without auxiliary input. It takes the concatenation of concepts as input and outputs the target sentences. (2) *Concept2Story:* In addition to the Node2Text baseline, we experiment with two representative methods from the controlled text generation literature. Plan-and-write (Yao et al., 2019) first generates storyline keywords, then uses the keywords to generate a story. We use the concept set and context to generate storyline keywords. Action-Plan (Fan et al., 2019) uses predicate-argument pairs as storyline. We adapt the KFCNet model to retrieve prototype sentences. All Concept2Story baselines are used in an iterative generation pipeline, to enable fair comparison to our method.

Evaluation Metric We evaluate systems against the K reference sentences provided by a dataset, by measuring the similarities between the machine-generated text and the gold references. Following CommonGen (Lin et al., 2020), we adopt widely-used automatic metrics for evaluating text generation, focused on (1) n-gram overlap: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee & Lavie, 2005), and (2) concept association: CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016). Lin et al. (2020) reports that SPICE yields the best correlation with human judgments and thus we used it as the main evaluation metric.

Implementation Details For the main experiments, we develop the imagination module by continually pre-train a T5-large model over the caption, story, and vision SKGs. We then further adapt the imagination module over each task dataset annotated with the silver-standard SKGs by further fine-tuning. To train the verbalization module, we fine-tune T5-base and BART-large as two backend LMs. During training, we use both silver-standard SKGs and generated SKGs, while averaging the training loss associated with each of them. We use the Adam optimizer with weight decay $1e-2$. We search the optimal hyper-parameters based on the perplexity over the development set, where the learning rate is chosen from $\{2e-6, 1e-5, 3e-5, 1e-4\}$, the batch size is chosen from $\{8, 16, 32, 64, 128\}$.

4 Results and Analysis

We design experiments to answer the following questions: (1) Does contextualized scene imagination improve the performance of GCSR models? (2) Does imagination allow GCSR models to learn with less data? (3) How does each source of scene knowledge for pretraining affect the GCSR performance? (4) Do generated SKGs make common sense and correspond to the generated text?

4.1 Main Results

We compare our proposed approach with state-of-the-art neural text generation methods on two GCSR tasks to understand whether scene imagination helps GCSR. Table 3 shows the performance of different models on the CommonGen test set. We have the following observations. First, I&V drastically improves the vanilla T5-large model, demonstrating the effectiveness of the contextualized imagination module in GCSR. Second, our model also outperforms other models using dif-

Table 2: Performance of the compared methods on the Concept2Story tasks. Best results are bold-faced. We mark them with an asterisk if they exceed the second best with statistical significance (p-value < 0.05).

	Concept2Story-VIST						**Concept2Story-ROC**					
	T5-base			BART-large			T5-base			BART-large		
Model	BLEU-4	CIDEr	SPICE	BLEU-4	CIDEr	SPICE	BLEU-4	CIDEr	SPICE	BLEU-4	CIDEr	SPICE
Node2Text	20.64	25.41	58.55	18.52	22.91	55.48	23.31	29.32	57.66	20.60	26.09	53.80
Keyword	16.75	21.87	56.23	15.62	20.86	55.49	22.24	27.05	50.41	22.14	27.40	49.52
Action-Plan	17.84	22.77	57.11	16.20	21.10	54.77	21.15	27.32	56.14	20.45	26.29	54.32
Prototype	20.28	25.05	58.17	**22.81**	**26.93**	58.84	23.59	29.48	57.68	26.76	31.60	58.35
I&V	**21.05***	**25.78***	**59.21***	22.45	26.80	**59.11***	**26.77***	**32.33***	**60.63***	**28.30***	**33.40***	**60.39***

ferent auxiliary inputs, including prototypes (Re-T5 and EKI-BART), knowledge facts (KG-BART) and images (VisCTG), showing the benefit of SKGs over these knowledge sources. Moreover, our model underperforms KFCNet, which retrieves prototypes from a much larger corpus (over 70M sentences, vs. 1.7M in ours). Although the retrieved prototypes provide high coverage over the queried concepts (Li et al., 2021), their model is supervised to compose sentences that are very similar to those existing prototypes. It is thus unclear whether that model is conducting commonsense reasoning or just mimicking the prototypes. We filter out any collected SKGs that cover the concept sets in the downstream test data. This ensures that the imagination module is examined with its compositional generalization.

