SEE&TELL: Controllable Narrative Generation from Images

Anonymous submission

Abstract

We propose a visual storytelling framework that separates the act of 'seeing,' or observing the imagery, from 'telling,' or planning and explaining what is happening or what might happen next. We validated our framework by implementing a suite of computer vision and language generation algorithms that construct a holistic representation of the contents of an image. From this and a separate input that guides the story's direction, our system generated narrative plans using GPT-3's question-answering capabilities, and then realized those plans as expressive texts using a few-shot learning approach with GPT-3. We analyzed 64 generated stories by measuring the preservation, loss, and gain of visual information, and the willingness of a reader to take action to read more.

Introduction

There are many different ways to tell a story. By changing the narrating character or altering the tone, different tellings of the same underlying sequence of events can be realized in a variety of ways (Lönneker 2005). Key in these computational approaches to storytelling is the ability to distinguish between the story contents, i.e., all the known characters, locations, and possible events in the story world, and the story telling, i.e., the selection and presentation of information that reflect the desired state of the final story. We posit that this distinction between the story contents and story telling (also known as the *fabula* and *sujet* respectively (Propp 1968)) which has successfully been applied to textual storytelling, can similarly extend to visual storytelling. By reimaginging the *fabula* as a portfolio of story knowledge from the visual affordances in imagery, different events can be derived, selected, and generated as an expressive textual telling of a story.

We propose a framework called SEE&TELL that separates the act of 'seeing,' or observing the imagery, and 'telling,' or planning and explaining what is happening or what might happen next. Our implementation used a suite of computer vision and natural language generation algorithms that constructed a holistic and computational representation of the contents of an image that were seen. From this and a separate input that guides the story's direction (called a stage direction; examples shown in Table 1), our system constructs a narrative plan using GPT-3's question-answering capabilities (generated narrative plans shown in Table 1), and finally embellishes the plan into an expressive text using GPT-3 and a few-shot learning approach (generated expressive texts in Table 1). The top story features a crowd enjoying the parade, a rather literal interpretation of the image, while the bottom

Stage direction: People are lining the streets to see the parade.
Generated Narrative Plan: The goal of the people is to see the parade. The people will achieve their goal by the people will gather on the street and line up to see the parade. But the people could not achieve their goal if the weather was bad and the parade was cancelled. The weather must be good and the parade must go on.
Generated Expressive Text: The big day had finally arrived! The people gathered in front of the supermarket to see the parade. They all stood in line, chatting and laughing as they waited. If it rained or if there was a flood, they'd miss the parade!

Stage direction: A thief is observing the scene prior to a planned robbery.
Generated Narrative Plan: The goal of the thief is to rob the group of people in the background. The thief will achieve their goal by the thief will steal one of the umbrellas. But if the thief does not have the patience to wait for the right moment. The group of people in the background must move towards the thief.
Generated Expressive Text: The thief looked in from afar. It was time for the thief to strike. The group was walking down the street, ready for a fun night on the town. The thief had the perfect mission: grab one of their umbrellas, and run away. A fool-proof plan. However, this was a group of people who knew each other well. They knew how to fend for themselves in dire situations. If necessary, they could work together to save their belongings from harm.

Table 1: Two different stage directions (manual input) for a single image, showing different generated narrative plans and expressive texts from our system.

story imagines this as a potential venue for a robbery, with the thief targeting the umbrellas in the crowd.

Following prior definitions of creativity as a solution or idea that has significant novelty and practical usefulness (Wallas 1926; Turner 1993), we posit that creativity in visual storytelling can be measured in part by the novelty that the story presents with respect to the image contents, as well as its impact to the reader. We develop two scales that attempt to remove the variability in interpreting Likert scales which have traditionally been used to measure visual storytelling texts.

The paper contributions are as follows: 1) we propose a narrative-inspired framework for visual storytelling; 2) we implement this framework using computer vision, natural language generation, and large language models; and 3) we conduct annotation and analysis of creativity in visual storytelling, as measured by visual novelty and usefulness.

Related Work

Prior works in visual storytelling have trained neural models to generate stories from human example, relying on large, crowdsourced datasets, such as VIST (Huang et al. 2016; Yao et al. 2019). To influence the direction of the generation, some models take as input a title, genre, or prompt (Yao et al. 2019; Jana 2019; Ji et al. 2022), a mechanism that commonly appears in non-image neural-based story generation as well (Fan, Lewis, and Dauphin 2018; Wang, Durrett, and Erk 2020). These neural approaches to visual storytelling, however, are not based in formal theories that model plot, whereas prior works in non-visual, non-neural storytelling have been successful in generating stories as the progression, fulfillment, and failure of character goals and desires (Pérez y Pérez and Sharples 2001; Turner 1993; Meehan 1977; Riedl and Young 2005). The lack of plot model in visual storytelling may lead the stories to be boring, erratic, or irrelevant to the image itself. No such theory for visual storytelling exists to our knowledge, however, recent work has been conducted with neural models on non-image storytelling that attempt to keep the generator 'on track' by issuing a series of questions that together form a coherent plot (Castricato et al. 2021). The feasibility of using large language models for commonsense reasoning has also been explored, showing how large language models can both answer questions about direct facts of a new articles, as well as imagine the possibilities if something went differently (Summers-Stay, Bonial, and Voss 2021). Our work leverages the use of purposeful and sequential question-answering approaches for visual storytelling, and is the first visual storytelling framework to our knowledge that combines the advances of neural models with a narrative framework to generate plot and expressive story tellings.

