ARL-TR-9106 • Nov 2020

DEVCOM
ARMY RESEARCH
LABORATORY

Maintaining Consistency and Relevancy in Multi-Image Visual Storytelling

by Aishwarya Sapkale and Stephanie M Lukin

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9106 • Nov 2020

Maintaining Consistency and Relevancy in Multi-Image Visual Storytelling

by Aishwarya Sapkale
University of Maryland, Baltimore County, Department of Computer Science

Stephanie M Lukin
Computational and Information Sciences Directorate, DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE	***Form Approved*** ***OMB No. 0704-0188***

1. REPORT DATE *(DD-MM-YYYY)* November 2020	**2. REPORT TYPE** Technical Report	**3. DATES COVERED (From - To)** June 2020-September 2020

4. TITLE AND SUBTITLE Maintaining Consistency and Relevancy in Multi-Image Visual Storytelling	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Aishwarya Sapkale and Stephanie M Lukin	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) DEVCOM Army Research Laboratory ATTN: FCDD-RLC-IT 12025 Waterfront Drive, Playa Vista, CA 90094	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TR-9106
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Contact author email: <stephanie.m.lukin.civ@mail.mil>

14. ABSTRACT

This report proposes an approach for visual storytelling across multiple images that prioritize two aspects of narrative generation: 1) the retention of *narrative consistency* between clauses in the generated story; and 2) the retention of *relevancy* between the generated story and the images from which it was derived. We take a structured approach to multi-image visual storytelling that centers around the middle image in a sequence of three. Acting as the focal point, or climax of the narrative, the plot points surrounding this are selected using events from the Atlas of Machine Commonsense (ATOMIC) corpus for if–then reasoning about daily activities, and then the selected events are subsequently grounded to the images. The result is an architecture that, given an author goal to guide the story in the form of a prompt, will generate a short narrative that retains a narrative arc and does not deviate from the content of the images.

15. SUBJECT TERMS

narratives, visual storytelling, computer vision, natural language processing, narrative theories

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Stephanie M Lukin
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	35	**19b. TELEPHONE NUMBER (Include area code)** 310-448-5396

Contents

List of Figures

List of Tables

1. Introduction

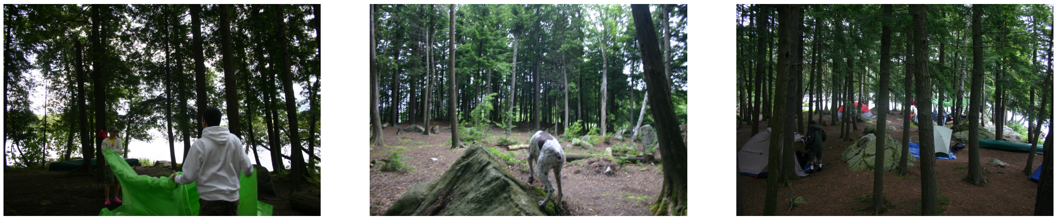
The emerging field of visual storytelling provides for the exploration of an expansive and creative space. The storyteller must determine not only what is seen in an image or sequence of images, but also what conclusions can be extrapolated from the visual snapshot with respect to the larger story world. This problem space is a strong candidate to facilitate efficient and effective communication between a human and an agent when they are not in the same space, and the agent is the one remote and out in the field observing an environment. Because the human is not at this location, they have to rely on the agent for situational awareness. Due to bandwidth constraints or information overload, the agent may be unable to directly send images or stream video; they must convey information in another way. Thus, a new problem space is formed where an agent generates mission-specific narratives over a sequence of images that it sees in that remote environment.

This potential for richness and support in human-agent teaming comes with a number of challenges, some unique to visual storytelling, and some prevalent in the traditional text form, including the assurance that the generated text is consistent and logically structured, and relevant to the larger and untold events of the story world. We focus on these particular challenges in the visual medium as we delve into the multi-image storytelling space:

1. The retention of *narrative consistency* between clauses in the generated story
2. The retention of *relevancy* between the generated story and the images from which it was derived

Stories about images are more than captions; they are a creative interpretation of the images and an extrapolation of what might be happening in the images, as well as what led to them and what might happen after them. Thus, these image-derived stories can be generated along a narrative arc following Aristotle's dramatic theory[1]: setup, rising action, climax, falling action, and resolution. Of equal importance to the structure of the story is the relevancy with respect to the images from which it was derived. The generated stories should be relevant across *all* the images, not excluded because they do not fit a predetermined plot point in the narrative. Thus, the events in the stories should be grounded to the visual information from the images.

In order to focus on narrative consistency, that is, the degree to which a story is coherent and makes sense given the characters' motivations and the logic of the storyworld itself, we utilize the Atlas of Machine Commonsense (ATOMIC) corpus for if–then reasoning.[2] This dataset contains 24,000 everyday events revolving around people, and possible causes and effects of these events. Additionally, it identifies the agents, themes, and resulting mental states for each event. In order to focus on relevancy between the images and the generated text, we obtain a simple conceptual description of the images via caption generation, and use cosine similarity to compare the semantic space of the visually grounded terms from the caption to a subset of possible events. Finally, we elicit a storyline prompt that acts as an author goal to guide the generation of the narrative. The result is a short narrative that retains a narrative arc and does not deviate from the contents of the images. Figure 1 shows a sequence of images, the storyline prompt, and the generated output of the proposed system.



(a) image$_1$ **(b) image$_2$** **(c) image$_3$**

Input Storyline	This dog is lost in the woods.
Generated Narrative	This dog goes into the woods. She is lost in the woods. She feels scared and she wants to find their way. Then, she rescued by forest rangers.

Fig. 1. Image sequence, input storyline, and generated narrative. (The incorrect pronoun resolution is discussed in Section 4).

The rest of this report is outlined as follows: Section 2 describes previous work on narrative structures and the recent approaches to visual storytelling. Section 3 describes the storytelling affordances we leverage, and how they are incorporated into our structured approach. Section 4 describes our architecture, and Section 5 is a preliminary evaluation to automatically extract storylines from a collection of texts, and our planned crowdsourcing evaluation on the generated narratives. We conclude in Section 6 with future work.

2. Related Work

Narratologists have long since made the distinction between the content of the story and its telling—between "who sees" and "who tells".[3] Named *fabula* and *sujet*, respectively, by Propp,[4] this distinction has motivated the study of how to computationally model each component separately, treating the *fabula*, or the contents of the storyworld, as "building blocks" that can be selected and rearranged to create more complex narratives,[5] and the *sujet* as something that can be adapted, altered, or tailored to the audience's needs[6–8] or the storyteller's desires.[9–12] Additionally, the story points selected from the *fabula* can be reordered or rearranged to evoke different reader responses, for example, suspense.[11,13–15] Part of the decision of what to talk about is guided by goals of the system itself.[16] Character goals drive the generation based on individual character plans and desires,[17] whereas author goals are externally defined.[18] Our narratives are guided by author goals.

