Visually Grounded Story Generation Challenge

Xudong Hong[124], **Asad Sayeed**[3], **Khushboo Mehra**[24], **Vera Demberg**[24] **and Bernt Schiele**[14]

[1]Dept. of Computer Vision and Machine Learning, MPI Informatics
[2]Dept. of Language Science and Technology and Dept. of Computer Science, Saarland University
[3]Dept. of Philosophy, Linguistics, and Theory of Science, University of Gothenburg
[4]Saarland Informatics Campus, Saarbrücken
{xhong,kmehra,vera}@lst.uni-saarland.de
schiele@mpi-inf.mpg.de, asad.sayeed@gu.se

Abstract

Recent large pre-trained vision-and-language models have achieved strong performance in natural language generation. However, most previous generation tasks neither require coherent output with multiple sentences nor control the output text by grounding the output in the input. We propose a shared task on visually grounded story generation, where the input is an image sequence, and the output is a story that is conditioned on the input images. This task is particularly challenging because: 1) the output story should be a narratively coherent text with multiple sentences, and 2) the protagonists in the generated stories need to be grounded in the images. We aim to advance the study of vision-based story generation by accepting submissions that propose new methods.

1 Introduction

Vision-based language generation (VLG) is to generate text from visual input. It is a challenging but interesting task because it requires joint vision and language modeling. Recent large pre-trained vision-and-language models (VLMs) like GPT-4 (OpenAI, 2023) or MiniGPT-4 (Zhu et al., 2023) have shown great success on several multi-modal tasks, such as image captioning (Vinyals et al., 2016), visual question answering (Goyal et al., 2017) and visual dialog generation (Das et al., 2017).

Despite recent breakthroughs, current tasks only require models to predict a label or generate short texts (i.e., less than 30 words). It is unclear whether the newest VLMs can generate coherent texts with multiple sentences from visual input. On the contrary, humans can produce long and localy coherent texts from the same visual input. To investigate machine intelligence, we need a task that is more similar to human behavior (Bubeck et al., 2023).

Several previous tasks have been proposed to test the capabilities of VLMs to handle longer output, such as visual paragraphs (Krause et al., 2017), localized narratives (Pont-Tuset et al., 2020), and video captioning (Voigtlaender et al., 2023). However, these tasks are designed for literal descriptions where sentences are independent of each other, rather than for coherent text. Coherence is a fundamental property of human language. In particular, local coherence, which refers to the relations between entities in context, affects language comprehension and production. Local coherence is essential for vision and language (V&L) research because: **1.** It has many applications in vision and language tasks. For example, a better model of local coherence can improve the performance of text-to-image retrieval (Park and Kim, 2015). **2.** Modeling coherence is a prerequisite for modeling event knowledge as events center around entities. Better event modeling improves vision and language pre-training (Zellers et al., 2021, 2022).

Story generation is a well-studied task in natural language generation, widely used for testing whether large pretrained models can track entities (Paperno et al., 2016) and generate locally coherent texts. Unlike image captions, stories contain several characters and events involving recurrent characters and their interactions with each other and the environment. In addition, *characters* and *relevant content* are among the most critical aspects of story writing (Goldfarb-Tarrant et al., 2020). We argue that story generation is a suitable benchmark for testing whether VLMs can generate coherent texts.

In this work, we propose a new shared task, Visually Grounded Story Generation (**VGSG**), which requires the VLMs to generate stories with protagonists grounded on images. We aim for coherent and visually grounded stories with high diversity. This task is particularly challenging for two reasons: **1.** The protagonists in the generated stories need to be grounded in the images, meaning that their actions and descriptions should be consistent with the