Table 2 shows the experimental results by I&V on the two Concept2Story datasets using T5-base and BART-large as the backend respectively. Among most evaluation metrics, our method outperforms Node2Text and baselines with other intermediate representations incorporated in the same backends. This demonstrates that our imagination module can provide contextualized scene imagination that are more helpful in guiding long narrative generation.

Table 3: Performance comparison with the top-ranked, published models on the official CommonGen test set. *Note that KFCNet uses a much larger corpora (over 70M) to retrieve prototypes and on average less than one concept in the concept sets is not covered (Li et al., 2021), while we filter out any SKGs that contain concept sets that overlap with CommonGen dataset.

Model	**BLEU-4**	**CIDEr**	**SPICE**
KFCNet (Li et al., 2021)*	**43.62**	**18.85**	**33.91**
RE-T5 (Wang et al., 2021)	40.86	17.66	31.08
VisCTG (Feng et al., 2021b)	36.94	17.20	29.97
SAPPHIRE Feng et al. (2021a)	37.12	16.90	29.75
KG-BART Liu et al. (2020)	33.87	16.93	29.63
EKI-BART Fan et al. (2020)	35.95	17.00	29.58
T5-base (our implementation)	33.81	15.79	28.34
T5-large (our implementation)	32.85	15.76	28.38
T5-large (reported)	31.96	15.13	28.86
I&V (T5-base)	40.16	17.44	30.57
I&V (T5-large)	<u>40.57</u>	<u>17.71</u>	<u>31.29</u>

4.2 Performance Analysis

How does the knowledge source affect GCSR? We perform an ablation study in order to understand how effectively each source of SKGs contributes to the imagination. Specifically, we use each of the following SKG sources to pre-train an imagination module using T5-large as the backend: the silver-standard SKGs extracted from the training set from the downstream task (Task-AMR), and the external SKGs: Caption-AMR, Story-AMR, and VG-SceneGraph (§2.2). For CommonGen, we do not further fine-tune the imagination module in order to distinguish the contributions from each knowledge source more clearly. For Concept2Story (ROCstories), we conduct further fine-tuning using the task-AMR. Since this task provides the context as input, we find it helpful to adapt the imagination module with the task dataset.

The results are shown in Table 4 and we have the following observations. For CommonGen, the contribution comes mostly from the SKGs based on Caption-AMR while being less from VG-SceneGraph. This may due to the fact that VG-SceneGraph is biased towards spatial relations and attributes of objects. For Concept2Story, we find both Story-AMR and Caption-AMR to be helpful for continual pretraining. The former teaches the model to generate contextualized imagination which is necessary for story generation in particular while the latter teaches the model about general commonsense knowledge. For both datasets, the imagination modules that are pre-trained over all the SKG instances yield significantly better results than the ones trained on the task-AMR datasets. This validates our intuition that different sources of SKGs contain complementary commonsense knowledge, and they should be used together for machine imagination.

Table 4: Performance of our method using different SKG sources to train the imagination module, with T5-large as the backbone LM.

	CommonGen (in-house)			Concept2Story-ROC		
Knowledge Source	BLEU-4	CIDEr	SPICE	BLEU-4	CIDEr	SPICE
Task-AMR	28.87	15.74	31.22	23.14	29.25	57.91
Caption-AMR	32.21	16.14	32.16	23.77	29.76	58.46
Story-AMR	23.73	13.51	27.53	24.17	30.10	58.59
VG-SceneGraph	21.00	13.36	29.07	22.84	25.33	53.96
All-SKG	**33.27**	**16.95**	**33.49**	**26.77**	**32.33**	**60.63**

Table 5: SPICE performance of our method using different sizes of T5 as backbone for the imagination module.

Dataset / Backbone LM	**T5-base**	**T5-large**
CommonGen (in-house)	32.00	33.49
Concept2Story-ROC	59.56	60.63

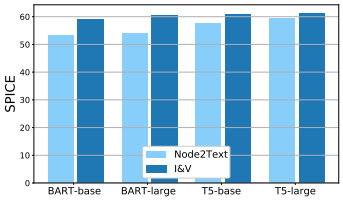

Figure 4: Ablation study on backbone LM sizes of our verbalization module and Node2Text using the Concept2Story-ROC dataset.

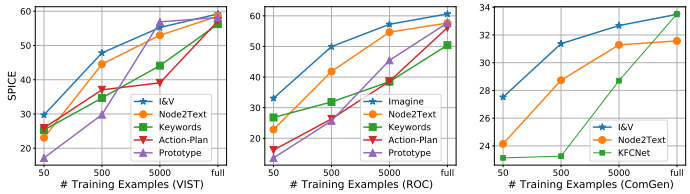

Figure 5: Performance study on the low-resource training setting. SPICE scores on the three benchmark datasets with different number of training examples are reported for comparison.