Though not exclusive, there are differences between modeling the *structure* of a narrative and the modeling *representation* used to generate narratives. For the latter, various methods exist, including Transform-Recall-Adapt Methods (TRAMs),[19,20] graphical structures,[21–23] or neural networks.[24,25] This work focuses on the distinction of the former, that is, the narrative arc and shape the plot takes as the narrative unfolds. One such guide is Aristotle's dramatic theory, and Freytag's subsequent refinement[26] into the five-act dramatic structure, a pyramid consisting of exposition, rising action, climax, falling action, and resolution. Another type of narrative analysis construct is Labov and Waletzky's oral narratives,[27] in which they categorize narrative clauses that serve different functions, including action (clauses containing a causal relationship), orientation (clauses setting the background of providing traits of characters), and evaluation (clauses describing the emotional responses of characters to the events).[28] We leverage both as inspiration in our approach.

The division of *fabula* and *sujet* and the focus on narrative structure have not yet been made in visual storytelling approaches or methodologies. Emerging in recent years with the advancement of computer vision algorithms, tested approaches are more akin to caption generation or simple description, rather than a narrative along the lines of prior textual-based narrative generation. The Pix2Story framework* based on Kiros et al.[29] identifies the central theme of a single image; however, as

*https://azure.microsoft.com/en-us/blog/pix2story-neural-storyteller-which-creates-machine-generated-story-in-several-literature-genre/

the narrative unfolds, it devolves further away from the image, and the plot itself becomes tangled.[*] The Visual Storytelling (VIST) dataset considers the challenge of a sequence of images,[30] but their data collection and generated stories are single sentences that are generic, in that the sentence could apply to a number of images, and is not necessarily grounded to the presented image. Lukin et al.[31] describe a new data collection effort that attempts to overcome these hurdles by separating out and isolating the object identifications and image description task from the narrative writing task, the results of which we utilize in our evaluation.

3. Structured Approach

In the medium of visual storytelling, we treat image sequences as the visual *fabula* from which we can extract the story we wish to tell. The ATOMIC corpus serves as our world knowledge base, where facts and observations from the images can be fully understood. To guide and ground the narrative generation, we require a storyline, a single sentence that serves as a prompt or author goal for the generation. First, we describe the affordances we leverage for our structured and narratively inspired approach, and then how each piece is brought together.

3.1 Storytelling Affordances

3.1.1 Image Sequences

We use sequences of three images for our approach (exemplar sequence in Fig. 1). The images themselves come from three different sources, but are curated under an ongoing data collection effort following the pipeline described in Lukin et al.[31] In addition to the curated image sequences, the dataset consists of crowdsourced writing prompts that demonstrate the steps that humans take while coming up with a story based on a sequence of images. We further discuss this pipeline and the writing prompts in Section 5, as we utilize them in evaluating our approach.

The first image sequence source is derived from the VIST dataset.[30] The VIST task invites crowdworkers to write a narrative over a sequence of five images, which were constructed from Flickr albums. These albums were further downselected from five images to three, preserving the temporal aspect, resulting in 100 image sequences of three images each.

[*] Appendix B describes our attempt to utilize Pix2Story for this task.

The second image sequence source is created from the Visual Genome dataset.[32] These images are also extracted from Flickr; however, Visual Genome does not inherently divide the images into albums. To create image sequences from this data, the images were randomly sampled and the subsequent images sorted by Visual Genome ID and manually examined for consistency within a scene. An additional 100 sequences of three images were formed in this way.

The final image sequence source is created from the images taken over the course of the human-robot experimentations conducted in Bonial et al.[33] and Marge et al.,[34] where a participant verbally instructs a robot in a navigation task of an unknown environment. The images were randomly sampled within a participant. One hundred image sequences were formed from this data source.

3.1.2 ATOMIC Dataset

The Atlas of Machine Commonsense (ATOMIC) dataset[2] serves as our resource for plot points in the narrative. ATOMIC consists of inferential knowledge to reason about causes and effects. The dataset focuses on everyday situations resolving around people, which complements the typical image sequences from the VIST and Visual Genome data sources. The core of ATOMIC are *base events*. Their derivative causes and effects are represented by if–then relations stored in a natural language triple of the <*event1, relation, event2*> structure. For example, <*"PersonX pays PersonY a compliment", "effect", "PersonY will return the compliment"*> represents the base event of a person (X) paying another person (Y) a compliment, and that it has the effect where the other person Y will return the compliment. The base events have another representation as well, a *prefix pair*: a verb–direct object tuple extracted from the natural language text.[*] The prefix pair for the above example is <*pays, compliment*>. ATOMIC has 877,000 triples based on 24,000 base events.

ATOMIC enhances its causal relations by introducing nine inferential dimensions. Each dimension represents a type of if–then knowledge that answers questions about the base events. The relations are categorized as "causes", "effects", and "stative". The causes primarily focus on the intents and needs of the agent, forming the *xIntent* and *xNeed* inference dimensions. For instance, in Table 1, for the base event *"PersonX is lost in the woods"*, one of the annotations for dimension *xNeed* is *"PersonX goes into the woods"*.

[*] The assumed subject in prefix pairs is PersonX, as all base events revolve around PersonX.

Table 1. Examples of inference dimensions in ATOMIC. (Dimension *xIntent* is blank for this particular event. Dimensions *xAttr*, *oReact*, *oWant*, and *oWant* not utilized in this work and not shown in the table).[2]

Event	**Type of causal relation**	**Inference examples**	**Inference dim.**
"PersonX is lost in the woods"	Causes	PersonX goes into the woods –	xNeed xIntent
	Effects	PersonX will feel scared PersonX will want to find their way PersonX will be rescued by forest rangers	xReact xWant xEffect

An event might also affect the one causing it as well as others involved. The dimensions under the effects category mainly focus on the after effects (*xEffect & oEffect*), wants (*xWant & oWant*) and reactions (*xReact & oReact*) of the agent (*x*) and other (*o*) people involved in the event. For the base event in the example above, the annotation for *xReact* is *"PersonX will feel scared"*, *xWant* is *"PersonX will want to find their way"* and *xEffect* is *"PersonX will be rescued by forest rangers"*. The final category on statives are not utilized in this work, nor are the dimensions relating to PersonO; only the inference dimensions for the causes and effects on PersonX are utilized for generating our narratives.

3.1.3 Storylines

We define storylines as a single sentence conveying the central event, or climax, of the narrative. These sentences serve as the author goal and act as an input prompt to the system to help guide the direction in which to focus on. We stipulate that the storyline must include an action verb, an object, and a character with some form of agency (the ability to think or act).

3.2 Approach

The image sequences are temporal in nature, a property that we leverage in our approach. As described above, the storyline serves as the central action of the narrative, and thus we anchor the storyline to the middle image (image$_2$), and its ATOMIC causes and effects to image$_1$ and image$_3$, respectively (Fig. 2).