Evaluation of stories generated about images commonly involves Likert scales for creativity, relevancy, and interest, e.g., (Huang et al. 2016; Wang et al. 2018), as metrics such as BLEU, which require a gold-standard and discourage diversity, are not appropriate. However Likert scales are not without their own shortcomings, as pointed out by Sai et al. and Howcroft et al., who comment on the uninterpretability of human judgments (Sai, Mohankumar, and Khapra 2020), and who state that "...the difference between 'very disfluent' & 'disfluent' is [not] the same as the distance between 'slightly disfluent' & 'slightly fluent' on a 6-point scale" (Howcroft and Rieser 2021) (pg. 1). Hu et al. (2020) create separate visual storytelling definitions for relevancy, coherence, and expressiveness, however their definitions are intrinsic to their implementation. In our study to follow, we propose new metrics that follow the advice of Sai et al. and Howcroft et al. to link scores to a concrete definition and ensure that raters do not have different interpretations of the numbers on the scale. We design these annotations to be applicable for any visual storytelling system evaluation.

SEE&TELL Visual Storytelling Framework

Our visual storytelling framework called SEE&TELL separates the act of 'seeing,' or observing the environment, from 'telling,' or explaining what is happening or what might happen next. This division is inspired by the differentiation of *fabula* and *sujet*, and is depicted in Figure 1. First, to 'see', we propose that a holistic, computational representation of a SCENE can be formulated by enumerating over the entities and attributes, i.e., visual information, that appear in an image. Entities, or THINGS, include identification of physical objects, their relative spatial locations, and traits such as color or material. Attributes, or the MOOD, assess the location and intangible attributes, such as the lighting, the weather, and activities taking place.

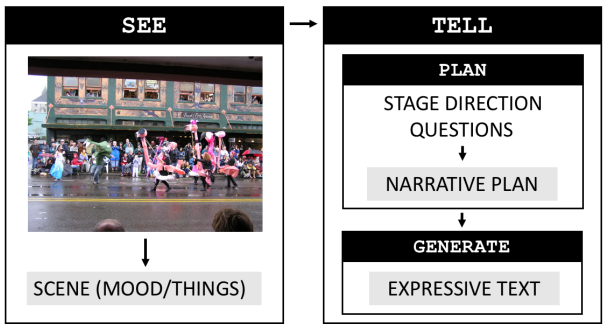

Figure 1: SEE&TELL visual storytelling framework.

The process for 'telling' is twofold, involving planning and generating. To plan, we require a particular story focus. For this, we introduce a stage direction to establish the starting point and main characters. Our usage of the term 'stage direction' is from theater, referring to an "instruction written into the script of a play, indicating stage actions, movements of performers, or production requirements" (Stage Direction. 2018)[1]. Our framework then plans by asking and answering questions that interrogate the information 'seen' and provided in the stage direction. These questions probe at the characters' motivations and challenges in accomplishing

[1]Many prior visual and non-visual storytelling works use the term 'prompt,' which serves to prompt the system in its generation. However, with 'prompt' now used as a standard in large language models, we introduce a term more closely tied to storytelling.

them, and in sequence, can follow different dramatic structures, e.g., Aristotelian structure (Aristotle (330 BC) 1997). The answers to these questions form a narrative plan: all the information regarding the plot with respect to the image. The final stage of 'telling' transforms the outline of the narrative plan into an expressive text, the realization of a story telling. This 'tell' module is reminiscent of modular natural language generation systems (Reiter and Dale 2000).

Implementing SEE&TELL

We implemented our proposed framework using a suite of algorithms, manual stage directions, and questions, which all have the ability to be interchanged. Our selected components are described below, and examples shown in Figure 2.

THINGS. We defined THINGS as entity masks, labels, center of mass, depth, and color.

- Entity detector: Detectron2 provided masks, labels, and center of mass (Wu et al. 2019). We used the Large Vocabulary Instance Segmentation (LVIS) model, trained on more than 1,200 instances representative of the 'long tail' of entities (Gupta, Dollar, and Girshick 2019).
- Depth Estimator: K-means clustering was performed over the masks' depths as estimated by DiverseDepth to cluster the entities into the foreground and the background (Yin et al. 2020).
- Color Estimator: Color Thief averaged the pixels within a mask (Dhakar 2020), following Eum et al.'s process (Eum, Han, and Briggs 2020). The resulting HEX value was mapped to a language description from webcolors CSS3_HEX_TO_NAMES.[2]

MOOD. We defined the MOOD as the location, scene attributes, and a single sentence summary.

