MMCOMET: A Holistic Multimodal Commonsense Knowledge Graph

Anonymous ACL submission

Abstract

We present **MMCOMET**, the first multimodal commonsense knowledge graph (MMKG) that integrates physical, social, and eventative knowledge. This new resource addresses a major limitation of existing MMKGs in supporting complex reasoning tasks like image captioning and storytelling. MMCOMET extends the ATOMIC2020 knowledge graph to include a visual dimension, through an efficient image retrieval process, resulting in over 900K triples. Through a standard visual storytelling experiment, we show that our holistic approach enables generating richer and more contextually aware stories.

1 Introduction

Commonsense knowledge graphs (KG) are multi-relational graphs that consist of statements about everyday concepts (e.g., *dinner party*, *coffee machine*). Each statement in the graph is usually encoded as a $<$ `head`, `relation`, `tail` $>$ relationship, where `head` is a concept, `tail` is either a concept or a natural language phrase, and `relation` is an informative label describing how the `head` and `tail` are related. Recently, KGs have been widely applied to various downstream applications, ranging from information retrieval (Guu et al., 2020) to natural language generation (NLG) tasks, such as image captioning (Zhou et al., 2019), visual question answering (Gardères et al., 2020) and storytelling (Wang et al., 2024). Applications that integrate KGs into their framework typically query an external existing KG to search for concepts related to the model's input prompt. This extra information is then injected into the model as an additional feature to help improve commonsense reasoning and contextual understanding (Chang et al., 2020; Xu et al., 2022; Sun et al., 2022; Zou et al.). For such applications, the model's output quality is directly linked to the quality of the main source of knowledge (Razniewski et al., 2024). Therefore, there is a need for the KGs to have *high expressive power*, large coverage over a variety of information and the knowledge retrieved needs to be directly aligned with the downstream task.

Figure 1: Automated visual storytelling:
Baseline: *Family members enjoyed leisurely moments together. Grandpa shared memories during the trip*;
Ours: *Family spent the day relaxing in the boat, enjoying beer. Grandpa took grandson on his lap as he drove, with grandson's parents watching nearby.*
More examples in Table 7.

Most existing prominent commonsense KGs, such as ConceptNet (Liu and Singh, 2004), focuses on hypernymy (Hertling and Paulheim, 2017) relations (e.g. $<$ `chocolate cake`, `IsA`, `cake` $>$) or physical information about a given concept (e.g. $<$ `cake`, `MadeOf`, `flour` $>$). While these relations are useful for basic information retrieval tasks, such knowledge is insufficient for more complex tasks, such as storytelling, which require reasoning about real-world social situations and activities. In image captioning, using current commonsense properties about physical objects lacks the nuanced understanding to interpret the social context. This has motivated other works to construct KGs focusing on *eventive* and *social* commonsense knowledge. One prominent example is ATOMIC2020 (Hwang et al., 2021), a KG containing commonsense information about human interactions, reactions, and desires that specific events could trigger.

Previous research, particularly on NLG tasks like

dialogue generation (Cai et al., 2023) and visual storytelling (Wang et al., 2024), that have incorporated such commonsense into their frameworks have achieved richer and more diverse outputs. However, one key limitation remains prevalent: knowledge acquisition fundamentally involves multiple modalities and is not just limited to text. This has brought attention to multimodal knowledge graphs (MMKG), with several existing MMKGs exploiting visual data to express KG concepts (Ferrada et al., 2017; Alberts et al., 2021). However, these MMKGs are again, limited to describing simple factual relationships which are not highly informative for tasks that require comprehensive social reasoning. For instance, in the story segment shown in Figure 1, our method tells a more engaging and context-aware story. Instead of just saying that the family enjoyed their time together, our story (i) specifies that they were *relaxing* in the boat and *enjoying beer*, which paints a clearer picture of the setting and activities (**physical knowledge**), (ii) includes a specific event, namely that *Grandpa taking his grandson on his lap as he drove*, which makes the scene more dynamic (**eventive knowledge**), and (iii) acknowledges the social aspect by mentioning that the *grandson's parents were watching nearby*, which hints at family relationships (**social knowledge**). By capturing these details, our MMKG enables the storytelling model to provide a fuller sense of the scene, the people involved, and their interactions.

In this work, we construct MMCOMET, the first MMKG that combines physical, eventive, and social commonsense knowledge into a single large-scale graph, covering both the natural language and visual modalities. Specifically, we build our graph by expanding ATOMIC2020, which contains 1.33M commonsense statements related to social, physical, and event-driven concepts. We enhance our KG with a rich visual dimension by retrieving highly relevant images from existing high-quality datasets (for physical concepts) and using web search (for social and eventive concepts). This results in the first holistic multi-modal commonsense knowledge graph with over 900K tuples [1].

A summary of our main contributions: **1)** We present the first multimodal commonsense knowledge graph covering social interactions, physical attributes, and eventive aspects of everyday concepts. **2)** For searching for the matching visual modality to the commonsense textual statements, we propose a novel computationally efficient approach using existing image datasets. Our approach uses approximately 60 times fewer similarity-matching calculations per sample than standard brute-force searches. **3)** We propose a new baseline model to exploit text and image modalities using social- and event-specific commonsense relations. **4)** We conduct extensive experiments using our knowledge graph on a standard multimodal downstream task, namely visual storytelling, and show the clear advantage of using MMCOMET to generate higher quality stories.

[1] MMCOMET can be found at: URL (upon acceptance).

2 Related Work

2.1 Text-only Knowledge Graphs

Basic Shared Knowledge. The aspiration to automate commonsense knowledge (CSK) has been a longstanding goal in AI (Lenat, 1995; McCarthy, 1959), aiming to equip machines with structured knowledge about everyday concepts. These structured collections, known as commonsense knowledge graphs (CSKG) or commonsense knowledge bases (CSKB), include projects like ConceptNet (Liu and Singh, 2004; Speer and Havasi, 2013; Speer et al., 2017), which leverages human crowdsourcing to collect commonsense statements across predefined relations (e.g., `IsA`, `UsedFor`, `CapableOf`). Other projects, such as WebChild (Tandon et al., 2014, 2017), Quasimodo (Romero et al., 2019; Romero and Razniewski, 2020), TupleKB (Mishra et al., 2017), and Ascent (Nguyen et al., 2022), use automated methods based on handcrafted extraction patterns or open information extraction (Niklaus et al., 2018) from large text corpora like QA forums, books, image tags, and the Web. TransOMCS (Zhang et al., 2020) applies statistical methods and neural-based learning to assess the plausibility of extracted statements. ATOMIC (Sap et al., 2019) collects inferential CSK via large-scale crowdsourcing, while ATOMIC2020 (Hwang et al., 2021) leverages generative AI (Vaswani et al., 2017). Using COMET (Bosselut et al., 2019), ATOMIC2020 generates commonsense tuples from subject-relation pairs derived from ConceptNet and ATOMIC, expanding relations to cover social, eventative, and physical categories (see Table 5). The COMET-ATOMIC2020 framework (Hwang et al., 2021) demonstrates that training smaller models like BART (Lewis et al., 2019) on high-quality

Multimodal Knowledge Graph	Domain or Focus	Size	Multimodal	Image Source
IMGpedia (Ferrada et al., 2017)	Encyclopedic	442M	Entity, Relation	WMC
ImageGraph (Oñoro-Rubio et al., 2017)	Encyclopedic	560K	Entity	WSE
MMKG (Liu et al., 2019)	Encyclopedic	814K	Entity	WSE
Richpedia (Wang et al., 2020)	Encyclopedic, Geographic	172M	Entity, Relation	WSE, WP
VisualSem (Alberts et al., 2021)	Encyclopedic, Multilingual	1.5M	Entity	WP, ImageNet
MMEKG (Ma et al., 2022)	CS (Event)	934M	Entity, Tuples	imSitu
TIVA-KG (Wang et al., 2023)	CS (Physical)	1.4M	Entity, Tuples	WSE
MMpedia (Wu et al., 2023)	Encyclopedic	19M	Entity	WSE, WP
AspectMMKG (Zhang et al., 2023)	Encyclopedic	645K	Entity	WSE, WP
MMCOMET (*ours*)	CS (Social, Physical, Eventive)	989K	Entity, Relation	WSE, OID

Table 1: An overview of existing MMKGs, compared to ours. CS: Commonsense; WSE: Web Search Engine; WP: Wikipedia; WMC: Wikimedia Commons; OID: open-source image datasets from captioning/storytelling tasks.