In Labov and Waletzky's analysis of oral narratives, we note that orientation tends to occur towards the beginning of the story, action heavily in the middle, and evaluation towards the end. Therefore, in addition to simply mapping the storyline, causes, and effects to the images, we imply that orientation-type clauses will correspond with to image$_1$, action with image$_2$, and evaluation with image$_3$. Orientation clauses are formed by utilizing *xIntent* and *xNeed* dimensions from the causes, both mapped to image$_1$. Actions are formed by considering both the storyline and the reaction of the character(s), utilizing the *xReact* dimension by mapping it to image$_2$. Finally, evaluations are formed based on the effects to show the consequences of image$_2$ by mapping the *xWant* and *xEffect* dimensions to image$_3$.

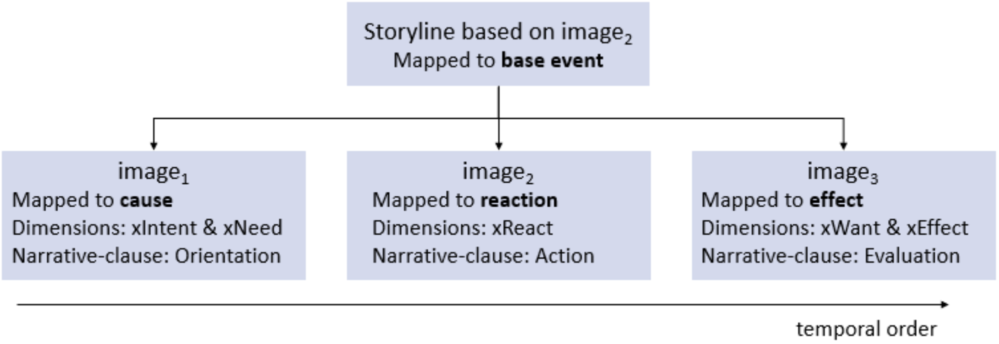


Fig. 2. Structured approach

As an example, the image sequence in Fig. 1 takes on an image$_2$-derived storyline 'This dog is lost in the woods.' This storyline is considered as the base event from ATOMIC, and its dimension inferences are retrieved. Then, we map the causes to image$_1$ by selecting an annotation from *xIntent* and *xNeed* dimensions each which best relates to image$_1$. Similarly, we map the reactions of agent (i.e., *xReact*) to image$_2$, and the effects (i.e., *xWant* and *xEffect*) to image$_3$. The result is a simple narrative with one main event and its cause and effect grounded to the image sequence. For this example, the system generates the narrative, *"This dog goes into the woods. She is lost in the woods. She feels scared and she wants to find their way. Then, she rescued by forest rangers."* The next section describes how these mappings are implemented.

4. Architecture

We implement our structured approach for narrative generation by employing a modular framework. We take inspiration from the Natural Language Generation (NLG) pipeline,[35] which is divided into content planning, sentence planning, and surface realization. Our architecture is depicted in Fig. 3. As input, we take the sequence of three images and a storyline based on the image$_2$. The storyline is fed to the **Base Event Extraction** module, which performs the first part of content planning by selecting the relevant base event from ATOMIC and gathering the relevant annotations. Next, the **Dimension Extraction** module maps the previously extracted causes and effects to the corresponding images. This is also a form of content planning. The result is a set of natural language events and dimension inference annotations extracted directly from ATOMIC.

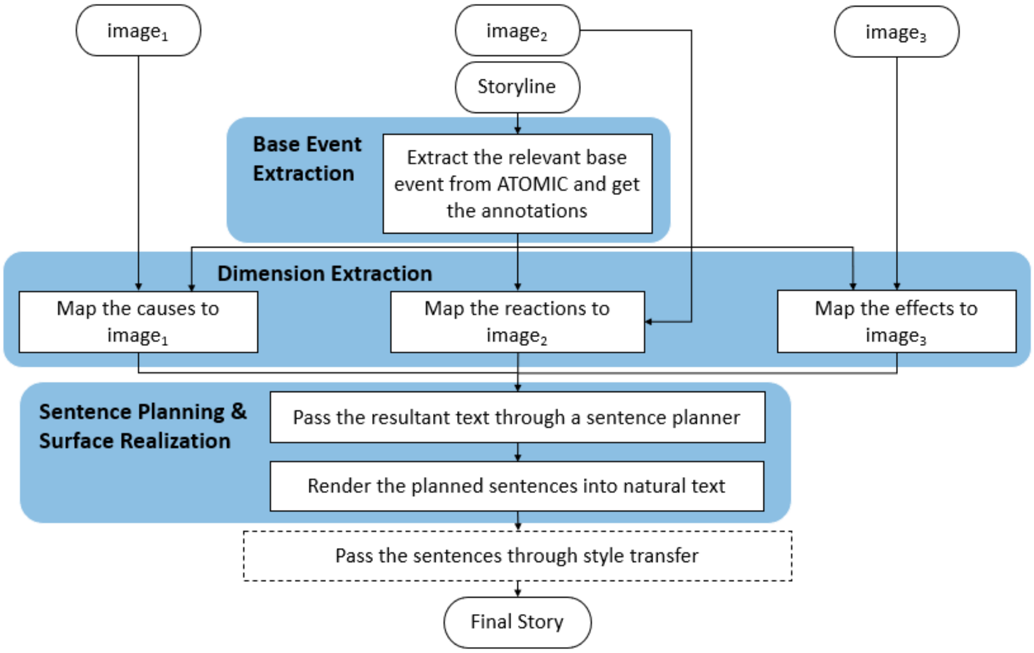

image$_1$
image$_2$
image$_3$
Storyline
Base Event Extraction
Extract the relevant base event from ATOMIC and get the annotations
Dimension Extraction
Map the causes to image$_1$
Map the reactions to image$_2$
Map the effects to image$_3$
Sentence Planning & Surface Realization
Pass the resultant text through a sentence planner
Render the planned sentences into natural text
Pass the sentences through style transfer
Final Story

Fig. 3. Structured architecture

The **Sentence Planning** module is then responsible for examining this intermediate representation, identifying the main character, reconciling pronouns, and performing aggregation with simple conjunctions. Finally, the **Surface Realization** renders the final story.

As the primary focus of this work is on the narrative planning, the natural language output is simple sentences. We envision future work that would provide more complex narratives with respect to different writing styles like comedy, suspense, and so on. We leave this for future work (dashed box in the Fig. 3).

The resultant text is a story relevant to the image sequence that retains narrative consistency. We now discuss each module in depth.

4.1 Base Event Extraction

This module is designed to achieve *narrative consistency* by extracting the relevant base event from the ATOMIC dataset and gathering the dimension inference annotations based on the input storyline. Figure 4 shows a detailed flowchart for this process.