- Location Predictor: Places365 predicted the location pictured in an image (Zhou et al. 2017).
- Non-Tangible Attribute Predictor: The Scene Understanding (SUN) predictor identified non-tangible attributes, such as function, action, and ambient lighting (Patterson and Hays 2012; Patterson et al. 2014; Xiao et al. 2016, 2010).
- Summarizer: MMF generated a one-sentence caption describing the image (Singh et al. 2020).

SCENE. We generated a SCENE by compiling the extracted visual information from the THINGS and MOOD into natural language. We used PySimpleNLG[3], a python port of SimpleNLG (Gatt and Reiter 2009) to create the following:

> This is [LOCATION] with [LIGHTING]. It is an [SPATIAL ATTRIBUTES] where [ACTIVITIES] can take place here. Here we see [LIST ENTITIES IN THE FOREGROUND AND THEIR COLOR]. In the background is/are [LIST ENTITIES IN THE BACKGROUND AND THEIR COLOR]. [SUMMARY].

STAGE DIRECTION. We designed and tested eight different stage directions. One type we call 'spontaneous,' in that it is image-dependent and drawn from impromptu human description of the image from crowdsourced data. The other seven were fixed, and inspired by non-image story generation works that utilize writing prompts of some kind. The fixed are listed below, as well as an example of a spontaneous type. In parentheses is a key phrase which is later used to refer each type.

- A secret agent must take down a terrorist mastermind. (Riedl and Young 2005) (agent)
- A Lady of the Court named Jennifer is in love with a knight named Grunfeld. (Turner 1993) (lady)
- The Mage, the Warrior, and the Priest go on an adventure. (Fan, Lewis, and Dauphin 2018) (mage)
- A princess is in love with a man whose values clash with her own. (Pérez y Pérez and Sharples 2001) (princess)
- A robot is looking for an evacuation route after a natural disaster. (Lukin, Hobbs, and Voss 2018) (robot)
- A search and rescue team is looking for a missing person. (Lukin, Hobbs, and Voss 2018) (search)
- A thief is observing the scene prior to a planned robbery. (thief)
- People are lining the streets to see the parade. (Lukin, Hobbs, and Voss 2018) (spontaneous)

QUESTIONS. We designed and tested four questions to serve as the minimum number of points that would conform to Aristotelian 3-act structure: exposition, rising action, climax, falling action, denouement (the final two elements are combined into a single question). Each question explicitly included the subject(s) mentioned from the stage direction. The questions progressively build upon the results of the prior.

1. What is the goal of [THE SUBJECT]?
2. What will [THE SUBJECT] do to achieve their goal?
3. What could prevent [THE SUBJECT] from achieving their goal?
4. What must happen next in order for [THE SUBJECT] to achieve their goal?

These questions differ from prior visual question-answering (VQA) tasks where questions can be answered directly from the image, e.g., counting objects or identifying object co-locations (Antol et al. 2015); our questions require commonsense and external knowledge not found in the images. Our framework can inherently control the narrative's complexity and length by adding more questions without compromising the computational load.

NARRATIVE PLAN. We generated narrative plans by sequentially answering the questions, building upon the prior answers. We chain prompts for GPT-3[4], where each prompt contains SCENE, stage direction, question, and prior answers. The complete narrative plan is the concatenation of each answer, resulting in a four-sentence text. Below is the GPT-3 template with variables in square brackets, and a "Q/A" prompt, following the prefix prompt approach

[2] https://pypi.org/project/webcolors/1.3/

[3] https://github.com/bjascob/pySimpleNLG

[4] We used the davinci model, temperature of 0.5, and top p of 1.