CSK can outperform larger LLMs, such as GPT-3 (Brown et al., 2020), in generating *plausible* commonsense statements. **Special Commonsense Knowledge.** One special commonsense project integrates data from *7 sources* to construct a hyper-relational CSKG for joint usage (Ilievski et al., 2021). Candle (Nguyen et al., 2023) focuses on scaling *cultural* commonsense knowledge, while Uncommonsense (Arnaout et al., 2022) and ANION (Jiang et al., 2021) emphasize compiling lists of salient *negative* commonsense statements.

Unlike text-only CSKGs, our multimodal graph incorporates visual knowledge, expanding the scope of commonsense understanding and enabling a wider variety of downstream tasks, such as image captioning and visual storytelling.

2.2 Multimodal Knowledge Graphs

While text-only KGs support AI applications like question answering, multimodal knowledge graphs (MMKGs) enhance understanding by integrating diverse data types, enabling richer representations and broader applications such as visual reasoning.

Encyclopedic Multimodal Knowledge. Most existing MMKGs focus on general knowledge (e.g., $<$ `France`, `HasCapital`, `Paris` $>$), featuring prominent entities from various classes like people, countries, and books. IMGpedia (Ferrada et al., 2017) uses images from Wikipedia Commons and statements from DBpedia Commons. ImageGraph (Oñoro-Rubio et al., 2017) is based on FreeBase15K (Bordes et al., 2013), combining Convolutional neural networks (CNNs) and KG embeddings (KGE) methods to answer visual-relational queries. The numerical and visual MMKG (Liu et al., 2019) links entities from FB15K, DBpedia15K, and YAGO15K (Suchanek et al., 2007), enriching them with images and numeric literals. Richpedia (Wang et al., 2020) focuses on geographic information from Wikidata, linked to descriptions and images. VisualSem (Alberts et al., 2021) emphasizes multilinguality with images and descriptions in up to 14 languages. MMpedia (Wu et al., 2023) combines textual statements from DBpedia and curated images. AspectMMKG (Zhang et al., 2023) enhances entity understanding by incorporating aspect-related images. While these MMKGs focus on documented facts, our goal is to capture everyday commonsense that humans implicitly use but rarely verbalize.

Commonsense Multimodal Knowledge. Few projects focus on mining multi-modal commonsense knowledge, with MMEKG (Ma et al., 2022) being a notable example. It integrates textual and visual information about *events* using an efficient extraction pipeline and induction strategy, organizing millions of concept events with relations like temporal, causal (e.g., `Co-occur`), and hierarchical (e.g., `SubclassOf`). It utilizes imSitu (Yatskar et al., 2016) for visual knowledge on human activities. TIVA-KG (Wang et al., 2023) covers text, audio, images, and video modalities, with text represented by word embeddings [2], video and image encoded with ResNet (He et al., 2016), and audio using VGGish (Hershey et al., 2017).

While TIVA-KG [3] focuses on physical concepts and MMEKG on event concepts; our proposed MMCOMET contains information about physical, eventive, and social concepts. This allows a holistic understanding of the nuances and subtleties of real-world social interactions. An overview comparing different existing MMKGs with ours is in Table 1.

[2] https://github.com/commonsense/conceptnet-numberbatch

[3] Note that while TIVA-KG also contains eventative and social concepts, the majority of its tuples are about the *physical* commonsense domain (more than 75%).

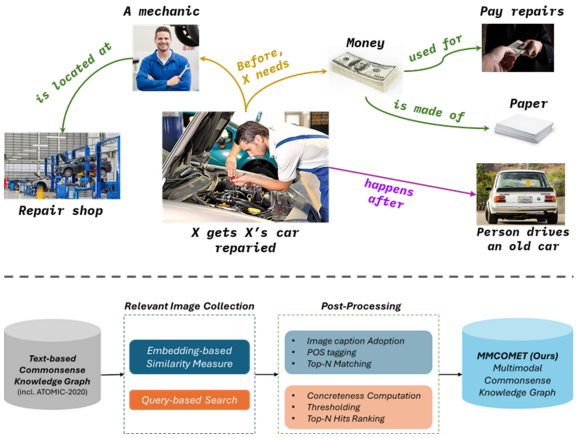

Figure 2: A tiny subset of MMCOMET(top) and its Construction Pipeline(bottom).

3 Method

3.1 MMCOMET Graph Construction

We describe the method used for constructing MMCOMET. MMCOMET's basic topology is extracted from the ATOMIC2020 (Hwang et al., 2021) commonsense knowledge graph, a 'text-only' KG covering social, physical, and eventive aspects of everyday inferential knowledge. The graph contains 23 commonsense relations of which there are 7 physical-entity relations, 9 social-interaction relations and 7 event-centred relations relating to common everyday human experiences. Descriptions and corresponding examples of each relation are presented in Table 5 (Appendix G). In total, ATOMIC2020 has ∼1.33 million tuples (in the form of $<$ `head`, `relation`, `tail` $>$), whereby a small subset of 172K commonsense reflecting qualitative human experiences was integrated from ConceptNet (Liu and Singh, 2004) via manual elimination. The remaining tuples which focus more on social-interaction and event-centered knowledge are annotated by human workers.

To incorporate visual modality into ATOMIC2020, we collect matching images for the `head` and `tail` phrase for each tuple. Two methods are proposed for automatic image collection: 1) **Similarity Matching** through comparing the `head`/`tail` phrase textual embedding with the candidate images' embeddings from existing datasets, and 2) via a **Web Search** using the `head`/`tail` phrase as the input prompt. The former approach is utilised more for physical relations (e.g. `ObjectUse`, `AtLocation`), which usually involves finding images depicting certain objects or tangible entities. Conversely, web search is more suitable for collecting images related to social and event-centered knowledge (e.g. `xWant`, `xReact`) as such knowledge is typically more abstract and requires depictions of human interactions. Therefore, we utilise the large corpora of images available on the Web to search for these cases. The overall pipeline is visualised in Figure 2 and the two methods are described as follows.

Similarity Matching (Retrieving salient physical MMCSK): In this approach, we use images from captioning and visual storytelling datasets. Firstly, we leverage CLIP (Radford et al., 2021) to convert the commonsense `head` or `tail` phrase to a textual embedding. Images in the corpus are also each converted to an image embedding with CLIP. For each commonsense phrase, we then find the top-n matching images by searching for the image embeddings that obtain the highest cosine similarity score with the commonsense phrase embedding. However, performing an exhaustive search to compute similarity scores between each commonsense embedding with *all* image embeddings is computationally inefficient and does not scale well to large image corpora.