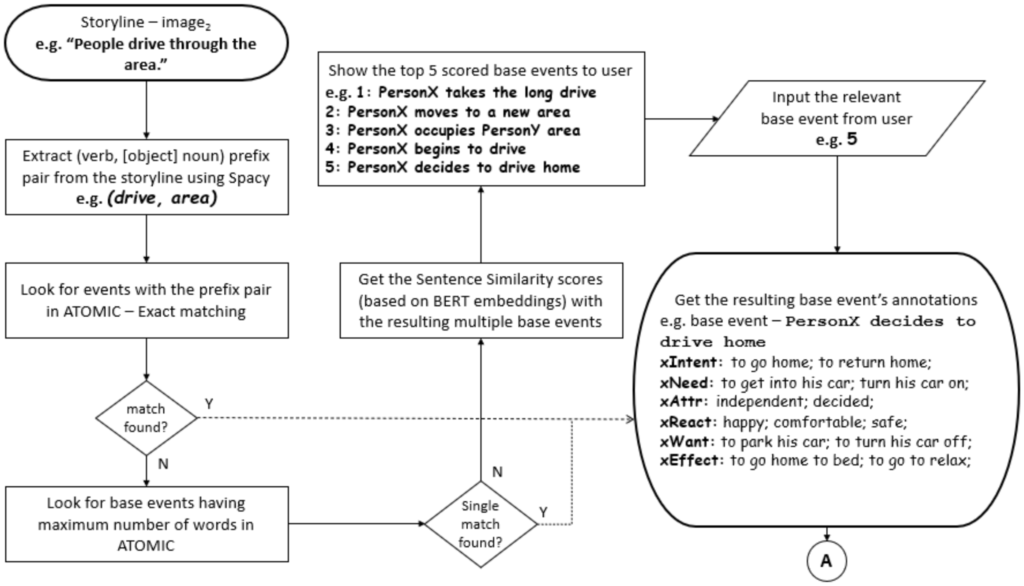

Storyline – image$_2$
e.g. "People drive through the area."
Extract (verb, [object] noun) prefix pair from the storyline using Spacy
e.g. (drive, area)
Look for events with the prefix pair in ATOMIC – Exact matching
match found?
Y
N
Look for base events having maximum number of words in ATOMIC
Single match found?
Y
N
Get the Sentence Similarity scores (based on BERT embeddings) with the resulting multiple base events
Show the top 5 scored base events to user
e.g. 1: PersonX takes the long drive
2: PersonX moves to a new area
3: PersonX occupies PersonY area
4: PersonX begins to drive
5: PersonX decides to drive home
Input the relevant base event from user
e.g. 5
Get the resulting base event's annotations
e.g. base event – PersonX decides to drive home
xIntent: to go home; to return home;
xNeed: to get into his car; turn his car on;
xAttr: independent; decided;
xReact: happy; comfortable; safe;
xWant: to park his car; to turn his car off;
xEffect: to go home to bed; to go to relax;
A

Fig. 4. Base Event Extraction module

First, the prefix pair from the input storyline is extracted using Spacy.[*] For example, the storyline 'This dog is lost in the woods' results in a prefix pair *(lost, woods)*. This prefix pair is then compared against the prefix pairs of all base events in the ATOMIC dataset. If an exact match is found, that is, the verb and object of the storyline prefix match an event prefix in ATOMIC, the corresponding event is selected as the most representative of the storyline (follow the 'Y' and the dashed line in Fig. 4). For example, an exact match is found in ATOMIC on the base event *"PersonX is lost in the woods"* with the prefix pair of *(lost, woods)*. After an exact

[*] https://spacy.io/

match, the dimension inference annotations are retrieved and passed to the Dimension Extraction module.

If an exact match on the prefix pair is not found, then we follow a two-step process. First, a down-selection is performed to find a subset of ATOMIC events with the highest number of overlapping words compared to the storyline. This may result in multiple events each having the same number of overlapping words. Then, from the selected events, a sentence similarity is performed using the cosine distance between BERT embeddings[36] of each event and the storyline[37] (solid lines in Fig. 4). More likely than not, this results in multiple matching events; to select a single base event with which to proceed, we employ either a random or interactive approach. With the random approach, the system remains fully automatic, and a base event is randomly selected from a list of the top five semantically similar base events. In the interactive mode, a user can be prompted to make the selection.

Figure 4 shows an example of when an exact match is not found. The initial down-selection is performed by comparing the storyline 'People drive through the area' against all the base events in ATOMIC. Then, the list is ranked by cosine similarity, and the top five matching events are *"PersonX takes the long drive"*, *"PersonX moves to a new area"*, *"PersonX occupies PersonY's area"*, *"PersonX begins to drive"* and *"PersonX decides to drive home"*. In the figure, the user has selected event 5. Similar as to the case of an exact match, the dimension inference annotations are retrieved from the selected event and passed to the Dimension Extraction module.*

4.2 Dimension Extraction

This module is designed to maintain *relevancy* between the generated narrative and the images. We utilize caption generation as a proxy for a brief, textual description of highlights in the image, resulting in a simple sentence describing the image in terms of objects and activities. We ground these captions to the ATOMIC dimensions for generating the causes and effects, as well as the orientation, action, and evaluation-type clauses. This process is depicted in Fig. 5.

The input to the Dimension Extraction module is the base event as determined from

*Recent advances have been made to automatically generate inference dimensions for unseen events.[38] Our work uses the human annotations in the initial exploration of this space, but the event selection component can be replaceable in the future.