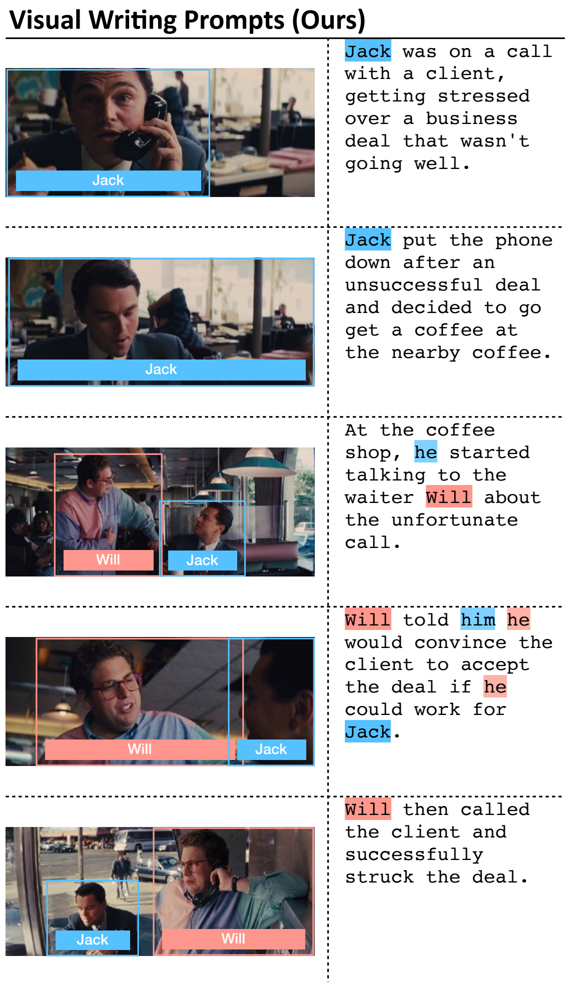

Figure 1: Example of Visual Grounded Story Generation on Visual Writing Prompts dataset. The dataset has recurring characters across all five images and sub-stories. Each occurrence of a character in a sub-story has a bounding box in the corresponding image, which grounds the textual appearance to visual input.

visual information provided. **2.** The output story needs to be a coherent text, meaning that it should have a clear beginning, middle, and end, and flow logically from one sentence to the next.

We hope that this task will help the exploration of VLG by encouraging participants to propose new methods that generate coherent and visually grounded stories. We welcome submissions from researchers around the world who are interested in tackling this exciting challenge. We also seek for researchers who are interested to join the organization of this shared task.

2 Related Work

VLG with Coherence. One relevant task is Visual Storytelling (Huang et al., 2016), where the input is a sequence of images and the output is a coherent story. Another task that requires some sort of coherence in the generated text is movie description (Rohrbach et al., 2015), where the input is a video clip from the movie and the output is the corresponding text description of the scene. Chandu et al. (2019) propose a dataset of procedural text from recipes with instructional images, but characters are not explicitly annotated. Unfortunately, the local coherence of the generated text is not evaluated in either of these tasks (Mitchell et al., 2018).

Visual Story Generation. Most of the previous tasks for visual story generation have several limitations: there is no sequence of events behind the images (Park and Kim, 2015; Huang et al., 2016) or the dataset is limited in scale (Xiong et al., 2019). None of them can be used for evaluating visual grounding. Mitchell et al. (2018) hosted the first shared task of visual story generation. But there are no automatic evaluations of either coherence or visual grounding. Our shared task is the first to jointly evaluate the coherence and visual grounding of generated stories.

3 Task Description

We define the VGSG task as follows: given a sequence of images (like the first column of Figure 1) the system needs to generate a coherent short story conditioned on the image sequence (like the second column of Figure 1). In addition, the generated story should contain the characters seen in the image sequence.

The VGSG shared task focuses on coherent and visually grounded stories with high diversity.

3.1 Datasets

To evaluate the submissions, we will use two datasets that provide grounding annotations for characters:

Visual Writing Prompts (VWP; Hong et al., 2023b), a vision-based dataset that contains 2K image sequences aligned with 12K human-written stories in English.[1] Each image is corresponding to a part of a story. Instances of each protagonist are annotated with the character's name (see Figure 1).