How does the backbone LM size affect our moperformance? We also ablate the LM architecture of the imagination module and the verbalization module respectively to see how our method work with different pre-trained LMs. For the imagination module, we use T5-base and T5-large. This is to investigate how the capacity of LMs affects the learning of scene knowledge. The results are shown in Table 5. Compared to T5-large, we observe a slight performance drop for T5-base, which indicates that larger LMs are able to encode our rich set of SKG instances in a more expressive manner. For the verbalization module, we use BART-base/large and T5-base/large. The results are shown in Figure 4. We observe that compare to baseline, our method consistently yields a better performance regardless of what LM architecture is used.

Does imagination allow models to learn (faster) with less data? Next, we study how the indirect supervision provided to the imagination module help the system effectively learn with limited task-specific training data. Accordingly, we conduct a low-resource experiment where we randomly sample $\{50, 500, 5000\}$ training and development examples from each dataset. For each data size, we use 5 random seeds to obtain 5 different training and development splits. On each split, we train and test with random initialization of 3 seeds, and we report the average on the total 15 ways of results. In this study, the imagination module is fixed untrainable after continual pretraining and is not fine-tuned over the sampled task datasets.

Figure 5 shows that our model consistently outperforms the baselines, and the performance gain is larger when less training data are used. This indicates that rich sources of SKGs provide practical forms of indirect supervision to complement limited task-specific training data. The robustness of our model in low-resource settings also justifies the need for including contextualized SKGs as an intermediate representation, which further enhances the verbalization module to generate plausible sentences even with little training data.

4.3 Human Evaluation on Generated SKGs

We conduct human evaluation on the SKGs generated by our imagination module to examine their two aspects: (1) whether the SKGs organize concepts in a way that follows common sense and (2) whether the SKGs align with the generated text. We ask three human annotators to score 100 SKG instances generated by our imagination module for CommonGen and Concept2Story datasets by answering "yes/no" to the above two questions. The detailed annotation criterions and guideline can be found in the appendix A.2. Table 6 shows the evaluation results where we get a fair inner agreement

measured by Fleiss Kappa. We observe that the generated SKGs generally follow human common sense in a high degree across three datasets, which demonstrates the effectiveness of training the imagination module to learn useful commonsense knowledge with vast indirect supervision from different resources and modalities. Moreover, the SKGs are well-aligned with the generated text, which indicates that the verbalization module consistently follows the guidance of the imagination module when generating sentences.

5 Related work

Knowledge-Enhanced GCSR Recent works (Liu et al., 2020; Li et al., 2021) on GCSR propose to retrieve external knowledge to enhance the text generation. Prototype-based models, including EKI-BART (Fan et al., 2020), Re-T5 (Wang et al., 2021), and KFCNet (Li et al., 2021) retrieve massive prototype sentences from external corpora (over 70M) like visual captions and Wikipedia as auxiliary input to the LM. Though the retrieved prototype sentences provide high coverage on the concepts, their model is supervised to compose sentences that are very similar to those existing prototypes. It is thus unclear whether their models are conducting commonsense reasoning or only mimicking the prototypes. KG-BART (Liu et al., 2020) incorporates the embedding of relational facts about the concepts from ConceptNet into both the encoders and decoders of the BART architecture (Lewis et al., 2020). As there could be multiple relations between two concepts, it is unclear how to select the relation that fits a given context (Fadnis et al., 2019). Our imagination module infers the relations between concepts by taking all the concepts into consideration and organizes them in a coherent way.

Table 6: Human evaluation on the generated SKGs regarding common sense (CS) and alignment with text (AL). IAA is Fleiss' Kappa inter-annotator agreement.

	CS	AL	IAA
CommonGen	77.00	66.67	0.288
VIST	79.33	68.00	0.252
ROC	83.67	75.67	0.158

Content Planning Our method is also related to prior works (Goldfarb-Tarrant et al., 2020) that propose intermediate representations as a way to "plan ahead" before generating long narratives. Plan-and-write (Yao et al., 2019) generates chains of keywords as a storyline, but do not consider relations between keywords (concepts) as we do. Action-plan (Fan et al., 2019) takes a step further by using predicate-argument with semantic role labeling, but still does not involve all the concepts in a sentence. Moreover, these methods are limited to obtaining supervision from task-specific datasets, whil we gather effective indirect supervision signals from rich multi-source, multi-modal SKG representations without the need for additional annotations.