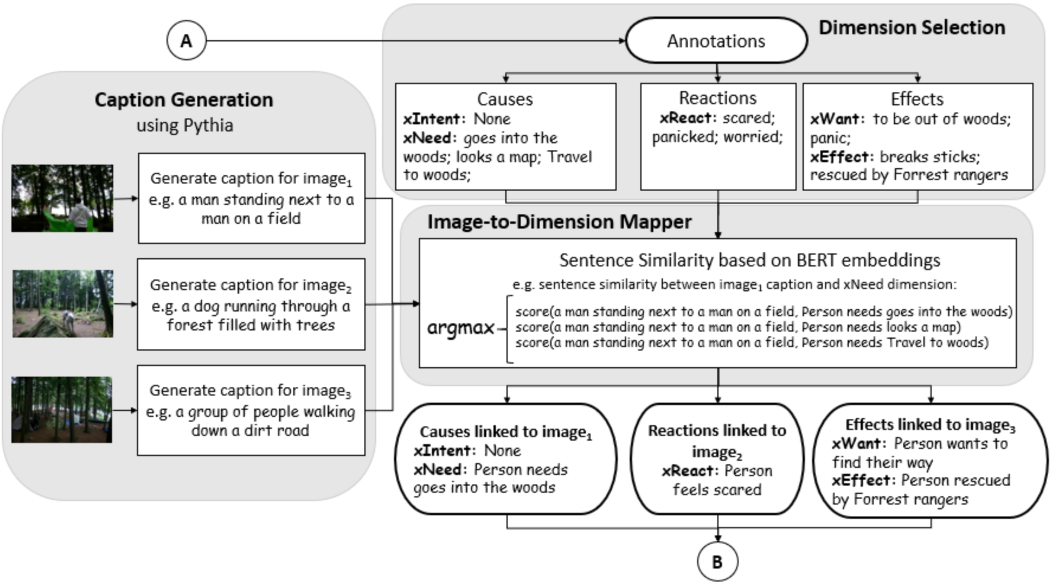

A
Annotations
Dimension Selection
Causes
xIntent: None
xNeed: goes into the woods; looks a map; Travel to woods;
Reactions
xReact: scared; panicked; worried;
Effects
xWant: to be out of woods; panic;
xEffect: breaks sticks; rescued by Forrest rangers
Caption Generation
using Pythia
Generate caption for image1
e.g. a man standing next to a man on a field
Generate caption for image2
e.g. a dog running through a forest filled with trees
Generate caption for image3
e.g. a group of people walking down a dirt road
Image-to-Dimension Mapper
Sentence Similarity based on BERT embeddings
e.g. sentence similarity between image1 caption and xNeed dimension:
argmax
score(a man standing next to a man on a field, Person needs goes into the woods)
score(a man standing next to a man on a field, Person needs looks a map)
score(a man standing next to a man on a field, Person needs Travel to woods)
Causes linked to image1
xIntent: None
xNeed: Person needs goes into the woods
Reactions linked to image2
xReact: Person feels scared
Effects linked to image3
xWant: Person wants to find their way
xEffect: Person rescued by Forrest rangers
B

Fig. 5. Dimension Extraction

the Base Event Extraction, and the associated dimension inference annotations. The first step of this module is **Dimension Selection**, a separation of the inference dimension annotations into Causes (*xIntent* & *xNeed*), Reactions (*xReact*), and Effects (*xWant* & *xEffect*). Simultaneously, captions are generated for each image; this work utilizes the Pythia caption generator.[39,40]

Next, the captions and annotations are passed to the **Image-to-Dimension Mapper**. The mapper pairs the caption and the annotations according to the predetermination that image$_1$ corresponds to causes, image$_2$ to reactions, and image$_3$ to effects. The cosine distance is computed between the BERT embeddings for the caption and their respective annotation dimension pairs. The highest scoring annotation is selected. We prepend the strings *"Person needs", "Person wants"*, and so on, to each annotation as we compare them with captions as they are complete sentences (e.g., "goes into the woods" becomes "Person needs goes into the woods"). Using complete sentences instead of the raw annotations leads to better similarity scores.

For example, the caption for image$_1$ in Fig. 1 is *"a man standing next to a man on a field"*. We pair this caption with all the annotations for *xNeed* dimension in the following way:

- ("a man standing next to a man on a field", "Person needs goes into the woods”)

- (“a man standing next to a man on a field”, “Person needs looks a map”)
- (“a man standing next to a man on a field”, “Person needs travel to woods”)

For the above example, the highest scoring pair is *“Person needs goes into the woods”*.

We follow the same procedure for all the dimensions and form a temporal order between them starting by causes (*xIntent* and *xNeed*), then affectual states (*xReact*), and finally effects (*xWant* & *xEffect*). The storyline itself is inserted between the causes and statives. Thus, the output of this module for the above example is an intermediary text plan: *“Person needs goes into the woods. This dog is lost in the woods. Person feels scared. Person wants to find their way. Person rescued by forest rangers.”* [*]

4.3 Sentence Planning and Surface Realization

Several sentence planning operations are performed on the text plans output by the Dimension Extractor: identification of the primary character in the storyline and the insertion of appropriate pronouns, and discourse planning, to include conjunctions and aggregation. Following this, the modified text plans are rendered into natural language sentences by a surface realizer.

The main character of the storyline is first identified by extracting the first noun phrase from the sentence using Spacy. For instance, for the storyline, *“This dog is lost in the woods”*, the first noun phrase is *“This dog.”* Therefore, the main character for this storyline is *“This dog.”* This identification works on plural nouns, such as extracting *“various people”* from the storyline, *“There are various people walking on a sidewalk.”* Next, the grammatical number of the character is determined (singular and plural for the above examples, respectively) using the Inflect library.[†] For determining the correct pronoun for plural nouns, the selected pronoun is *“they”*, as it is a stipulation of our storylines that the character has agency. For simular nouns, we check for common gender nouns and add *“he”* or *“she”* accordingly. If we are

[*] For this particular example, there are no annotations for *xIntent* as the ATOMIC dataset states that the people indicated that this event is not performed intentionally.

[†] https://github.com/jazzband/inflect

unable to determine a gender, for instance the noun *"dog"*, we randomly select the *"he"/"she"* pronoun.

Note the incorrect pronoun resolution of *"their"* with the *"she"* in the generated sentence *"She feels scared and wants to find their way."* The clause *"find their way"* is extracted directly from ATOMIC and concatenated with the pronoun we determined, but any pronouns in the ATOMIC clause are not resolved at this time. We leave this direct alteration of ATOMIC clauses to future work.

For refining the sentences with correct grammar and adding conjunctions, we use SimpleNLG.[41] All our sentences have the same structure with the subject as the character or pronoun, verb, and the object which is the annotation. We use this structure to form clauses in SimpleNLG. We also add the features for plural numbers and past tense as required. We also use complementary and coordinated phrases to avoid some repetition. Finally, we add a random discourse connective like *"later"* or *"then"* for the sentence that mentions the effect to conclude our narrative.

With the agentive character, gender pronoun, and discourse relations identified, the text plans can be rendered into natural language text. Note that in the Dimension Extraction module, 'Person <needs>' was already added to the text plan as part of the intermediary representation for adequate comparison to the captions. At this stage, the *"Person"* lexeme from the first sentence is replaced with the extracted character from the storyline, and all subsequent lexemes given the pronoun. Finally, SimpleNLG renders these plans, generates tense and number, and performs the discourse planning rendering. Thus, the final system output for our running example is: *"This dog goes into the woods. She is lost in the woods. She feels scared and she wants to find their way. Then, she rescued by forest rangers."*

5. Storyline Extraction and Planned Evaluation

To evaluate the validity of our structured approach, we will generate hundreds of narratives based on sequences of images. However, each image sequence requires a relevant storyline prompt. Due to time constrains, we were unable to conduct a separate crowdsourcing experiment to collect new storylines, thus we turned to an approach that automatically extracts sentences from an existing dataset that fit our storyline criteria of agency and action. This extraction and our planned evaluation are described in the next sections.

5.1 Automated Storyline Extraction

As mentioned in Section 3.1.1, the image sequences from Lukin et al.[31] have additional crowdsourced writing prompts demonstrating the process humans take as they come up with a story based on a sequence of images. There are three writing prompts: **Object Identification**, in which annotators answer the question “what is here?” for each individual image. This is not an exhaustive labeling task; instead, annotators are encouraged to identify the most important objects relevant to their story. **Single Image Inferencing**, in which annotators answer the question “what happens here?” for each image. This is most similar to a description writing task. Responses are between two and four sentences. **Multi-image Narration**, in which annotators answer the question “what has happened so far?” for images seen until the current one. Data collection is presently underway, but of the total 300 image sequences, 146 with 5 annotations each have been made available.