VIST-Character by Liu and Keller (2023) which has visual and textual annotations for recurring characters in 770 stories from the test split of the

[1] https://vwprompt.github.io/

Name	Image Genre	Story Genre	Story Source	# Story	# image per Story	# token per Story
VWP	movie	short story	crowdworker	12 K	[5, 10]	83.7
VIST	photo	short story	crowdworker	50 K	5	57.6
Travel blogs	photo	blog	blogger	10 K	1	222.3‡
MSA	movie	movie synopsis	fan	5 K	92	129

Table 1: Statistics of datasets. Numbers with ‡ are obtained from a small sample of the Disney split of the dataset that is available in their repository.

VIST dataset (Huang et al., 2016), along with an importance rating of all characters in any story.[2] We only use it for evaluation.

We also evaluate on these datasets:

Visual Storytelling (VIST; Huang et al., 2016) is a widely used dataset with 50K image-story pairs.

Travel blogs (TB; Park and Kim, 2015) are two datasets with 10K image sequence-story pairs extracted from travel blogs of visiting New York City or Disneyland.

Movie Synopses Associations (MSA; Xiong et al., 2019) contains movie synopses from 327 movies where there are 4494 scenes aligned with corresponding paragraphs in synopses.

These data sets are publicly available so there's a risk of exposure to the participants. To ensure a fair comparison and make the task more challenging, we collect additional data following the data collection process of these works combine with selected subsets as blind test sets. The statistics of all the datasets are in Table 1.

3.2 Tracks

The VGSG shared task contains three tracks: **Strict Track** focuses on exploring Language and Vision Mapping methods and Language Generation models through a controlled experiment. We provide extracted visual features from a pre-trained vision model, which participants can only use as input to train their models with the provided dataset.

Open Track aims to test the state-of-the-art of the task. Participants can use all kinds of resources, including pre-trained models and additional text or vision-only datasets. However, they cannot use other vision and language datasets apart from the provided dataset.

Grounding Track is based on the Open Track, but participants are required to submit a mapping of all entities in the generated text and provided characters (see Figure 2 for an example). The submissions to this track will be evaluated on the VIST-Character dataset (Liu and Keller, 2023).

3.3 Schedule

We propose the following tentative schedule:

Dec 1st, 2023 We will announce the joint task at the INLG 2023 conference (if accepted), with data available on the task's dedicated website. This is the point when individuals can sign up for the task.

Feb 1st, 2024 The submission is opened. Participants can submit their systems to the organizers.

May 1st, 2024 Submission ends at this point and organizers start the process of automatic evaluation on blind test sets and human evaluation of the systems.

Jun 1st, 2024 The VGSG shared task comes to a conclusion. The organizers will submit reports regarding participant performance and overall challenge outcomes to the INLG 2024 conference and will present these findings at the event. The previously concealed test set will be released to the public.

	Jack	Will
Jack was on a call with a client, getting stressed over a business deal that wasn't going well.	1	−1
Jack put the phone down after an unsuccessful deal and decided to go get a coffee at the nearby coffee.	1	−1
At the coffee shop, he started talking to the waiter Will about the unfortunate call.	1	1
Will told him he would convince the client to accept the deal if he could work for Jack.	1	1
Will then called the client and successfully struck the deal.	−1	1

Figure 2: Example a matching matrix between entities in the generated story and the character in the images.

[2]https://github.com/iz2late/VIST-Character

4 Evaluation

We will perform both automatic and human evaluations for the submissions. The scripts for all automatic metrics will be provided after the submission system is open; human evaluation will be conducted after all submissions have been received. We will release the annotator instructions and source code of all metrics after the shared task.