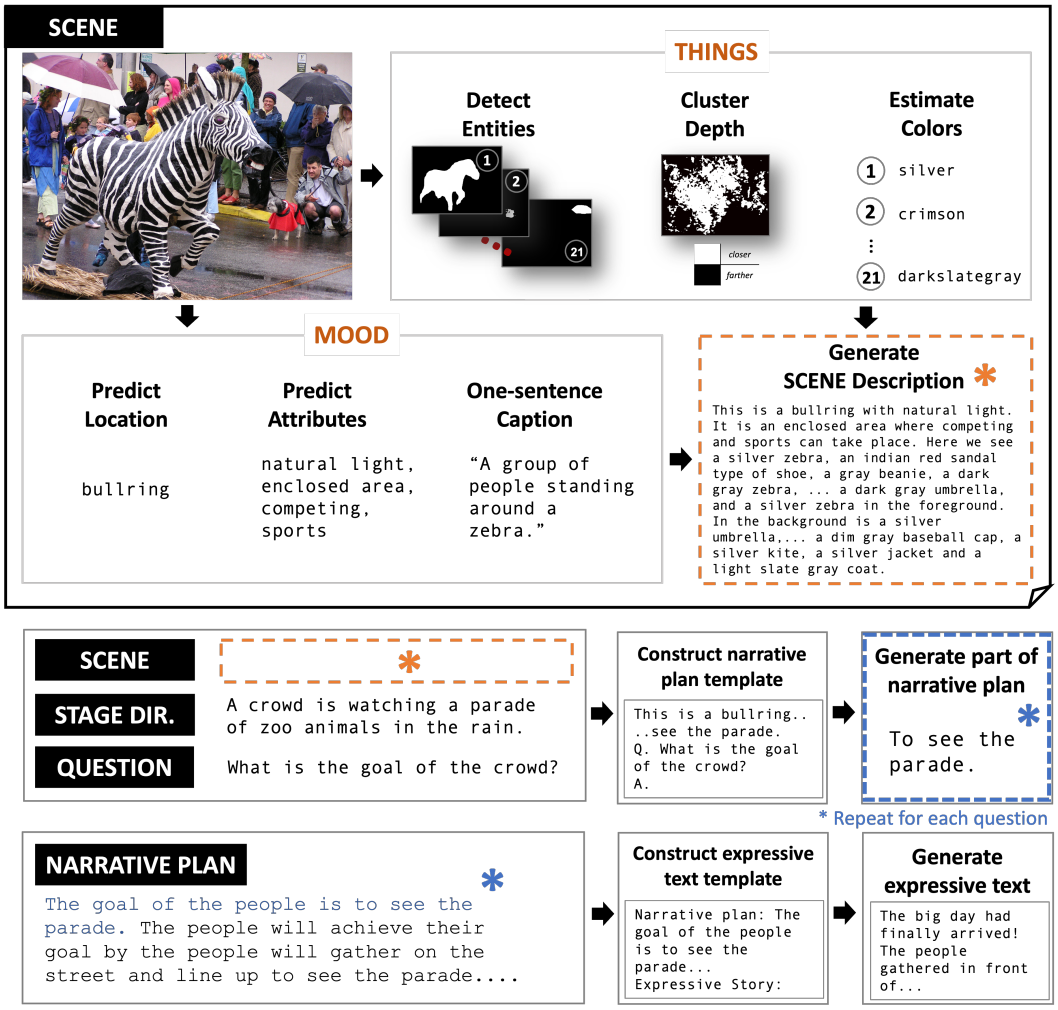

Figure 2: **SEE&TELL implementation with examples.** System output from the THINGS and the MOOD components are fed into the SCENE description generator. The compositional SCENE is paired with a stage direction and question to generate one part of the narrative plan. This process is repeated for each question. The final narrative plan is assembled and given to GPT-3 using a few-shot learning paradigm to generate an expressive text of that plan.

described in Liu et al. (2021):

> Read the description below and answer the question based on what you read.
>
> [SCENE]. [STAGE DIRECTION]. [PRIOR NARRATIVE PLANS]
>
> Q. [QUESTION].
>
> A.

By chaining prompts, GPT-3 answers only one question at a time, allowing it to better focus on the requested generation and mitigate deviations. Below is the GPT-3 prompt filled in with the SCENE, the stage direction (underlined), the chained narrative plan (i.e., the answers to the previous questions, in *italics*), and the GPT-3 answer to this question (in **bold**):

> Read the description below and answer the question based on what you read.
> This is a stage/outdoor with natural light. It is an open area where congregating and socializing can take place. Here we see an indian red umbrella, a light sky blue plastic bag, a light slate gray umbrella and a light pink umbrella in the foreground. In the background is a dark gray boot. A group of people walking down a street. People are lining the streets to see the parade.
> *The goal of the people is to see the parade. The people will gather on the street and line up to see the parade.*
> Q. What could prevent the people from achieving their goal?
> A. **The people could not achieve their goal if the weather was bad and the parade was canceled.**

The final narrative plan was the concatenation of all four answers, as shown in Table 1.

EXPRESSIVE TEXTS. We generated expressive texts that realized a particular narrative plan using GPT-3 and a few-shot learning paradigm[5]. Two sets of demonstrations were given to GPT-3, where each demonstration consisted of a narrative plan generated from SEE&TELL, and a story written by an author of this paper to show the type of translation of material from narrative plan to expressive text. We chose a few-shot approach because of the diversity in the narrative plans, and selected one narrative plan with a spontaneous stage direction and another with the 'robot'. We crafted a new GPT-3 prompt following the example below (narrative plan to be converted into expressive text in *italics*):