In order to automatically collect storylines, we examine the Single Image Inferencing writing prompt responses for image_2, as part of this task involves describing the central action about what is happening in a single image. As the storyline is expected to be a single sentence, we search for the sentences that have a character with agency performing some action. For instance, the storyline, ‘This dog is lost in the woods’ corresponds to the base event, ‘PersonX is lost in the woods’. Replacing the person variable with ‘dog’ can be done without changing the meaning of the sentence. We examine all the individual sentences from the Single Image Inferencing responses, and filter the storylines based on character agencies. This is implemented by searching for ‘person’, ‘animal’ or ‘people’ in the hypernym paths of the character in the storylines using NLTK WordNet.[*]

After filtering, we automatically identified 321 storylines with an agentive character (derived from 144 unique image sequences). Most of the storylines come from the Visual Genome (47%) and VIST (38%) datasets, which is expected as the images from these datasets are taken by people in their day-to-day lives. Despite the fact that its images do not contain people, 15% of the storylines were derived from the human-robot exploration dataset (examples in Appendix A).

Our system additionally supports an interactive mode where a user can provide the storyline as input, and non-interactive mode that uses the extracted storylines as

[*] https://www.nltk.org/howto/wordnet.html

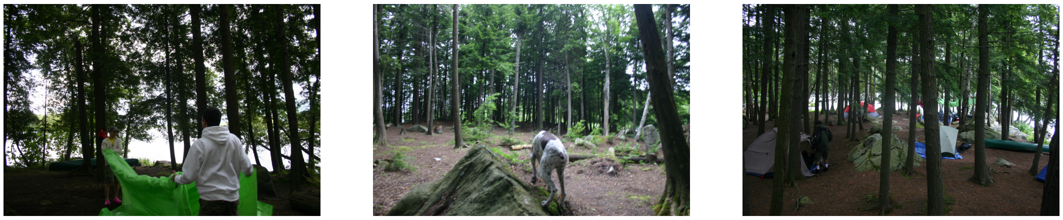
discussed above. This gives the user flexibility to either use the one provided or try out new storylines on the same image sequence.

5.2 Planned Crowdsourced Evaluation

We have begun to draft a crowdsourcing experiment for Mechanical Turk that evaluates our 321 generated narratives along several dimensions: ***relevancy***, which is one of our goals that would help identify if the story is related to the image; ***concreteness***, which focuses on what the image conveys rather than its general descriptions. Furthermore, we plan to analyze the narrative for ***consistency***, to see if the story is coherent (i.e., the sentences in the narrative form a story rather than mere descriptions of the images). Wang et al.[42] also suggest *expressiveness* as a metric, one that is certainly relevant to storytelling in general; however, expressive language generation is not the aim of our current research, so it is not applicable at present.

Our generated narratives will be compared with the stories in the VIST dataset, the closest existing dataset to these sequential visual narratives, and a simple caption baseline concatenated for each image (examples in Fig. 6). The Mechanical Turk experiment will show annotators the image sequence along with the three texts, and will be structured either as an absolute ranking or Likert rating task for relevancy, concreteness, and consistency. Absolute ranking would force the annotators to place one text above another, while the Likert scale would indicate 'how much' along the dimensions, which may or may not reveal an absolute rank. We expect our generated stories would be more relevant and concrete, as well as coherent compared to the captions that just give the content of the image lacking all the narrative features, and VIST stories that are relevant to the images, but often lack narrative consistency and concreteness.

In addition to a comparative evaluation, each story will be individually scored along a 5-point Likert scale for the same dimensions of relevancy, concreteness, and consistency. A human–authored narrative (example shown in Fig. 6) will also be scored and regarded as a top-line for this task.



(a) image$_1$ **(b) image$_2$** **(c) image$_3$**

Our Generated Narrative	This dog goes into the woods. She is lost in the woods. She feels scared and she wants to find their way. Then, she rescued by forest rangers.
VIST Narrative	People setting up the tent for the camping trip. My dog exploring the sights. The camping ground area surrounded by trees.
Caption Generation	a man standing next to a man on a field. a dog running through a forest filled with trees. a group of people walking down a dirt road.
Human Authored Narrative	As the friends fold up the tarp they look around and notice that their dog is no longer sitting next to them. They realize that once again the dog has wandered off to go on his own adventure. After they finish folding the tarp they will go after him if he hasn't find his way back. they are not worried because their dog is very competent at finding his way home. They can hear some slight rustling in the woods nearby that they know must be coming from the dog, and though they can't see him they know he is close.

Fig. 6. Comparison of our generated narrative, VIST, caption generation, and a human authored narrative

6. Conclusions and Future Work

This report presents a novel approach to multi-image visual storytelling that takes advantage of structure to generate narratives. It focuses on both narrative consistency among the subsequent clauses of the story, and relevancy with respect to the image sequence. This approach is a step towards focused and grounded narratives based on a visual *fabula* guided by author goals.

We utilize ATOMIC for its vast network of interconnected if–then events that could be followed and utilized as plot points. Additional commonsense knowledge and if–then datasets exist, such as the Situations With Adversarial Generations corpus (SWAG)[43] and the Corpus of Plausible Alternatives (COPA),[44] both multiple choice corpora where, given an event, the goal is to choose the most likely result of that action. Future work can explore the feasibility of utilizing or supplementing these resources in our proposed approach alongside ATOMIC.

As the primary focus of this work is the generation of events, the *sujet* is simple, and may not lend itself to score highly in terms of expressiveness and stylistic variation one would expect in narratives. This can be explored in future work, taking inspiration from other structured approaches for imposing stylistic variation on meaning representations[45,46] and style transfer or personalization tasks.[47–49]

The iterative element introduced in Section 5 serves as more than an evaluative aid; interactivity can function as a mechanism for narrative engagement by inviting the user to co-author the narrative together. Co-construction of personal blogs and movie scenes using TF-IDF retrieval-based approach has successfully been explored.[50,51] Our framework additionally allows for flexibility in expanding upon and shaping the narrative formed from the *fabula*. So far, we have focused on one event and one narrative-clause per image. However, the potential exists to dive deeper into each image, curating additional orientation-related events for image_1 with a blend of action to keep the reader involved, and subsequently build additional action and evaluation clauses in image_2 and conclude with further evaluation in image_3.

This work serves as a fundamental stepping stone to developing an autonomous agent that can understand what it sees, and describe it in such a way for a remote teammate to gain situational awareness. The immediate impacts of the presented approach are in maintaining the logical consistency expected in a situational report about a particular environment, and in the interpretation of the scenery in the image itself. In contrast to previous vision and language works, these narratives are not captions and do not simply list what is in a scene, nor purport that the images are isolated from one to the next. Thus, this work is one of the first to address this complex problem space with this particular scenario and end-goal in mind.

7. References

1. Aristotle (330 BC). The poetics. Dover, New York; 1997.

2. Sap M, Le Bras R, Allaway E, Bhagavatula C, Lourie N, Rashkin H, Roof B, Smith NA, Choi Y. ATOMIC: An atlas of machine commonsense for if–then reasoning. In: AAAI Conference on Artificial Intelligence; 2019.