Machine Imagination There are also some prior works exploring machine imagination in different tasks. Lin & Parikh (2015) proposes to generate images as visual evidence for solving commonsense question answering. Elliott & Kádár (2017) improve multi-modal translation by training the model to translate a sentence and imagine via jointly learning a visually-grounded representation. VisCTG (Feng et al., 2021b) retrieves Google images to visually ground the Concept2Sentence generation. Aforementioned works directly captures images to enrich the generation, while in this work, we investigate relational graphs (SKG) as the representation of machine imagination which focuses on reasoning with structured knowledge captured in both language and vision.

6 Conclusions

This paper proposed to enhance neural architectures for GCSR with an intermediate imagination layer. We divided the GCSR process into two steps: imagination, which generated a plausible scene knowledge graph for a given set of concepts, and verbalization, which transformed this scene graph into a fluent sentence that corresponds to human common sense. The method was trained with diverse scene knowledge graphs derived from both text and vision modalities. Our experiments demonstrated the ability of the proposed method to perform GCSR effectively, by describing plausible scenes, and efficiently, by requiring less training data. The image caption graphs proved most beneficial to learn from. Future work should investigate the impact of imagination on interactive commonsense tasks, like dialogue generation, and include scene graphs from the audio modality.

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pp. 382–398. Springer, 2016.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pp. 178–186, 2013.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4762–4779, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1470. URL https://www.aclweb.org/anthology/P19-1470.

Desmond Elliott and Akos Kádár. Imagination improves multimodal translation. *arXiv preprint arXiv:1705.04350*, 2017.

Kshitij Fadnis, Kartik Talamadupula, Pavan Kapanipathi, Haque Ishfaq, Salim Roukos, and Achille Fokoue. Heuristics for interpretable knowledge graph contextualization. *arXiv preprint arXiv:1911.02085*, 2019.

Angela Fan, Mike Lewis, and Yann Dauphin. Strategies for structuring story generation. *arXiv preprint arXiv:1902.01109*, 2019.

Zhihao Fan, Yeyun Gong, Zhongyu Wei, Siyuan Wang, Yameng Huang, Jian Jiao, Xuanjing Huang, Nan Duan, and Ruofei Zhang. An enhanced knowledge injection model for commonsense generation. *arXiv preprint arXiv:2012.00366*, 2020.

Steven Feng, Jessica Huynh, Chaitanya Prasad Narisetty, Eduard Hovy, and Varun Gangal. SAPPHIRE: Approaches for enhanced concept-to-text generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pp. 212–225, Aberdeen, Scotland, UK, August 2021a. Association for Computational Linguistics. URL https://aclanthology.org/2021.inlg-1.21.

Steven Y Feng, Kevin Lu, Zhuofu Tao, Malihe Alikhani, Teruko Mitamura, Eduard Hovy, and Varun Gangal. Retrieve, caption, generate: Visual grounding for enhancing commonsense in text generation models. *arXiv preprint arXiv:2109.03892*, 2021b.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4319–4338, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.351. URL https://aclanthology.org/2020.emnlp-main.351.

Michael Wayne Goodman. Penman: An open-source library and tool for AMR graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 312–319, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.35. URL https://aclanthology.org/2020.acl-demos.35.

H Paul Grice. Logic and conversation, syntax and semantics. *Speech Acts*, 3:41–58, 1975.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1233–1239, 2016.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2019.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. URL https://arxiv.org/abs/1602.07332.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

Haonan Li, Yeyun Gong, Jian Jiao, Ruofei Zhang, Timothy Baldwin, and Nan Duan. Kfcnet: Knowledge filtering and contrastive learning network for generative commonsense reasoning. *arXiv preprint arXiv:2109.06704*, 2021.

Yikang Li, Pulkit Goel, Varsha Kuppur Rajendra, Har Simrat Singh, Jonathan Francis, Kaixin Ma, Eric Nyberg, and Alessandro Oltramari. Lexically-constrained text generation through commonsense knowledge extraction and injection. *arXiv preprint arXiv:2012.10813*, 2020.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1823–1840, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.165.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Xiao Lin and Devi Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2984–2993, 2015.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S Yu. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. *arXiv preprint arXiv:2009.12677*, 2020.

Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. Gpt-too: A language-model-first approach for amr-to-text generation. *arXiv preprint arXiv:2005.09123*, 2020.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1098. URL https://aclanthology.org/N16-1098.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *arXiv preprint arXiv:1607.00970*, 2016.

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*, pp. 311–318, 2002.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? 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)*, pp. 2463–2473, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://www.aclweb.org/anthology/D19-1250.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 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. *arXiv preprint arXiv:1910.10683*, 2019.

Lei Sha. Gradient-guided unsupervised lexically constrained text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8692–8703, 2020.

Ruth Tincoff and Peter W Jusczyk. Some beginnings of word comprehension in 6-month-olds. *Psychological science*, 10(2):172–175, 1999.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Han Wang, Yang Liu, Chenguang Zhu, Linjun Shou, Ming Gong, Yichong Xu, and Michael Zeng. Retrieval enhanced model for commonsense generation. *arXiv preprint arXiv:2105.11174*, 2021.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro Szekely, and Xiang Ren. Connecting the dots: A knowledgeable path generator for commonsense question answering. *arXiv preprint arXiv:2005.00691*, 2020.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7378–7385, 2019.

Jiaxuan You, Rex Ying, Xiang Ren, William Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *International conference on machine learning*, pp. 5708–5717. PMLR, 2018.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Fspool: Learning set representations with featurewise sort pooling. In *International Conference on Learning Representations*, 2019.

Pei Zhou, Rahul Khanna, Seyeon Lee, Bill Yuchen Lin, Daniel Ho, Jay Pujara, and Xiang Ren. Rica: Evaluating robust inference capabilities based on commonsense axioms. *arXiv preprint arXiv:2005.00782*, 2020.

A Appendix

A.1 Rules for mapping visual scene graphs to SKG

For relations that are annotated as verbs by VisualGenome, we break the relationship (subject, relation, object) into (relation, :ARG0, subject) and (relation, :ARG1, oject). For other popular relations, we conduct the following mapping:(subject, be, object)→(subject, domain, object), (subject, displace, object)→(subject, possible, object),(subject, have/of, object)→(subject, part, object),(subject, with, object)→(subject, poss, object),(subject, on/behind/at/under/along/in/..., object)→(subject, location, object)

Table 7: The most common relation types in SKG instances and their example triplets.

Relation types	Examples
ARG1	(play, ARG1, guitar)
ARG0	(play, ARG0, man)
ARG2	(ask, ARG2, girl)
Location	(play, Location, stage)
Time	(play, Time, sing)
Op1	(down, Op1, stair)
Part	(dog, Part, ear)

A.2 HUMAN EVALUATION CRITERION FOR GENERATED SKG'S

When facilitating the human evaluation study on our generated Scene Knowledge Graphs, we randomly sampled $K = 100$ samples of *concept sets* $\{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^{100}\}$ from the ROC, VIST, and CommonGen datasets. For each concept set, we provide both the relations and text that are associated with the concept set. Altogether, the concept set, relations, and text form a single line. With this in mind, the goal of the human evaluation study is to score each line by two dimensions, i.e., *CommonSense* and *Alignment*.

When measuring *CommonSense*, we ask the human evaluators whether the relation triplets in each line make commonsense to them. We note that each triplet is a relationship between two concepts in the form of (concept1 [relation] concept2). Further, the relation types for these triplets are from the Abstract Meaning Representation (**AMR**) language. For each line, the *CommonSense* score can take on a binary value of either 0 or 1, meaning the relations in the line make commonsense to the evaluator.

In addition to measuring *CommonSense*, we also measure *Alignment* between the provided relation triplets and the text in a line. For example, if the relations are "(throw[:ARG0] person), (throw [:ARG1] ball)" and the sentence is "A person throws a ball", then they are aligned. However, if the text says nothing about someone throwing a ball, then they are not aligned. Similar to *CommonSense*, we score *Alignment* with a binary metric system of 0 or 1.

We conducted this study with 3 individuals who are studying Computer Science, though not all individuals were aware of SKG's prior to the human evaluation. We aggregated the scores for each evaluator by averaging their *CommonSense* and *Alignment* scores for each dataset, and this provided us with overall average scores for *CommonSense* and *Alignment* w.r.t. our task datasets. Additionally, we computed Fleiss' Kappa Score to understand the overall agreement and disagreement of the scores across evaluators. We leveraged the *krippendorff* package in Python to compute the Fleiss' Kappa Score, and overall, we found that our evaluators fairly agreed with eachother across ROC, VIST, and CommonGen.