> Narrative Plan: The goal of the robot is to find an evacuation route for the people. The robot will achieve their goal by the robot will find the evacuation route for the people. But the robot could not find the evacuation route for the people. The robot must find the evacuation route for the people.
> Expressive Text: The robot was programmed with only one goal in mind: the safety of people. In dire situations, the robot was responsible for scouting out potential evacuation routes. If the robot couldn't secure a route, it would have utterly failed. It makes it its duty to always be on guard.
> ###
> Narrative Plan: The goal of the people is to have fun and enjoy themselves. The people will achieve their goal by they will walk down the street and have fun. But the people are walking down the street. The people must walk down the street and have fun.
> Expressive Text: Finally, it was time for fun. The group met up on the street, ready to walk and have fun. The worst thing would be if someone didn't join them, so they made sure everyone came along. They absolutely wanted to celebrate together.
> ###
> Narrative Plan: *The goal of the princess is to get married to the man she loves. The princess will achieve their goal by use a light pink umbrella. But the princess could not achieve their goal because she did not use a dark red umbrella. The princess must use a dark red umbrella.*
> Expressive Text:

GPT-3 followed the provided pattern and generated an expressive text based on the narrative plan at the end of the prompt, shown in Table 1.

Creativity Metrics for Visual Storytelling

We designed evaluation metrics for visual storytelling that were linked to theories of creativity and eliminate variability in interpreting the scores. To map to the novelty aspect of creativity, we modeled the relationship of new vs. grounded information with respect to the image. Instead of using a 5-point Likert scale, we crafted quantifiable measures that can be scored by trained annotators, and in the future, learned by automated scoring systems. To map to the usefulness aspect of creativity, we selected an 'engagement' application, and designed a scale to measure how effective the stories are at engaging a reader, rather than a generic interest Likert scale.

Visual Novelty

We measured novelty as the consecutive, step-wise preservation, loss, and gain of visual information throughout the visual storytelling pipeline. Figure 3 shows examples of how visual information is passed, lost, and gained through each stage of generation.

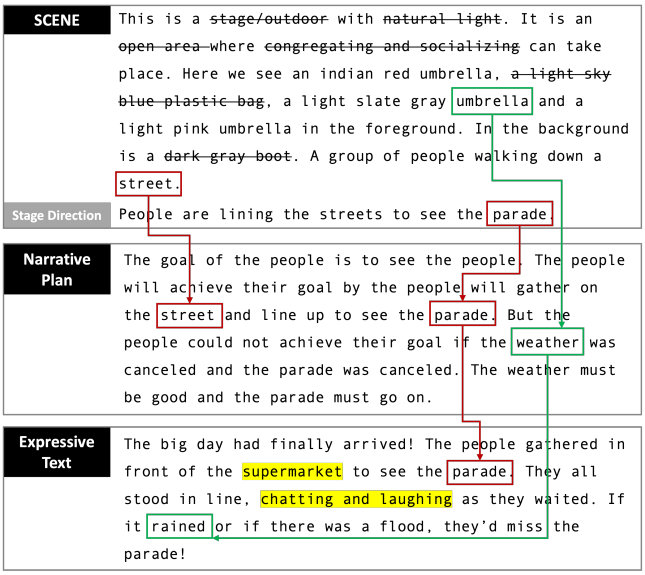

Figure 3: **Preservation, loss and gain of visual information.** Explicit and implicit preservation are annotated in red and green, respectively. Losses are crossed out, and gains are highlighted in yellow.

Preservation is the forward propagation of visual information in the SEE&TELL framework. In Figure 3, outlined in a red box, we see that *street* is present in the SCENE and appears again in the narrative plan, and that *parade* in the stage direction appears in the narrative plan, and again in the expressive text. Loss is the dropping of visual information such that is does not propagate forward in the framework. A number of visual information present in the SCENE in Figure 3 do not appear in the narrative plan, including *a light sky blue plastic bag* and *dark gray boot* (crossed out in the figure). Finally, gain is the introduction of new visual information not present in the immediate prior text in the framework. We note that no new information was introduced in the narrative plan in this example, but that *supermarket* and *chatting and laughing*[6] were introduced in the expressive text (highlighted in yellow in the figure).

[5] We used the davinci model, temperature of 0.8, and top p of 1.

[6] While *congregating and socializing* did appear in the SCENE, because these concepts were not carried over into the narrative plan, they then counted as novel concepts in the expressive text.

We note that while some pairs were word-for-word overlaps, e.g., *street*, others used similar terms implied to be related, for example, *umbrella*, *weather*, and *flood.* Both explicit and implicit matches (represented as red and green boxes in the figure respectively) were considered.