3. Genette G. Narrative discourse: An essay in method. Cornell University Press; 1983.

4. Propp VI. Morphology of the folktale. University of Texas Press; 1968.

5. Abbott HP. The Cambridge introduction to narrative. Cambridge University Press; 2008.

6. Thorne A. The press of personality: A study of conversations between introverts and extraverts. Journal of Personality and Social Psychology; 1987.

7. Thorne A, McLean KC. Telling traumatic events in adolescence: A study of master narrative positioning. Connecting Culture and Memory: The Development of an Autobiographical Self; 2003.

8. Aylett RS, Louchart S, Dias J, Paiva A, Vala M. FearNot!–An experiment in emergent narrative. In: International Workshop on Intelligent Virtual Agents; 2005.

9. Lönneker B. Narratological knowledge for natural language generation. In: European Workshop on Natural Language Generation; 2005.

10. Montfort N. Curveship: An interactive fiction system for interactive narrating. In: Workshop on Computational Approaches to Linguistic Creativity; 2009.

11. Callaway CB, Lester JC. Narrative prose generation. Artificial Intelligence. 2002;139(2):213–252.

12. Lukin SM, Walker MA. A narrative sentence planner and structurer for domain independent, parameterizable storytelling. Dialogue & Discourse. 2019;10(1):34–86.

13. Bae BC, Young RM. Suspense? Surprise! Or how to generate stories with surprise endings by exploiting the disparity of knowledge between a story's reader and its characters. In: Joint International Conference on Interactive Digital Storytelling; 2009.

14. Ware SG, Young RM. CPOCL: A narrative planner supporting conflict. In: Artificial Intelligence and Interactive Digital Entertainment Conference; 2011.

15. Niehaus J, Young RM. A computational model of inferencing in narrative. In: AAAI Spring Symposium: Intelligent Narrative Technologies II; 2009.

16. Wardrip-Fruin N. Expressive processing: digital fictions, computer games, and software studies. 2009.

17. Meehan JR. TALE-SPIN, an interactive program that writes stories. In: International Joint Conferences on Artificial Intelligence; 1977.

18. Lebowitz M. Creating characters in a story-telling universe. Poetics. 1984;13(3):171–194.

19. Turner SR. MINSTREL: A computer model of creativity and storytelling [thesis]. University of California at Los Angeles; 1993.

20. Tearse BR. Skald: Exploring story generation and interactive storytelling by reconstructing Minstrel [thesis]. UC Santa Cruz; 2018.

21. Winer D, Young RM. Discourse-driven narrative generation with bipartite planning. In: International Natural Language Generation Conference; 2016.

22. Li B, Lee-Urban S, Riedl MO. Toward autonomous crowd-powered creation of interactive narratives. In: Intelligent Narrative Technologies Workshop; 2012.

23. Elson DK. Modeling narrative discourse [thesis]. Columbia University; 2012.

24. Fan A, Lewis M, Dauphin Y. Hierarchical neural story generation. In: Association for Computational Linguistics; 2018.

25. Wang S, Durrett G, Erk K. Narrative interpolation for generating and understanding stories. arXiv preprint arXiv:2008.07466; 2020.

26. Freytag G. Die technik des dramas. S Hirzel; 1890.

27. Labov W, Waletzky J. Narrative analysis: Oral versions of personal experience. 1997;7:3–38.

28. Swanson R, Rahimtoroghi E, Corcoran T, Walker M. Identifying narrative clause types in personal stories. In: Special Interest Group on Discourse and Dialogue; 2014.

29. Kiros R, Zhu Y, Salakhutdinov R, Zemel RS, Torralba A, Urtasun R, Fidler S. Skip-Thought vectors. arXiv preprint arXiv:1506.06726; 2015.

30. Huang TH et al. Visual storytelling. In: North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016.

31. Lukin S, Hobbs R, Voss C. A pipeline for creative visual storytelling. In: Workshop on Storytelling; 2018.

32. Krishna R et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision. 2017;123(1):32–73.

33. Bonial C et al. Laying down the yellow brick road: Development of a wizard-of-oz interface for collecting human-robot dialogue. arXiv preprint arXiv:1710.06406; 2017.

34. Marge M, Bonial C, Foots A, Hayes C, Henry C, Pollard K, Artstein R, Voss C, Traum D. Exploring variation of natural human commands to a robot in a collaborative navigation task. In: Workshop on Language Grounding for Robotics; 2017.

35. Reiter E, Dale R. Building natural language generation systems. Cambridge University Press; 2000.

36. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805; 2018.

37. Narayanan S. Semantic similarity in sentences and BERT. Medium; 2019 (accessed August 20, 2020). https://medium.com/analytics-vidhya/semantic-similarity-in-sentences-and-bert-e8d34f5a4677.

38. Bosselut A, Rashkin H, Sap M, Malaviya C, Celikyilmaz A, Choi Y. COMET: Commonsense transformers for knowledge graph construction. In: Association for Computational Linguistics; 2019.

39. Singh A, Natarajan V, Jiang Y, Chen X, Shah M, Rohrbach M, Batra D, Parikh D. Pythia-A platform for vision & language research. In: Systems Modeling Language Workshop; 2018.

40. Singh A, Natarajan V, Shah M, Jiang Y, Chen X, Batra D, Parikh D, Rohrbach M. Towards VQA models that can read. In: IEEE Conference on Computer Vision and Pattern Recognition; 2019.

41. Gatt A, Reiter E. SimpleNLG: A realisation engine for practical applications. In: European Workshop on Natural Language Generation; 2009.

42. Wang X, Chen W, Wang YF, Wang WY. No metrics are perfect: Adversarial reward learning for visual storytelling. In: Association for Computational Linguistics; 2018.

43. Zellers R, Bisk Y, Schwartz R, Choi Y. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In: Conference on Empirical Methods in Natural Language Processing; 2018.

44. Maslan N, Roemmele M, Gordon AS. One hundred challenge problems for logical formalizations of commonsense psychology. In: AAAI Spring Symposium Series; 2015.

45. Dušek O, Novikova J, Rieser V. Findings of the E2E NLG challenge. In: International Conference on Natural Language Generation; 2018.

46. Oraby S, Reed L, Tandon S, Sharath T, Lukin S, Walker M. Controlling personality-based stylistic variation with neural natural language generators. In: Special Interest Group on Discourse and Dialogue; 2018.

47. Rao S, Tetreault J. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In: North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018.