As such, we additionally utilise the image captions and employ a part-of-speech tagger to tag nouns from the captions. This step was motivated by the idea that nouns can be seen as representative of the image's overall theme (Wang et al., 2024). Then, we collect the images corresponding to each noun tag to create a dictionary as follows: $\{Noun_1 : \{image_1, image_2, ..\}, Noun_2 : \{image_3, image_4, ...\}, ...\}$. Here, it is noted that an image can belong to multiple noun keys in the dictionary. We found 37,219 unique nouns with each noun tag containing on average 347 images. Next, the CLIP textual embeddings of the unique noun tags are computed. Finally, for the image retrieval process, we first obtain the top matching noun tags by comparing the similarity score between the commonsense phrase embedding and all noun embeddings. Then, we obtain the image subset corresponding to the matching noun tags and only search the image embeddings from this image subset to find the top 10 matching images corresponding to the `head`/`tail` phrase.

Web Search (Retrieving salient social and eventative MMCSK): For commonsense `head`/`tail` phrases that contain more abstract terms (like emotions and feelings) which are more

challenging to visually ground, we utilise web search to enhance the matching images. This is particularly useful for social and event-based commonsense. Specifically for each phrase, we extract the unique words after removing stopwords and lemmatisation. A numerical score for each lemmatised word is then obtained from a crowdsourced dataset that contains concreteness ratings for 40K most common lemmas (Brysbaert et al., 2014). Here, the concreteness score is a rating given to a word where a higher rating means that the word's concept refers to something that exists in reality which one can directly experience through their senses. Higher ratings tend to be associated with tangible nouns (e.g. '*computer*', '*flower*'), whereas more abstract terms (e.g. '*justice*', '*honour*') will have lower scores. After obtaining the concreteness for each unique word in the commonsense `head/tail`, the average of scores across all words is taken to compute a final concreteness score for each commonsense phrase. We then adopt web search to find the top 10 images from Google Images [4] for the commonsense phrases that produced a *concreteness score lower* than a pre-determined threshold of 4. This threshold was chosen as the average concreteness score of the social and event-centric relations was found to be less than this value (see Table 4). In total, ∼415K commonsense phrases required web search, approximately 72% of the total number of unique commonsense phrases in ATOMIC2020.

To obtain the most relevant results, we make slight adjustments to the commonsense phrases when web searching. Specifically, for social and event-based commonsense, we add the word '*person*' in front of the search phrase (if the phrase did not already contain that word) to ensure that the retrieved image depicted a person or people. Moreover, we find that the search engine sometimes returns images that only contain text, particularly for more abstract phrases. As Google allows filtering out images containing a specific term by adding '*-term*' to the query, we also append '*-text*' to filter out such images. For instance, searching the phrase '*person assesses the business strategy*' does not give useful results, whereas adjusting the query to '*person assesses the business strategy -text*' returns images of people in a business setting.

Lastly, we combine the web-matched images with the images found from the open source captioning and storytelling datasets. Specifically, for each commonsense `head/tail` phrase that utilised web search, we find the CLIP embeddings of the web images and filter out any that have a low similarity score with the commonsense phrase using a threshold of 0.15. Then, all the retained web images plus the images from the captioning/storytelling datasets are combined and the top 15 images were retained. Sample multimodal commonsense statements are shown in Appendix K.

4 Experimental Setup[5]

4.1 Datasets

Conceptual Captions (Sharma et al., 2018): comprises of approximately ∼3.3 million images annotated with captions. Captions and images are web-crawled and cover various topics and domains.

Visual Storytelling Dataset (Huang et al., 2016): contains ∼81K unique images obtained from Flickr. Each album contains image sequences comprising 5 photos and 5 human-written stories, and each story usually comprises one sentence per image. The unique number of stories is ∼50K. About 67% of the images also have matching human-written captions. We used the story sentences and captions to create the noun tag dictionary.

COCO Captions (Lin et al., 2014): contains ∼164K images paired with 5 human-annotated captions. We use the training and validation set as the matching captions for the test set are not in public.

Flickr30K (Young et al., 2014): has 31K images collected from Flickr, each with 5 reference captions provided by human annotators.

Concreteness Ratings 40K (Brysbaert et al., 2014): presents concreteness ratings for approximately 40K most common English lemmas/expressions, crowdsourced from over 4000 participants. Concreteness measures how much a word relates to something tangible and perceptible.

4.2 Human Evaluation: MMCOMET Quality

We conduct a human evaluation to check whether the collected images closely match the commonsense phrase. For this, 60 commonsense phrases (20 heads, 40 tails) were randomly sampled for each of the 19 relations, resulting in 1140 commonsense phrases being manually checked. The 1140 samples were divided among 3 annotators, and for

[4] `https://images.google.com/`

[5] Implementation details can be found in Appendix A

each sample, the annotator was given the top 7 images corresponding to the commonsense head/tail phrase. Out of the 7 images, they recorded how many of the images 'fully' and partially' matched the phrase. A 'full match' means that all concepts mentioned and implied by the commonsense phrase are present in the image, whereas a 'partial match' means that some important concept mentioned in the commonsense phrase is absent from the image (however, the overall theme of the image is still related to the phrase). Specific examples of fully and partially matched images are presented in Appendix D. For this study, we also ensure that 50% of the sampled commonsense utilised web search. Furthermore, we ask annotators to rate the concreteness level of the commonsense phrase as either 'Low', 'Medium' or 'High' where 'High' means that the commonsense phrase can be easily visually grounded whereas 'Low' means that the phrase contains abstract/intangible terms, thus making it harder to visualise. As each 1140 samples was presented with 7 images, the annotators analysed 7980 commonsense-image pairs altogether. After the study, we computed the proportion of samples with more than 5 out of 7 fully and partially matched images.

4.3 Downstream Task: Visual Storytelling

Method: Visual storytelling (VST) requires a model to generate a human-like and visually grounded story given a sequence of 5 images (Huang et al., 2016). We explore whether using the images from MMCOMET can improve commonsense reasoning in VST models. Specifically, we perform experiments with SCO-VIST (Wang et al., 2024), a social-interaction commonsense-enhanced VST framework that employs external knowledge relating to socially-triggered situations and reactions to generate high quality stories. The 3 stages of SCO-VIST are outlined in detail in Appendix B. For our experiments, we adopt the same approach as SCO-VIST but enhance the story graph by converting it into a '*multi-modal*' graph. In SCO-VIST, each node in the graph only has a textual modality (e.g. a caption description or commonsense phrase). We incorporate image features into the story graph by using the generated commonsense phrase to query MMCOMET to find the matching image. For simplicity, if the commonsense phrase is not found in MMCOMET, we just delete the node from the story graph. The CLIP features of the top found image is then concatenated with the textual features for the commonsense nodes in the story graph while for the caption nodes, we directly use the image features from the image sequence prompt. The story graph edge weights are then recalculated by taking the cosine similarity of adjacent nodes and the following steps in Stage 3 remain the same to obtain the storyline. We experiment decoding the story with large vision language models and feed in the MMCOMET and VIST images in addition to the textual storyline. Appendix E provides a visual depiction of the framework.

Evaluation Metrics: We follow Wang et al. (2024) and use RoViST (Wang et al., 2022), an unreferenced metric set for visual storytelling consisting of 3 scores that target visual grounding (RoViST-VG), language coherence (RoViST-C) and repetition (RoViST-NR). Moreover, we consider other unreferenced NLG metrics such as Perplexity and UNION (Guan and Huang, 2020). For analysing semantic similarity, SPICE (Anderson et al., 2016), BLEURT (Sellam et al., 2020) and MoverScore (Zhao et al., 2019) is further adopted.