We model preservation, loss, and gain by first defining SC, SD, NP, and ET as the sets of visual information contained in the SCENE, stage direction, narrative plan, and expressive text respectively. The preservation-ratio from SCENE to narrative plan ($SC \rightarrow NP$) is defined as the overlap of visual information in SC and NP, divided by the total visual information in SC:

$$preserve_{SC \rightarrow NP} = \frac{|SC \cup NP|}{|SC|} \quad (1)$$

The loss ratio from $SC \rightarrow NP$ is defined as the visual information in SC that do not appear in NP, divided by the total visual information in SC:

$$loss_{SC \rightarrow NP} = \frac{|SC - NP|}{|SC|} \quad (2)$$

The gain-ratio from SCENE and stage direction to narrative plan ($SC, SD \rightarrow NP$) is defined as the visual information that appear in NP yet do not appear in SC or SD, divided by the total visual information in NP:

$$gain_{SC,SD \rightarrow NP} = \frac{|(NP - SC)| + |(NP - SD)|}{|NP|} \quad (3)$$

Using the same formula for preservation and loss, we compute $preserve_{SD \rightarrow NP}$ and $loss_{SD \rightarrow NP}$ for stage directions and narrative plans, and $preserve_{NP \rightarrow ET}$ and $loss_{NP \rightarrow ET}$ for narrative plans and expressive text. We additionally compute $gain_{NP \rightarrow ET}$ for gains in the expressive text from the narrative plan.

Usefulness Measured by Reader Engagement

To measure reader engagement as the usefulness our visual storytelling system, we create a forced categorical choice to better align responses by raters across a number of stories. The following definition is provided to situate the rater prior to giving their score: "A story world is comprised of characters, actions, and plot progression. Please give your rating based on the content of the narrative you read, rather than the delivery (i.e., grammar, spelling, word choice)." The rater is then shown an expressive text with matching image on a computer screen, and is given three options from which they must select one:

- "The storyworld is interesting. I would take action to read more, e.g., click or scroll to the next page."
- "The storyworld is somewhat interesting, but I wouldn't go out of my way to read more."
- "The storyworld is NOT interesting at all."

Annotation and Experimentation

To develop a high reliability for annotating the visual novelty metrics, six images from the VIST dataset (Huang et al. 2016) were used to generate six stories from SEE&TELL, and were doubly annotated by authors of this paper for explicit and implicit pairs between the SC and NP sets. Four rounds of adjudication were conducted until a high consensus was reached. Interannotator agreement was computed using Krippendorff's α (Krippendorff 1980), and were $\alpha = 0.92$ and $\alpha = 0.82$ for explicit and implicit pairs respectively, representing high agreement.

After the final revision of the annotation guidelines, a new set of eight images from the VIST dataset were run through SEE&TELL for each of the eight stage directions, resulting in 64 unique stories. The visual novelty annotation was conducted by one of the same annotators involved in the development of the schema, and the reader engagement was conducted by a single volunteer with a degree in linguistics.

Analysis and Discussion

Preservation, Loss, and Gain of Visual Information

We report the total visual information per set and averaged across all 64 stories as follows: $|SC| = 10.3$, $|SD| = 1.8$, $|NP| = 3.0$, $|ET| = 4.5$. Using Equations 1- 3, we compute the preservation, loss, and gain-ratios and report them as percentages in Table 2. We observe that across all stories, 90% of the visual information in the SCENE was lost and did not appear in the subsequent narrative plan ($loss_{SC \rightarrow NP}$). While this number might seem high for a system designed for *visual* storytelling, we acknowledge that no such baseline exists for the average or 'ideal' percentage of preservation or loss from an image to a narrative plan. Having it now, we can begin to understand the range of values observed, and unpack the interpretations.

To understand what these scores mean, we recall what is happening in SEE&TELL at this stage in the framework: GPT-3 has generated a narrative plan that answered the given question. GPT-3 was also given a description of the SCENE and the stage direction to use as background knowledge for determining the answer. In doing so, it was picking the visual information from what it was given that best help it aAnswer the question. Therefore we can claim that on average, 10% of visual information in the SCENE ($preserve_{SC \rightarrow NP}$) correspond more highly to answering the questions than the other 90%.

A range of $preserve_{SC \rightarrow NP}$ scores across stage direction is observed, ranging from a low of 2% in the 'lady' stage direction, to the 'spontaneous' stories having the highest percent of 23% This makes intuitive sense, as the 'spontaneous' stage direction is directly tied to the image itself, and is more likely to contain SCENE references that better help GPT-3 answer the question about that stage direction. For other scenarios, such as the 'lady,' neither a 'Lady of the Court' nor 'knight' nor their immediate mental associations (e.g., a castle, a horse, etc.) are apparent in the images, and thus the SCENES are less likely to contain visual information for GPT-3 to answer the question.