4.1 Automatic Evaluation

We will use metrics in the following categories to evaluate the submissions:

Reference-based metrics including unigram (B-1), bigram (B-2), trigram (B-3), and 4-gram (B-4) BLEU scores (B; Papineni et al., 2002), METEOR (M; Banerjee and Lavie, 2005), ROUGE-L (R; Lin, 2004), and CIDEr (C; Vedantam et al., 2015), which were used in the previous visual storytelling shared task (Mitchell et al., 2018). We will also use BERTScore (BS; Zhang* et al., 2020) which is effective in text summarization.

Grounding To measure the correctness of referring expressions of human characters in stories, we will use the character-matching (CM) metric defined in (Hong et al., 2023a).

Event diversity we will use metrics used by Hong et al., 2023b (based on (Goldfarb-Tarrant et al., 2020)) including the unique number of verbs, verb-vocabulary ratio, verb-token ratio, percentage of diverse verbs not in the top-5 most frequent verbs and unique:total ratios of predicate unigram, bigram, and trigram.

Coherence following Hong et al., 2023b we will use the generative Entity Grid model to calculate the log-likelihood based on entity transitions in system outputs.

4.2 Human Evaluation

In natural language generation tasks, automatic metrics do not provide a full understanding of the quality of the generated text. Reference-based metrics, in particular, have been shown to not correlate well with human judgment. In addition, several important aspects of narratives such as creativity, and logical coherence are hard to judge using automatic evaluation. Therefore, we will also conduct a human evaluation for the submissions, focussed on narrativity (whether the generation is a story or simply a description of images), character grounding (correctness of referring expressions, model hallucinations), and coherence. The scale of the evaluation depends on the funding we have. We also encourage participants to perform their own human evaluation and include the results in their reports.

4.3 Baselines

Our baselines are:

Seq2Seq (Huang et al., 2016) is a simple but powerful model with an encoder-decoder architecture. Visual features are first projected with an encoder which is a feed-forward neural network, then fed to the decoder which is a pre-trained language model.

TAPM (Yu et al., 2021) is a Transformer-based model which adapts the visual features with pre-trained GPT-2.

Other V&L models We also include other vision and language models that are competitive on similar vision and language tasks like Cho et al. (VL-T5; 2021), Li et al. (BLIP; 2022) and Zhu et al. (MiniGPT-4; 2023).

5 Conclusions

This proposal introduces a novel shared task called Visually Grounded Story Generation, which necessitates that Visual Language Models formulate narratives with protagonists based on image inputs, ensuring the production of coherent and visually grounded stories with high diversity. The task poses dual challenges: the need for protagonists' actions and descriptions to align with the provided visual information and the requirement for the output story to logically progress with a clear beginning, middle, and end. By initiating this task, the authors aim to foster advancements in Visual Language Generation, inviting global researchers to contribute new methodologies that facilitate the creation of visually consistent, logically structured stories.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Khyathi Chandu, Eric Nyberg, and Alan W Black. 2019. Storyboarding of recipes: Grounded contextual generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6040–6046, Florence, Italy. Association for Computational Linguistics.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Xudong Hong, Vera Demberg, Asad Sayeed, Qiankun Zheng, and Bernt Schiele. 2023a. Visual coherence loss for coherent and visually grounded story generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023b. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Danyang Liu and Frank Keller. 2023. Detecting and grounding important characters in visual stories. In *37th AAAI Conference on Artificial Intelligence*. AAAI Press.

Margaret Mitchell, Ting-Hao 'Kenneth' Huang, Francis Ferraro, and Ishan Misra, editors. 2018. *Proceedings of the First Workshop on Storytelling*. Association for Computational Linguistics, New Orleans, Louisiana.

OpenAI. 2023. Gpt-4 technical report.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Cesc C Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. *Advances in neural information processing systems*, 28:73–81.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663.

Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. 2023. Connecting vision and language with video localized narratives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2461–2471.

Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. 2019. A graph-based framework to bridge movies and synopses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4592–4601.

Youngjae Yu, Jiwan Chung, Heeseung Yun, Jongseok Kim, and Gunhee Kim. 2021. Transitional adaptation of pretrained models for visual storytelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12658–12668.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.