5 Results

5.1 Human Evaluation Study

Table 2 displays the results of the human evaluation study. We show the percentage of commonsense cases with high concreteness that achieved at least 5 fully matched plus partial matched (FM+PM) and fully matched (FM) images across the samples using only web search (WSE), only open-source data (OID) as well as across both image sources (Both). Firstly, when considering both image sources, FM+PM was 90.7% and FM was 77.5% for all relations. While the FM score may seem relatively low, we emphasise that FM+PM which considers partially matched images is still quite high. Moreover, we note that partially matched images are still highly relevant to the commonsense phrase and may be useful when applied to a downstream task (as shown in examples of Figure 3). To assess the level of agreement between the 3 annotators, the intraclass correlation was also computed on a random sample of 50 cases. This was found to be 0.76 and 0.82 for FM+PM and FM respectively, indicating a very good level of agreement. Secondly, analysing the percentage differences between WSE and OID, we find that web search has greater matching accuracy compared to when using similarity matching with images from captioning/storytelling datasets. This is unsurprising given

	Web (WSE)		Open-source Data (OID)		Both	
Physical	**FM+PM**	**FM**	**FM+PM**	**FM**	**FM+PM**	**FM**
ObjectUse	100	73	96	76	98	74.5
AtLocation	100	92.9	100	100	100	96.4
MadeUpOf	100	100	89.3	89.3	94.1	94.1
HasProperty	92.6	92.6	88.5	84.6	90.6	88.7
CapableOf	95.8	87.5	95.8	70.8	95.8	79.2
Desires	100	95.7	100	72	100	83.3
NotDesires	100	92.6	88	72	94.2	82.7
Event	**FM+PM**	**FM**	**FM+PM**	**FM**	**FM+PM**	**FM**
isAfter	91.3	69.6	77.3	45.5	84.4	57.8
HasSubEvent	96.3	92.6	92.9	71.4	94.5	81.8
isBefore	87.5	62.5	88.9	63	88.2	62.7
HinderedBy	86.7	73.3	76.2	42.9	80.6	55.6
Causes	96.2	96.2	87.5	79.2	92	88
Reason	100	92	83.3	75	91.8	83.7
Social	**FM+PM**	**FM**	**FM+PM**	**FM**	**FM+PM**	**FM**
Need	90.5	81	85.7	71.4	88.1	76.2
Attr	92	80	89.5	78.9	90.9	79.5
Effect	95	90	61.9	42.9	78	65.9
React	90.9	86.4	61.1	55.6	77.5	72.5
Want	80.8	73.1	86.4	50	83.3	62.5
Intent	95	80	88.2	58.9	91.9	70.3
All Relations	94.5	85.2	87	69.7	90.7	77.5

Table 2: Results from the human evaluation study analysed across the relations showing the proportion of commonsense phrases successfully matched with the collected images. First column (WSE) shows cases that used web search as the image source, second columns (OID) shows cases that used the Similarity Matching approach with open-source image datasets and last column show all analyzed commonsense phrase samples. FM means 'Full Match' and PM means 'Partial Match'. Only commonsense annotated with a 'High' concreteness level are considered here (897 samples).

that the search engine is sensitive to mentions of exact terms in the commonsense phrase, resulting in images that can closely match the input query. In particular, for WSE, the FM+PM and FM rate for all relations was over 94% and 85% respectively, compared to 87% and 69% for OID. It is noted that while the matching rate for OID images is lower, the majority ($\sim$ 72%) of commonsense in MMCOMET utilises web search.

Lastly, considering the results across different relation categories, we discover that physical relations have a higher matching rate than social and event-based commonsense relations. Considering both image sources, on average, the FM+PM rate for the physical, event and social relations was 96.1%, 88.4% and 85%, respectively. Additionally, the FM rate is on average 10.5%, 17%, and 13.8% lower than the FM+PM rate respectively for physical, event and social relations. Physical relations being the easiest to match is unsurprising as these commonsense usually consist of single words which are tangible nouns (e.g. '*cat*', '*coffee*'). However, event and social commonsense tend to be longer phrases containing subtleties relating to human emotions and behaviours that are hard to visually capture. Specifically, based on FM+PM, when considering both image sources, `React` and `Effect` seem to be the most challenging with a matching rate of $\sim$77% . These relations also have low concreteness (see Table 4).

5.2 Visual Storytelling Results

Table 3 shows the results of the VST experiments. The first 2 rows are obtained from the SCO-VIST paper (Wang et al., 2024) and were chosen as baselines as they achieved the highest RoViST and R+S+B+M+U score. Here, SRL-pmi calculates the story graph weights by adopting Point Mutual Information (PMI) between adjacent nodes, and TGCN-SRL-cos utilises a TGCN (Temporal Graph Neural Network) to first extract the story graph node embeddings and their cosine similarities are then used to refine the story graph weights. For both models, BART (Lewis et al., 2019) was employed as the story decoder. The last 4 rows are our experiments, which predominately follow the same method as the original SCO-VIST framework (as described in Section 4.3). Specifically, VILT-pmi/cosine uses PMI/cosine similarity to obtain the graph weights, replaces BART with the VILT vision-language (V-L) transformer (Kim et al., 2021) and further, feeds the 5 VIST images as inputs in addition to the textual storyline. Meanwhile, the LLaVa experiments are like the VILT experiments but replace VILT with the LLaVA V-L transformer (Liu et al., 2024) and incorporate knowledge from MMCOMET. Specifically, LLaVA-cos-MMCOMET^{S} enriches the story graph with MMCOMET image embeddings and the input into the LLaVA story decoder is the storyline obtained from the multimodal story graph + 5 VIST images. And finally, LLaVA-cos-MMCOMET^{S+I} additionally feeds the MMCOMET images along with the MMCOMET enriched storyline and the 5 VIST images. Comparing the VILT experiments with the SCO-VIST experiments, we observe that using V-L models slightly improves the RoViST score to 79$\sim$80 with most of the improvement attributed to RoViST-C (coherency) and RoViST-NR (non-redundancy). Despite this, when considering the overall metric R+S+B+M+U, the VILT experiments still do not yield a higher score than the baselines with the best VILT experiment giving an overall score of 261.8, 1.6 points lower than the best baseline. However, when incorporating MMCOMET images into the story graph for storyline extraction (LLaVA-cos MMCOMET^{S}), we observe a significant improve-

Model	R-VG	R-C	R-NR	RoViST (R)	SPICE (S)	BLEURT (B)	MoverScore (M)	UNION (U)	Perplexity	R+S+B+M+U	Story L.
SRL-pmi	70.4	72.8	91.6	78.3	**11.5**	34.7	56.0	75.9	14.7	256.3	51.2
TGCN-SRL-cos	70.3	72.3	90.5	77.7	10.9	34.9	56.0	84.0	14.9	263.4	52.3
VILT-pmi	71.3	76.4	92.1	79.9	10.2	35.1	54.4	82.2	13.2	261.8	51.2
VILT-cos	70.2	76.3	94.2	80.2	9.3	34.3	52.1	83.3	12.1	259.2	50.1
LLaVA-cos-MMCOMET^{S}	73.3	78.4	95.8	82.5	9.9	35.2	54.2	82.6	**10.8**	264.4	50.8
LLaVA-cos-MMCOMET^{S+I}	**75.2**	**79.1**	**96.2**	**83.5**	10.2	**36.7**	**57.4**	**84.2**	11.2	**272.9**	53.4

Table 3: Automatic metrics and average story length (**Story L.**) for the 2 baselines (SRL-pmi and TGCN-SRL-cos) versus our experiments (bottom 4 rows). For Perplexity, a lower score is better. **R+S+B+M+U** is the sum of RoViST, SPICE, BLEURT, MoverScore and UNION. **R-** represents RoVist.

ment in all RoViST metrics, ultimately yielding an overall RoViST of 82.5, implying that better storylines can be extracted with a multimodal story graph. While SPICE, BLEURT, MoverScore and UNION remain similar with the VILT models, Perplexity outperforms significantly at 10.8. Finally, we observe the largest improvement in the LLaVA-cos-MMCOMET^{S+I} experiment which adds MMCOMET images to the input of the V-L model. In particular, RoViST-VG, C and NR improve by 1.9, 0.7 and 0.4, respectively, when compared with LLaVA-cos-MMCOMETS. BLEURT, MoverScore and UNION also outperformed the baselines and other experiments, resulting in the highest overall R+S+B+M+U of 272.9, thus illustrating the effectiveness of incorporating multimodal commonsense from MMCOMET.