The opposite trend is exhibited for $loss_{SD \rightarrow NP}$. Here, on average, there was only a 15% loss from stage direction to narrative plan. We recall that the characters from the stage direction GPT-3 was given also appear in the questions. Therefore, to answer the question, GPT-3 is more likely to

	SC→NP		*SD→NP*		*NP→ET*		*SC,SD→NP*	*NP→ET*	*Engagement with ET*		
	Preserve	Loss	Preserve	Loss	Preserve	Loss	Gain	Gain	IntTakeAct	IntNoAct	NoInt
all	**10%**	**90%**	**85%**	**15%**	**89%**	**11%**	**16%**	**63%**	**31%**	**52%**	**17%**
agent	5%	95%	100%	0	91%	9%	0	59%	50%	38%	13%
lady	2%	98%	100%	0	74%	26%	40%	33%	13%	75%	13%
mage	9%	91%	100%	0	100%	0	18%	80%	38%	50%	13%
princ.	9%	91%	71%	29%	83%	17%	8%	67%	0	38%	63%
robot	5%	95%	65%	35%	90%	10%	5%	57%	38%	38%	25%
spont.	23%	77%	77%	23%	100%	0	6%	81%	13%	88%	0
search	10%	90%	100%	0	91%	9%	6%	59%	25%	63%	13%
thief	16%	84%	100%	0	86%	14%	28%	69%	75%	25%	0

Table 2: **Average preservation, loss, and gain-ratios, and usefulness statistics across 64 SEE&TELL stories.** The first column represents which stage direction was used to generate the story.

include visual information from the stage direction in its answer, including the characters, and a $preserve_{SD \rightarrow NP}$ of 85% suggests that GPT-3 is generating answers as we expect. Most of the stories have a 100% preservation-ratio, with the 'robot' stories scoring the lowest at 65%. We may interpret here that the visual information included in stage directions is not unilaterally incorporated by GPT-3 in answering questions; some visual information may be more challenging to incorporate than others. We take from these analyses that the stage direction is a stronger influence on GPT-3 for answering the questions than the SCENE, yet there is still a range preserved from the scene.

When looking at $loss_{NP \rightarrow ET}$, we observe an average of 11%. At this point in the pipeline, GPT-3 is being told to transform a narrative plan into an expressive text, where the examples matched a number of visual information. An average $presere_{NP \rightarrow ET}$ of 89% suggests that GPT-3 is again doing as it is instructed.

To unpack the gain-ratios, we understand that when moving from SCENE and stage direction to narrative plan, GPT-3 is still answering the question, and deciding that some elements of the story world were unknown but necessary, and novelly introduced to fit the answer. We observe a gain of 16% in $gain_{SC,SD \rightarrow NP}$, representing a subset of visual information GPT-3 decided was important in answering the question but was not given in the SCENE or stage direction. When moving from narrative plan to expressive text, GPT-3 is learning from few-shot demonstrations to embellish the narrative plan. A $gain_{NP \rightarrow ET}$ of 63% suggests that it is indeed following those instructions to add more detail to the expression of the story.

Willingness to Read More

The three right-most columns in Table 2 show the rater engagement scores. Of the 64 stories, 31% were scored as being interesting where the rater was willing to take action to read more (IntTakeAct), 52% were scored as being somewhat interesting but not willing to take action to read more (IntNoAct), and the remaining 17% were scored as being uninteresting and unwilling to reading more (NoInt). While the majority of responses fall in the middle of the scale, we recall that this scale is not a 1–3 Likert scale where the interpretation of the middle score can be interpreted by raters and experimenters differently. The middle score on our scale states that the stories were somewhat interesting. Therefore, 83% of our generated expressive texts were deemed to be of interest to the reader, while for the top 31%, the reader would have gone out of their way to read more.

Engagement greatly varied by stage direction, with the 'thief' stories having 75% of stories score IntTakeAct, yet the 'princess,' having 63% score NoInt. The rater provided free-text responses with their scores, and commented on whether or not the story "fit" the picture, revealing that that may have been tied to their scoring. For example, for the top story in Table 1, the rater scored it as IntNoAction and wrote *The story fit the picture, but was not enthralling.* For the bottom story in the same table, the rater scored it IntTakeAct and wrote *Ominous as to who the thief could be in the picture.*

Conclusions and Future Work

This paper presented SEE&TELL as a visual storytelling framework that separates 'seeing' from 'telling' and implements a question-answering and few-shot approach to generate stories. Our evaluation revealed that our system preserves 10% of visual information from SCENE to narrative plan, and that 16% and 63% of the visual information in the narrative plan and expressive text respectively are novelly generated and embellished by GPT-3. Eighty three percent of the expressive texts were deemed interesting, with 31% being engaging.

Our visual novelty scores provided a baseline for the next step in this research to explore if preservation, loss, and gain-ratios can be methodologically altered and subsequently compared to examine what stories with different ratios look like. Of particular interest, is the correlation of these stories with rater scores to examine the interplay between visual novelty and usefulness, especially when determining if the 10% preservation-ratio we achieved in narrative plan and 89% in expressive text are the 'right' amount of preservation or not. Furthermore, we can explore if there is a 'breaking point' at which too much or little new visual information is deemed as boring or irrelevant.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*, 2425–2433.