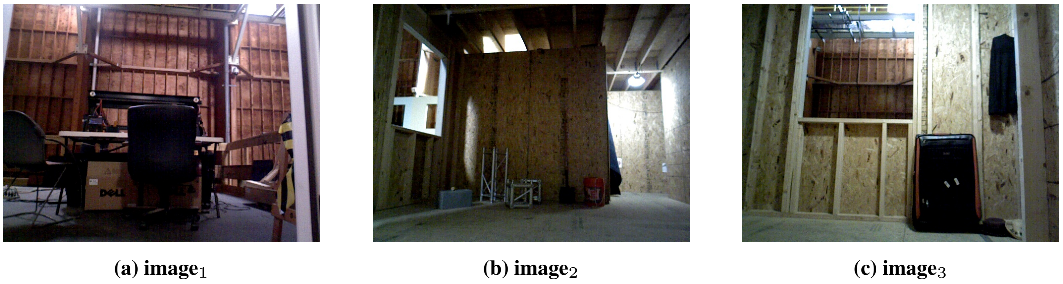
48. Mairesse F, Walker M. PERSONAGE: Personality generation for dialogue. In: Association of Computational Linguistics; 2007.

49. Rishes E, Lukin SM, Elson DK, Walker MA. Generating different story tellings from semantic representations of narrative. In: International Conference on Interactive Digital Storytelling; 2013.

50. Swanson R, Gordon AS. Say anything: A massively collaborative open domain story writing companion. In: Joint International Conference on Interactive Digital Storytelling; 2008.

51. Munishkina L, Parrish J, Walker MA. Fully-automatic interactive story design from film scripts. In: International Conference on Interactive Digital Storytelling; 2013.

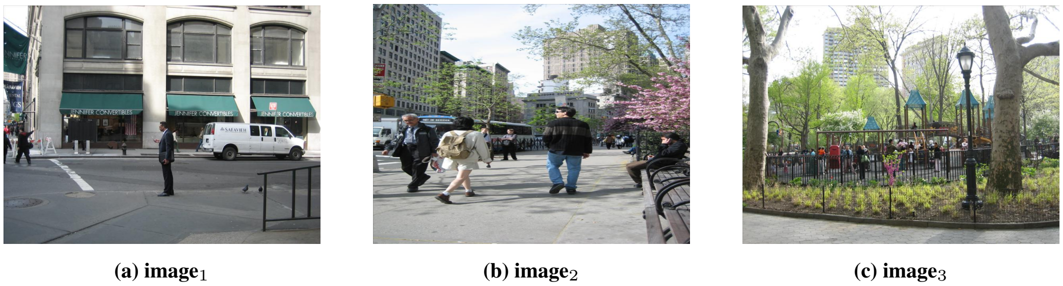
Appendix A. Dataset Examples



(a) image$_1$ **(b) image$_2$** **(c) image$_3$**

Input Storyline	Workers enter and add more things onto it.
Generated Narrative	Workers get a suitcase or bag because they wanted to get rid of persony. They enter and add more things onto it. They feel satisified and they want have the favor returned. Later, they felt sore after all of the packing.

Fig. A-1. Human-robot interaction dataset - image sequence, input storyline, and generated narrative. (The word "persony" in the first sentence is "PersonY", and is exactly as the annotator typed).



(a) image$_1$ **(b) image$_2$** **(c) image$_3$**

Input Storyline	There are various people walking on a sidewalk.
Generated Narrative	Various people need to leave his house because they wanted to get somewhere. There are they walking on a sidewalk. They feel comfortable and they want to get to a destination. Then, they walked into a cool looking store.

Fig. A-2. Visual Genome dataset - image sequence, input storyline, and generated narrative

Appendix B. Initial Approach

Our initial approach to multi-image visual storytelling was to leverage Microsoft (MS) Pix2Story. This is a neural model that generates a narrative based on a single image. We planned on generating individual stories for each image and then linking them together to form a creative story.

Pix2Story first generates a caption for the image using the Visual Semantic Embeddings trained on the MS Common Objects in Context (COCO) dataset of images and captions. Then it reconstructs the sentences around the caption using the continuity from the text passages it has been trained on, using Skip-Thought vectors. The Skip-Thought vectors are generally trained on a large number of passages; the passages could be from books, novels, movie scripts, short stories, and so on. Finally, it applies Style Shifting, which changes the caption style of the narrative into more story-like sentences.

We trained the Skip-Thought vectors on the Writing Prompts dataset. The Writing Prompts dataset is a collection of 300,000 human-written stories inspired on prompts from the Reddit's WritingPrompts forum. We selected this dataset as it could provide us with some form of context for generating the story that would potentially help when linking the three individual stories. The average length of stories is about 784 words. To form considerable length passages to train the Skip-Thoughts model we divided the stories into passages containing five sentences each.

Table B-1 shows the top five nearest captions and the generated story for image$_1$.[*] We observed that the captions have the objects from the images, but are not entirely relevant to the image. One possibility to overcome this was to replace the Visual Semantic Embedding module with a different, better performing caption generator. The generated language is poor; however, the narrative does contain the term 'forest' in it which shows that it does somewhat relate to the image and it does generate valid sentences.

We summarize the limitations as follows:

- The model requires a large amount of data for training. The Skip-Thoughts pretrained model on romance novels was trained on 14 million passages, whereas the WritingPrompts dataset only had 300,000 passages.

[*] The model was trained for 5 days on the available resources.

Table B-1. Captions and generated story using Pix2Story. (The spaces before and after punctuation marks are a byproduct of the generator).

Nearest Captions	A little girl is playing outdoors neat a tree . A young woman carrying a suitcase through the forest . A man and a woman standing near a tree . A white tent in the middle of a wooded area . A girl sits in a suitcase in the woods .
Generated Narrative	A few minutes later , I was on the ground , and I was in a forest . I was in a forest , and I was in a forest . I was in a forest , and I was in a forest . I was in a forest , and I was in a forest . I was in a forest , and I was in a forest . I was in a forest , and I was in a forest . I was in a forest , and I was in a forest . I was in a forest , and I was in a forest . I was in a forest , and I was in a forest . I was in a forest

- As the stories would be generated individually, linking would have been challenging. We do not know if we could connect three independent passages without taking away the essential actions or events that connect the stories to the images.

- The training configurations were not clearly specified in the documentation. The number of epochs to run to obtain reasonable results was undefined. Training the Skip-Thoughts encoder for 300,000 stories took 4 days for one epoch. Training Skip-Thoughts decoder took 1 day for completing one epoch.

Thus, we moved towards ATOMIC and our structured approach to overcome the linking limitation, as it provides us with the possible before and after events of actions in the images.

List of Symbols, Abbreviations, and Acronyms

ATOMIC –	Atlas of Machine Commonsense
BERT –	Bidirectional Encoder Representations
COCO –	Common Objects in Context
COPA –	Corpus of Plausible Alternatives
NLG –	Natural Language Generation
SWAG –	Situations With Adversarial Generations
TRAM –	Transform-Recall-Adapt Methods
VIST –	Visual Storytelling

1	DEFENSE TECHNICAL
(PDF)	INFORMATION CTR
	DTIC OCA
1	DEVCOM ARL
(PDF)	FCDD RLD DCI
	TECH LIB
1	DEVCOM ARL
(PDF)	FCDD RLD IT
	S M LUKIN
1	UNIVERSITY MARYLAND
(PDF)	A SAPKALE