5.3 Qualitative Analysis

To further illustrate the effect of using MMCOMET in visual storytelling, we show a selection of story frames in Table 7 (from Appendix J) with the outputs of the baseline model (SRL-pmi) and ours (LLaVA-cos-MMCOMET^{S+I}). The examples highlight how the storyteller model leverages MMCOMET's holistic commonsense, and produces more engaging narratives. Across all examples, the baseline provides simple descriptions lacking specificity and emotional depth. In contrast, our method generates stories with detailed context. For instance, in the *handmade goods* example, the baseline states "Merchants set up their booths", while our method enhances this by specifying that the booths were "vibrant" and filled with "handmade goods", capturing both the visual richness and the nature of the items being sold (**physical knowledge**). Similarly, in the *dad joke* example, the baseline provides the generic sentence "A father told a humorous joke". While this description is accurate, it lacks any emotional depth. Our version adds social nuance by describing the joke as "one of his typical, good-natured jokes". It recognizes the other individuals in the scene and their facial emotion, "that had everyone in stitches" (**physical and social knowledge**). In the *family gathering frame* (third example), the baseline only notes "It was a family gathering", whereas our method situates the event "in a restaurant where everyone enjoyed various delicacies". This added detail not only describes the physical setting but also suggests cultural or social practices associated with family gatherings, such as sharing meals (**eventive and social knowledge**). Another compelling example is the *birthday scenario*. The baseline provides a factual, but flat, statement: "Celebrated with a birthday cake". Conversely, our method enriches the narrative by describing a common birthday ritual: "Blowing out the candles of her birthday cake, she made a wish for happiness and adventure". This reflects an understanding of the event and the emotional aspirations tied to such moments (**eventive and emotional knowledge**). These examples demonstrate how our method integrates all types of commonsense knowledge to produce more context-rich narratives. By going beyond mere object and event recognition, our model captures the subtleties of human experiences, making the stories more engaging and authentic.

6 Conclusion

We introduced MMCOMET, the first multimodal commonsense knowledge graph that integrates physical, social, and eventive knowledge to enhance complex reasoning tasks like visual storytelling. By extending ATOMIC2020 with a visual dimension and over 900K triples, MMCOMET addresses key limitations of existing KGs and MMKGs, enabling richer, more contextually aware outputs. Experiments show significant performance improvements, with our model producing more detailed, coherent, and human-like narratives. MMCOMET sets a new foundation for multimodal reasoning, with future work aimed at expanding to additional modalities and applications.

7 Limitations

Cultural bias in visual data: While our work focuses on offering context-aware data, the image retrieval process may introduce cultural biases, affecting the diversity and fairness of the knowledge graph. For this, augmenting MMCOMET with cultural commonsense knowledge (Nguyen et al., 2023) is a much needed next step.

Contextual limitations due to KG incompleteness: Due the incompleteness of existing commonsense knowledge graphs (because of the open-world assumption (Razniewski et al., 2024)) and potentially highly nuanced or context-specific scenarios, our method might struggle with instances that require deeper world knowledge beyond the existing triples.

Ambiguity in visual data interpretation: Interpretation of visual data can be inherently subjective, with different observers potentially deriving different meanings from the same image. This poses a challenge for incorporating certain visual data into commonsense models. In scenarios where the visual context is highly complex or ambiguous, the model might produce reasoning that potentially misinterpret the underlying intent.

References

Houda Alberts, Ningyuan Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2021. Visualsem: a high-quality knowledge graph for vision and language. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*.

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

Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z Pan. 2022. Uncommonsense: Informative negative knowledge about everyday concepts. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*.

Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. 2023. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. In *Findings of the Association for Computational Linguistics*.

Ting-Yun Chang, Yang Liu, Karthik Gopalakrishnan, Behnam Hedayatnia, Pei Zhou, and Dilek Hakkani-Tur. 2020. Incorporating commonsense knowledge graph in pretrained models for social commonsense tasks. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*.

Sebastián Ferrada, Benjamin Bustos, and Aidan Hogan. 2017. Imgpedia: a linked dataset with content-based analysis of wikimedia images. In *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II 16*.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. Conceptbert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics*.

Jian Guan and Minlie Huang. 2020. Union: An unreferenced metric for evaluating open-ended story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *The 37th International Conference on Machine Learning*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore,

Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Sven Hertling and Heiko Paulheim. 2017. Webisalod: Providing hypernymy relations extracted from the web as linked open data. In *The 16th International Semantic Web Conference*.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Filip Ilievski, Pedro Szekely, and Bin Zhang. 2021. Cskg: The commonsense knowledge graph. In *The 18th International Semantic Web Conference*.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. "I'm Not Mad": Commonsense implications of negation and contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. 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*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV: 13th European Conference*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019. Mmkg: multi-modal knowledge graphs. In *The 16th European Semantic Web Conference*.

Yubo Ma, Zehao Wang, Mukai Li, Yixin Cao, Meiqi Chen, Xinze Li, Wenqi Sun, Kunquan Deng, Kun Wang, Aixin Sun, et al. 2022. Mmekg: Multi-modal event knowledge graph towards universal representation across modalities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

John McCarthy. 1959. Programs with common sense.

Bhavana Dalvi Mishra, Niket Tandon, and Peter Clark. 2017. Domain-targeted, high precision knowledge extraction. *Transactions of the Association for Computational Linguistics*.

Tuan-Phong Nguyen, Simon Razniewski, Julien Romero, and Gerhard Weikum. 2022. Refined commonsense knowledge from large-scale web contents. *IEEE Transactions on Knowledge and Data Engineering*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference*.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Daniel Oñoro-Rubio, Mathias Niepert, Alberto García-Durán, Roberto González, and Roberto J López-Sastre. 2017. Answering visual-relational queries in web-extracted knowledge graphs. *arXiv preprint*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*.

Simon Razniewski, Hiba Arnaout, Shrestha Ghosh, and Fabian Suchanek. 2024. Completeness, recall, and negation in open-world knowledge bases: A survey. *ACM Comput. Surv.*

Julien Romero and Simon Razniewski. 2020. Inside quasimodo: Exploring construction and usage of commonsense knowledge. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z. Pan, Archit Sakhadeo, and Gerhard Weikum. 2019. Commonsense properties from query logs and question answering forums. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Robert Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. *The People's Web Meets NLP: Collaboratively Constructed Language Resources*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*.

Yueqing Sun, Qi Shi, Le Qi, and Yu Zhang. 2022. JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Niket Tandon, Gerard De Melo, Fabian Suchanek, and Gerhard Weikum. 2014. Webchild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of the 7th ACM international conference on Web search and data mining*.

Niket Tandon, Gerard De Melo, and Gerhard Weikum. 2017. Webchild 2.0: Fine-grained commonsense knowledge distillation. In *Proceedings of ACL 2017, System Demonstrations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*.

Eileen Wang, Caren Han, and Josiah Poon. 2022. Rovist: Learning robust metrics for visual storytelling. In *Findings of the Association for Computational Linguistics*.

Eileen Wang, Soyeon Caren Han, and Josiah Poon. 2024. Sco-vist: Social interaction commonsense knowledge-based visual storytelling. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*.

Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. 2020. Richpedia: a large-scale, comprehensive multi-modal knowledge graph. *Big Data Research*.

Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. 2023. Tiva-kg: A multimodal knowledge graph with text, image, video and audio. In *Proceedings of the 31st ACM International Conference on Multimedia*.