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

Castricato, L.; Frazier, S.; Balloch, J.; Tarakad, N.; and Riedl, M. 2021. Automated Story Generation as Question-Answering. *arXiv:2112.03808*.

Dhakar, L. 2020. Color Thief. https://github.com/lokesh/color-thief.

Eum, S.; Han, D.; and Briggs, G. 2020. SomethingFinder: Localizing Undefined Regions Using Referring Expressions. In *CVPR Workshops*.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. *Hierarchical Neural Story Generation*. ACL.

Gatt, A.; and Reiter, E. 2009. SimpleNLG: A Realisation Engine for Practical Applications. In *ENLG*, 90–93.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *CVPR*. IEEE.

Howcroft, D. M.; and Rieser, V. 2021. What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think. In *EMNLP*, 8932–8939. Online and Punta Cana, Dominican Republic: ACL.

Hu, J.; Cheng, Y.; Gan, Z.; Liu, J.; Gao, J.; and Neubig, G. 2020. What makes a good story? Designing composite rewards for visual storytelling. In *AAAI*, volume 34, 7969–7976.

Huang, T.-H.; Ferraro, F.; Mostafazadeh, N.; Misra, I.; Agrawal, A.; Devlin, J.; Girshick, R.; He, X.; Kohli, P.; Batra, D.; et al. 2016. Visual storytelling. In *NAACL*, 1233–1239.

Jana, T. 2019. Pix2Story: Neural storyteller which creates machine-generated story in several literature genre. https://azure.microsoft.com/en-us/blog/pix2story-neural-storyteller-which-creates-machine-generated-story-in-several-literature-genre/.

Ji, Z.; Xu, Y.; Cheng, I.; Cahyawijaya, S.; Frieske, R.; Ishii, E.; Zeng, M.; Madotto, A.; Fung, P.; et al. 2022. VScript: Controllable Script Generation with Audio-Visual Presentation. *arXiv:2203.00314*.

Krippendorff, K. 1980. *Content analysis: An introduction to its methodology*. Sage publications.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv:2107.13586*.

Lönneker, B. 2005. Narratological knowledge for natural language generation. In *ENLG*.

Lukin, S. M.; Hobbs, R.; and Voss, C. R. 2018. A pipeline for creative visual storytelling. *arXiv:1807.08077*.

Meehan, J. R. 1977. TALE-SPIN, An Interactive Program that Writes Stories. In *IJCAI*, volume 77, 91–98.

Patterson, G.; and Hays, J. 2012. SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes. In *CVPR*.

Patterson, G.; Xu, C.; Su, H.; and Hays, J. 2014. The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. *IJCV*, 108(1-2): 59–81.

Pérez y Pérez, R.; and Sharples, M. 2001. MEXICA: A computer model of a cognitive account of creative writing. *JETAI*, 13(2): 119–139.

Propp, V. I. 1968. *Morphology of the folktale*. University of Texas Press.

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

Riedl, M. O.; and Young, R. M. 2005. Open-World Planning for Story Generation. In *IJCAI*, 1719–1720. Citeseer.

Sai, A. B.; Mohankumar, A. K.; and Khapra, M. M. 2020. A survey of evaluation metrics used for NLG systems. *arXiv:2008.12009*.

Singh, A.; Goswami, V.; Natarajan, V.; Jiang, Y.; Chen, X.; Shah, M.; Rohrbach, M.; Batra, D.; and Parikh, D. 2020. MMF: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf.

Stage Direction. 2018. Collins Dictionary. *Accessed November 14, 2022 [Online]*. HarperCollins.

Summers-Stay, D.; Bonial, C.; and Voss, C. 2021. What Can a Generative Language Model Answer About a Passage? In *Workshop on Machine Reading for Question Answering*.

Turner, S. R. 1993. *Minstrel: a computer model of creativity and storytelling*. University of California, Los Angeles.

Wallas, G. 1926. *The Art of Thought*, volume 10. Harcourt, Brace.

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

Wang, X.; Chen, W.; Wang, Y.-F.; and Wang, W. Y. 2018. *No metrics are perfect: Adversarial reward learning for visual storytelling*. ACL.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Xiao, J.; Ehinger, K. A.; Hays, J.; Torralba, A.; and Oliva, A. 2016. Sun database: Exploring a large collection of scene categories. *IJCV*, 119(1): 3–22.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 3485–3492. IEEE.

Yao, L.; Peng, N.; Weischedel, R.; Knight, K.; Zhao, D.; and Yan, R. 2019. Plan-and-write: Towards better automatic storytelling. In *AAAI*, volume 33, 7378–7385.

Yin, W.; Wang, X.; Shen, C.; Liu, Y.; Tian, Z.; Xu, S.; Sun, C.; and Renyin, D. 2020. DiverseDepth: Affine-invariant Depth Prediction Using Diverse Data. *arXiv:2002.00569*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million Image Database for Scene Recognition. *TPAMI*.