Yinan Wu, Xiaowei Wu, Junwen Li, Yue Zhang, Haofen Wang, Wen Du, Zhidong He, Jingping Liu, and Tong Ruan. 2023. Mmpedia: A large-scale multi-modal knowledge graph. In *The 22nd International Semantic Web Conference*.

Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. 2022. Human parity on commonsenseqa: Augmenting self-attention with external attention. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*.

Jingdan Zhang, Jiaan Wang, Xiaodan Wang, Zhixu Li, and Yanghua Xiao. 2023. Aspectmmkg: A multimodal knowledge graph with aspect-aware entities. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Yimin Zhou, Yiwei Sun, and Vasant Honavar. 2019. Improving image captioning by leveraging knowledge graphs. In *IEEE winter conference on applications of computer vision*.

Anni Zou, Zhuosheng Zhang, and Hai Zhao. Decker: Double check with heterogeneous knowledge for commonsense fact verification. In *Findings of the Association for Computational Linguistics*.

A Implementation Details

We filter out any tuples from ATOMIC2020 where the head or tail phrase was missing. Additionally, we remove the `isFilledBy` relations, whereby this tuple typically consisted of a head phrase with a blank space (e.g. '*PersonX connects ___ together*') and the annotator was required to fill in the blank with an entity (such as '*lego*', '*wires*'). As it would be difficult to adequately find a matching image for the head phrase due to the blank space, we decided to not include this relation type in MMCOMET. After filtering, MMCOMET consists of 989K commonsense tuples with 578K unique commonsense phrases over 19 relations. We used the NLTK toolkit's part-of-speech tagger (Bird et al., 2009) to tag the nouns from the human written image descriptions, to filter out stopwords and also for lemmatising the words. The *'clip-vit-base-patch32'* version of the CLIP model (Radford et al., 2021) was used as the text and image embedder. For commonsense phrases requiring web search, we used Serper's Search Engine Scraping API [6] to scrape top 10 images from the Google search engine, enabling autocorrect. Each commonsense phrase has 10-15 images (phrases using web search will tend to have more).

B SCO-VIST Visual Storytelling Framework

The 3 stages of SCO-VIST (Wang et al., 2024) are outlined are follows: 1) Given an image sequence prompt, the framework starts by using a pretrained captioning model to generate a caption for each image which serves as a literal description of the event depicted in the image. Comet-ATOMIC2020 (Hwang et al., 2021), a language model trained on the ATOMIC2020 graph is then used to obtain additional social and event-based commonsense. Given a head/source phrase and relation (e.g. eat a cake `Intent`), Comet-ATOMIC2020 is capable of producing a tail phrase on-demand (e.g. celebrate birthday). To query Comet-ATOMIC2020, the image captions are used as the head phrase to generate several commonsense tails relating to events that could potentially occur before and after the event represented in the current caption. In Stage 2), the generated commonsense phrases and image captions are then organised in a temporal weighted story graph where each node in the graph represents a possible plot point and the edge weights represent the likelihood of transitioning from one plot point to another. Finally in Stage 3), a shortest path searching algorithm is employed to identify the optimal storyline, which is subsequently fed into an NLG model for story decoding.

[6] https://serper.dev/

C Concreteness by Relation Type

We analyze the average concreteness scores per relation type and show the results in the first subtable of Table 4. As expected, the physical relations (blue highlighted) yield higher mean concreteness scores than the social-interaction (green) and eventive (orange) relation types with the exception of `isBefore` and `isAfter`. In particular, the social relations on average produced lower scores, most likely due to these commonsense containing more abstract terms. For instance, examining the relations with the lowest concreteness scores, the `React` relations typically involve human emotion descriptions (*curious*, *honoured*) while `Attr` consists of mostly adjective mentions (*irresponsible*, *unethical*). In the middle and right sub-table, we additionally report the proportion of `head/tail` commonsense phrases per relation type that required web search based on the concreteness threshold. Generally, the trend is similar with social and eventive relations requiring more web searches as attributed to their lower average concreteness scores. However, the trend is less apparent for the tail phrases with physical relations like `HasProperty`, `NotDesires/Desires` reporting over 70%, whilst their head counterpart yielded relatively low proportions. This is perhaps unsurprising as relations like `HasProperty`, and `NotDesires/Desires` typically mention a tangible noun entity in the head phrase, resulting in higher concreteness score for the head, whereas the tail phrase will often consist of a more abstract description. For example, for the `NotDesires/Desires` relations, desires are derived from human emotions and therefore, the tail commonsense phrase may sometimes be similar in nature to social-interaction commonsense e.g. $<$ `student`, `Desires`, `ask a question` $>$. Moreover, we find that `HasProperty` which aims to describe an entity's general characteristic will sometimes mention adjectives in the tail (e.g. $<$ `panda`, `HasProperty`, `quiet` $>$, thus making it harder to visually ground.

Relation	Avg. Concreteness	Relation	% Heads Searched	Relation	% Tails Searched
React (94228)	2.95	NotDesires	2.82	AtLocation	25.52
Attr (115815)	2.98	Desires	2.85	MadeUpOf	36.47
Causes (376)	3.29	AtLocation	18.78	isAfter	48.98
Intent (49312)	3.40	MadeUpOf	20.03	isBefore	53.50
Effect (113494)	3.53	HasProperty	22.77	CapableOf	59.55
Reason (334)	3.55	CapableOf	23.98	HinderedBy	64.72
HasSubEvent (12845)	3.58	ObjectUse	25.26	ObjectUse	67.65
Want (110831)	3.58	Reason	50.90	HasProperty	71.46
Need (89391)	3.65	Intent	51.76	HasSubEvent	73.03
HinderedBy (106647)	3.80	Need	54.11	NotDesires	74.03
HasProperty (5617)	3.80	React	54.85	Need	75.94
Desires (2737)	3.88	Effect	54.85	Reason	76.35
ObjectUse (165590)	3.90	isAfter	55.96	Effect	76.54
isBefore (23208)	3.91	isBefore	56.07	Causes	77.93
NotDesires (2838)	3.91	Want	56.15	Desires	79.65
isAfter (22453)	3.94	HinderedBy	56.19	Want	80.92
CapableOf (7968)	3.96	Attr	57.13	Intent	90.35
MadeUpOf (3345)	4.08	HasSubEvent	60.19	Attr	98.82
AtLocation (20221)	4.15	Causes	65.96	React	99.48

Table 4: The left sub-table shows the average concreteness scores for each commonsense relation (sample size in brackets) sorted in ascending order. The second and third sub-table shows the proportion of heads and tail phrases across each relation that required web search sorted in ascending order. Green, orange and blue highlighting indicates social, event and physical relations respectively.

D Fully vs. Partially Matched Images

Figure 3 shows examples of what we consider a full versus partial match for an image. An image is a full match if all the actions and concepts in the commonsense phrase are presented in the image. On the other hand, it is a partial match if some aspect is missing but the overall theme of the image still matches the commonsense phrase and will therefore still be useful if applied to the downstream task. In the first ('*PersonX gets in the car*') and third example ('*PersonX eats sandwiches for lunch*'), the left image is considered a full match as the images clearly show the action of a person entering a car and eating a sandwich respectively. Conversely, the images on the right are considered partial matches as for the first example, the image on the right just shows a person in the car but the action of 'getting into the car' is missing. Likewise, for the third example, the image on the right simply depicts sandwiches but the action of 'eating' is absent. For the second ('*PersonX gives toys for Christmas'*) and fourth example ('*PersonX loves to garden*'), an important noun or concept is missing from the partially matched images. Specifically, for the second example, although the action of 'giving' is present in the image on the right, the tangible noun 'toys' is not visually captured whereas the left image of a child holding a toy bear heavily implies she had received it from another person. Similarly, the image on the right in the fourth example is missing the important entity 'person' whereas the left image is a full match as it clearly depicts a person happy in his garden.

Figure 3: Examples of a fully matched image (left) and partially matched image (right) for each 4 different commonsense phrases.

E Multimodal Commonsense Enhanced Visual Storytelling

Figure 4 depicts how SCO-VIST's storygraph is enhanced with multimodal information from MMCOMET. Textual commonsense from the original story graph is used to query MMCOMET to find the matching visual modality. The visual embedding and text embedding is conatenated to create the multimodal story graph. Following SCO-VIST, the graph edge weights are computed by calculating cosine similarity of adjacent nodes, and the storyline is then extracted from the multimodal graph via a shortest path searching algorithm. The textual storyline plus the MMCOMET and VIST images are finally fed into a V-L model (LLaVA) for story generation.

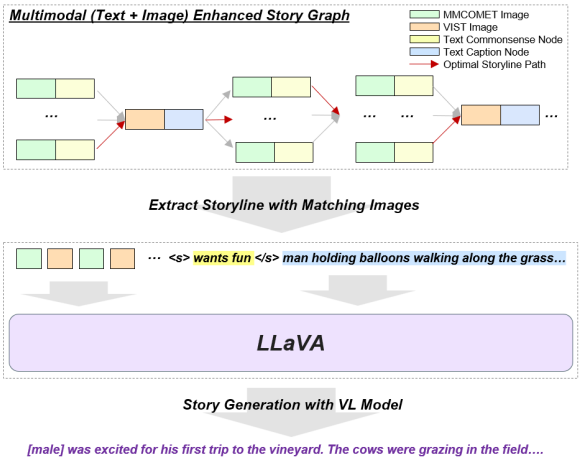

Figure 4: SCO-VIST framework enhanced with MMCOMET knowledge.

F Error Analysis

We analyze common errors that occurred in the process of finding matching images for the commonsense phrase. The 3 main types of errors and visual examples are depicted in Figure 5. One challenging issue that remains is finding matching images for non-visually grounding commonsense. For instance, the first 2 commonsense examples in Figure 5, '*rhetoric*' and '*leap year*' are very abstract terms and hence, difficult to visually ground even when employing web search. Consequently, the images obtained usually contains text or quotes as with the case for the matched images found for '*rhetoric*'. In other cases, the retrieved images may sometimes be related to movies or media where the title of the media is similar to the commonsense phrase such as with the case of '*leap year*' which seems to have retrieved images related to a movie that has the same name as the commonsense phrase.

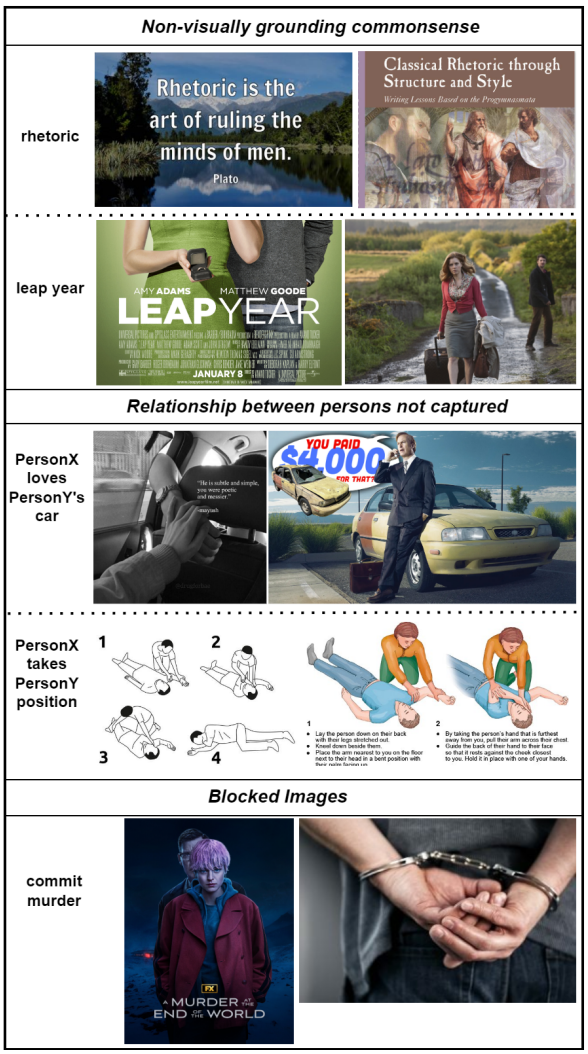

Figure 5: Examples of retrieved images for 3 different error cases found during the image collection stage.

The second issue found was that sometimes the relationship between people and concepts in the commonsense phrase is not sufficiently captured by the image. This typically occurs when the phrase mentions two people - '*PersonX*'and '*PersonY*'. For instance, the matching images found from '*PersonX loves PersonY's car*' either solely focus on the idea of 'love' as in the left image or the concept of 'car' as in the image on the right. However, the complex relationship between the first concept of a 'person loving' and the second concept, 'another person's car', is much more difficult to capture. Finally, there are cases where the commonsense phrase mentions violent or inappropriate topics. Inevitably, these images are most likely blocked by Google's web search engine such that the images

Relation	Description	Example (<head, relation, tail>)
ObjectUse	describes everyday affordances or uses of objects	bread ObjectUse make french toast
AtLocation	spatially describes the where an entity is likely to be found	bread AtLocation pantry
MadeUpOf	describes a part, portion or makeup of an entity	bread MadeUpOf dough
HasProperty	describes entities' general characteristics	bread HasProperty nice to eat
CapableOf	describe abilities and capabilities of everyday living entities	baker CapableOf coat cake with icing
Desires	desire of sentient entity	baker Desires quality ingredients
NotDesires	non-desire of sentient entity	baker NotDesires bad yeast
isAfter	events that can follow an event	X runs out of steam isAfter X exercises in the gym
HasSubEvent	provides the internal structure of an event	X runs out of steam HasSubEvent become tired
isBefore	events that can precede an event	X runs out of steam isBefore X hits the showers
HinderedBy	hindrances that obstruct the natural path of an event	X runs out of steam HinderedBy drinks too much coffee
Causes	causal relation between two events or entities	X runs out of steam Causes takes a break
xReason	provides a post-fact explanation of the cause of an event	X runs out of steam xReason did not eat breakfast
xNeed	describes a precondition for an agent to achieve the event	X runs out of steam xNeed do something tiring
xAttr	describes personas or attributes perceived by others given an event	X runs out of steam xAttr old
xEffect	actions that happen to agent X that may occur after the event	X runs out of steam xEffect drinks some water
oEffect	actions that happen to agent Y that may occur after the event	X votes for Y oEffect receives praise
xReact	emotional reactions of agent X in an event	X runs out of steam xReact tired
oReact	emotional reactions of agent Y in an event	X votes for Y xReact grateful
xWant	agent X's postcondition desires after an event	X runs out of steam xWant to get some energy
oWant	agent Y's postcondition desires after an event	X votes for Y xWant thank X
xIntent	defines the likely intent of an agent	X votes for Y xIntent to give support

Table 5: Different Relations in Comet-ATOMIC2020 and corresponding examples. Examples and definitions taken from the official Comet-ATOMIC2020 paper (Hwang et al., 2021).

obtained may not fully match the search phrase. For example, for the commonsense phrase '*commit murder*', we again obtained images related to movies/television shows as shown in the left image. At best, we tend to obtain images that are partial matches such as in the image on the right where the action is not explicitly shown but is implied.

G Comet-ATOMIC2020 Relation Types

Table 5 shows the different relations and their definitions, as well as corresponding examples in the original Comet-ATOMIC2020 graph. Green, orange and blue highlighting indicates social, event and physical relations respectively. Note that for simplicity, in this paper, we refer to xEffect and oEffect as Effect, xReact and oReact as React, and xWant and oWant as Want. The x prepended in front of the relation name is also removed for consistency e.g. xNeed is referred to as Need.

H Graph Statistics

Table 6 shows the number of unique nodes, total edges, unique web searched images, unique images acquired from open-source datasets and total unique images for each relation, and across all relations.

I Human Evaluation Information

We conducted our annotation process with three annotators: a PhD student, a Postdoctoral researcher, and an Associate Professor. These annotators were recruited based on their expertise in the relevant research to ensure high-quality annotations. Our annotators come from diverse geographical backgrounds—Asia, Australia, and Europe—reflecting a broad perspective in the annotation process. The human evaluation was conducted on a voluntary basis, and no monetary compensation was provided. Before participation, we provided clear instructions explaining the purpose of the evaluation and how the collected data would be used. The annotators were informed that their feedback would contribute to the assessment and refinement of our system's performance, and they participated with full awareness and consent.

Relation	# Nodes	# Edges	# WSE Images	# OID Images	# Total Images
ObjectUse	137,144	165,590	767,124	242,211	1,009,335
AtLocation	11,216	20,221	35,204	62,730	97,934
MadeUpOf	3,464	3,345	11,047	24,651	35,698
HasProperty	6,123	5,617	32,233	35,738	67,971
CapableOf	8,872	7,968	45,005	44,675	89,680
Desires	2,832	2,737	18,750	14,911	33,661
NotDesires	2,991	2,838	18,312	15,602	33,914
isAfter	36,181	22,453	162,312	97,386	259,698
HasSubEvent	10,670	12,845	67,522	42,541	110,063
isBefore	37,482	23,208	175,957	96,981	272,938
HinderedBy	103,844	106,647	588,983	164,975	753,958
Causes	531	376	3,299	3,471	6,770
Reason	441	334	2,622	2,888	5,510
Need	80,575	89,391	490,672	136,668	627,340
Attr	29,650	115,815	175,477	73,072	248,549
Effect	90,983	113,494	567,801	134,982	702,783
React	35,537	94,228	228,856	77,635	306,491
Want	107,268	152,989	689,798	146,314	836,112
Intent	45,771	49,312	292,494	92,549	385,043
All	571,404	989,408	3,476,098	435,695	3,911,793

Table 6: Graph statistics for MMCOMET showing total unique nodes, edges, web-searched images, open-source dataset images and total images across the different relations.

J Qualitative Anlaysis: Visual Storytelling with MMCOMET

Image	**Baseline** (SRL-pmi)	**Ours** (LLaVA-cos-MMCOMET^{S+I})
	Merchants set up their booths.	Merchants set up vibrant booths filled with **handmade goods**.
	A father told a humorous joke.	Dad then shared one of his typical, good-natured jokes that **had everyone in stitches**.
	It was a family gathering.	It was a family gathering **in a restaurant** where everyone enjoyed **various delicacies**.
	Celebrated with a birthday cake.	**Blowing out the candles** of her birthday cake, **she made a wish for happiness and adventure**.
	Seating provided a close-up perspective.	We were lucky enough to get to the **game early**, and our seats were magnificent.
	He longed for an outing on his new bicycle.	He planned to take a leisurely ride on his new bike **while enjoying the serene views of the lake**.
	We later enjoyed a meal together, savoring the wine.	We all sat together **at a long table** to enjoy the meal and wine.

Table 7: Selected story frames and their associated text to showcase the effect of MMCOMET in visual story telling.

K MMCOMET Qualitative Examples

Figure 6 and 7 show examples of several relations, each with 3 matching image examples for each of the head and tail are shown. Blue, orange and green arrows indicate physical, eventive and social interaction relations respectively. Images with red borders indicate images obtained from opensource datasets while images with blue borders indicate web searched images.

Figure 6: Examples of retrieved images for each head/tail for several relations.

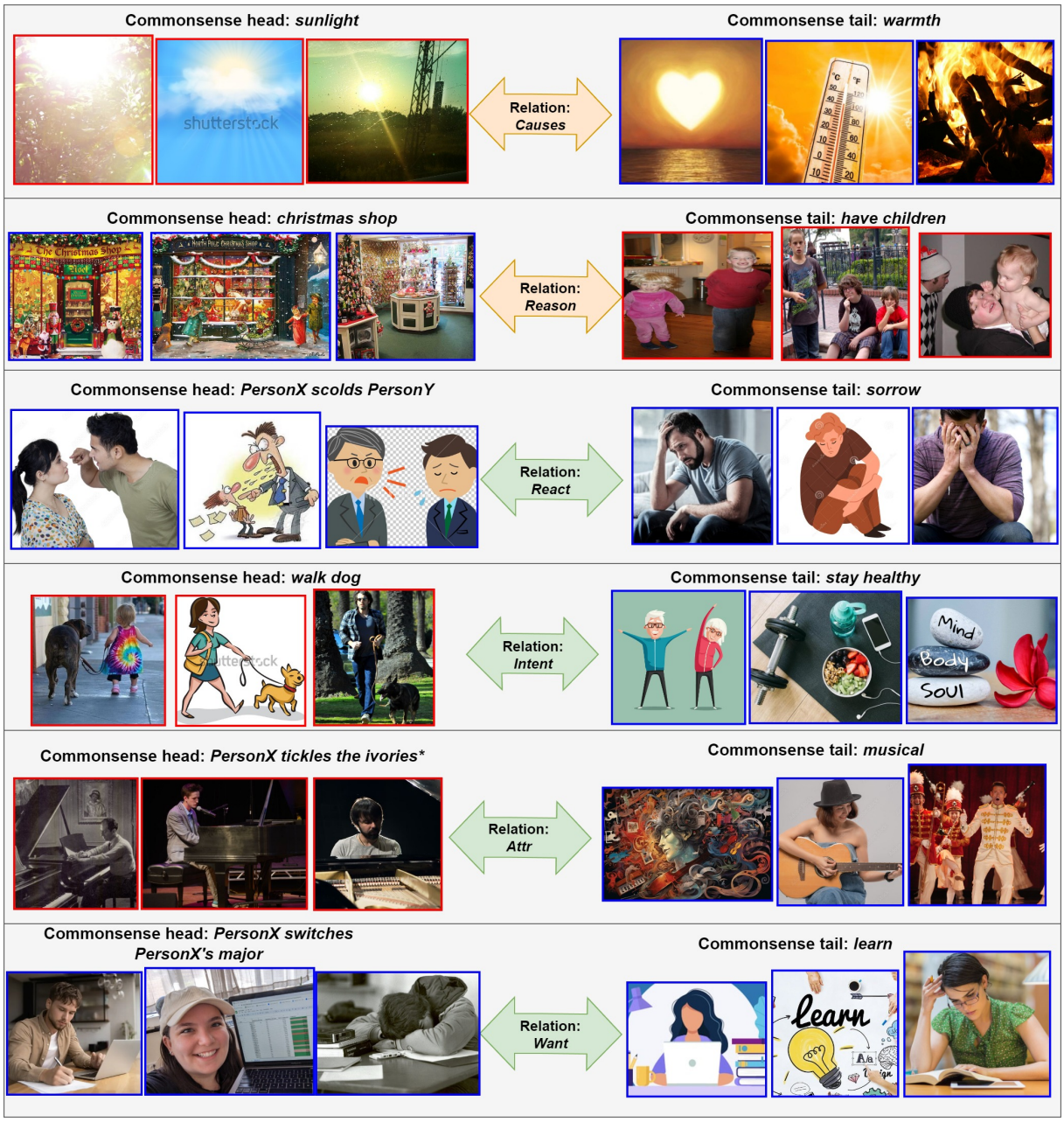

Figure 7: Examples of retrieved images for each head/tail for several relations.