On the Challenges in Evaluating Visually Grounded Stories

Aditya K Surikuchi*, Raquel Fernández and Sandro Pezzelle

Institute for Logic, Language and Computation
University of Amsterdam, Netherlands

Abstract

Producing stories grounded in visual content is an inherent trait of human intelligence and an integral aspect of interpersonal communication. With the surge of advanced vision-to-language models, there has been increased interest in developing and understanding the capabilities of models to generate visually grounded narratives. However, recent research has highlighted the challenges in evaluating model-generated stories. In this work, we study these evaluation limitations in the visually grounded story generation task by focusing on the recently released Visual Writing Prompts dataset and shared task. Through this study, we also explore the capabilities of several general-purpose vision-to-language foundation models for generating stories grounded in sequences of images. We observe that some recent models, such as Qwen2.5-VL, can generate stories that are coherent, consistent, and well-grounded in the visual data. Nevertheless, in line with the recent studies in this area, we find that the existing automatic evaluation metrics and methods are insufficient in fully capturing all the aspects essential for assessing model-generated stories. We believe our findings reinforce the evidence and arguments emphasizing the need for improvements to automatic approaches that can comprehensively evaluate and understand models for visual storytelling.

Keywords

visual storytelling, visually-grounded story generation, vision-to-language models, evaluation, NLG

1. Introduction

Given a sequence of multiple temporally ordered images as input, the visual storytelling or visually grounded story generation task requires models to generate plausible and compelling textual stories. Huang et al. [1] proposed this task and released the VIST dataset to facilitate the development of models that can generate stories based on the causal structure of the visual input sequence. Leveraging the VIST dataset, various modeling approaches have been proposed over the years [2, 3, 4, 5, 6, 7]. For evaluating stories generated by models, most work has predominantly resorted to using automatic reference-based n-gram metrics such as BLEU [8] and METEOR [9]. These metrics neither consider the visual input when assessing the generated stories nor do they account for the fact that several creative stories are plausible for a given image sequence. The research community has underlined this problem and proposed reference-free automatic evaluation metrics to measure the quality of stories along different dimensions such as visual grounding, coherence, and repetition [10, 11]. However, the latest work in this direction has shown that evaluating model-generated outputs in visual storytelling requires consideration of more aspects besides measuring the degree of visual grounding, coherence, and repetition [12].

To verify this argument, in this work, we focus on the recently released Visual Writing Prompts (VWP) [13] dataset from the Visually Grounded Story Generation challenge [14] and explore different aspects related to modeling and evaluation. First, with the VWP dataset, we train and generate stories using models that are shown to perform well on the VIST dataset. We then consider several vision-to-language foundation models designed for general-purpose tasks and use them to generate stories for the VWP dataset in a zero-shot manner. Using the evaluation framework proposed in Surikuchi et al. [12], we compare all the models and find that general-purpose VLMs achieve better results quantitatively in terms of the three different dimensions considered for assessment—*visual grounding*, *coherence*, and

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'25 Workshop, Lucca (Italy), 10-April-2025

*Corresponding author.

a.k.surikuchi@uva.nl (A. K. Surikuchi); raquel.fernandez@uva.nl (R. Fernández); s.pezzelle@uva.nl (S. Pezzelle)

repetition. However, through qualitative verification, we identify that the existing metrics do not fully capture all the aspects relevant for evaluating the quality of a story. These results are in line with the findings of recent studies and we support the arguments for exploring and considering more nuanced dimensions such as consistency of emotions and differentiating creative expressions from hallucinations for evaluating visual storytelling. Our code is available at: akskuchi/vwp-visual-storytelling.

2. VGSG Challenge

The Visually Grounded Story Generation (VGSG) shared task [14] was proposed to test the capabilities of AI models to generate coherent, grounded, and diverse short stories for sequences of images. It comprised three different tracks—`closed`, `open`, `grounding`—and used the Visual Writing Prompts (VWP) [13] dataset which we describe below. Further details pertaining to the challenge tracks, our approaches, and experiments are discussed in Section 3.

Data. Hong et al. [13] introduced the VWP dataset to overcome the various limitations present in other visual storytelling datasets such as VIST [1]. Primarily, image sequences in VWP are constructed to be semantically well-connected and centered around recurring characters such that they serve as meaningful writing prompts for human-annotators or AI models. VWP contains a total of 13213 sequences, each comprising 5 to 10 images from a curated set of frames obtained from the MovieNet dataset [15]. For the selected image sequences, stories were provided by Amazon Mechanical Turk crowd workers. The obtained text was then processed to anonymize the recognized named locations and characters using placeholders (e.g., [`female0`], ... ,[`femaleN`]). The overall dataset was split into 11778 training, 849 validation, and 586 test samples. VWP is shown to have more events and characters per story compared to the other visual storytelling datasets such as VIST. An example <image sequence, story> pair from the VWP dataset is shown in Figure 2.

Evaluation. Stories submitted to the challenge were evaluated using both automatic and human evaluation methods. Reference-based metrics such as BLEU [8], METEOR [9], ROUGE [16], and CIDEr [17] were used for comparing model-generated candidate stories with reference stories provided by humans. Stories were also evaluated in a reference-free manner along three dimensions—*coherence*, *character grounding*, and *diversity*. In the context of the shared task, coherence was defined in terms of entity transitions throughout the text and the Generative Entity Grid [18] metric was used for computing it. Character grounding scores were computed using the Character Matching [19] metric which measures the degree of match between the 'appearance' matrices of characters present in the image sequence and the generated text. Diversity of stories is measured as an average of various aspects such as unique number of verbs, verb-to-vocabulary ratio, verb-token ratio, and percentage of diverse verbs not in the top-5 most frequent verbs. An evaluation dashboard with reference-based metrics was made available during the shared task training phase to verify the stories generated by models for the VWP validation data split.[1] Furthermore, organizers also conducted human evaluation on the submitted stories and discussed their qualitative findings in the overall shared task summary report [20].

3. Our Approach

In this section, we discuss the various modeling approaches we used for the VGSG task both during and after the completion of challenge. Specifically, we participated in the `open` and `closed` tracks of the challenge and leveraged two models introduced in Surikuchi et al. [12]. We also consider two additional state-of-the-art vision-language models (VLMs) to shed light on their zero-shot capabilities for the visual story generation task. To comprehensively assess the quality of model-generated stories in terms of their closeness to corresponding human-written ones, we make use of the recently proposed

[1] https://huggingface.co/spaces/VGSG/TestVGSG

human-centric evaluation method—$\mathbf{d}_{HM}$ [12]. The remainder of this section describes the shared task tracks, the models we used, and the $\mathbf{d}_{HM}$ evaluation method.

3.1. Open Track

The objective of the `open` track was to test the current state-of-the-art of the VGSG task and allowed participants to use any pre-trained visual encoders and textual decoders. Therefore, for this track, we used the TAPM (+Llama 2) and LLaVA (*visual context*) models proposed in Surikuchi et al. [12]. Similar to LLaVA, we consider two additional state-of-the-art VLMs and use them off-the-shelf for this task.

TAPM (+Llama 2). Transitional Adaptation of Pretrained Models (TAPM) is an approach originally proposed by Yu et al. [7] for the visual storytelling task. It follows the visual encoder to language decoder architecture commonly used in models for visual storytelling. First, image-level and object-level features of the input image sequence are obtained using pre-trained ResNet [21] and FasterRCNN [22] respectively. These features are passed through the visual encoder, and the representations of the image at each temporal position are pooled together with the features of images at the neighboring positions for improved context. The context from the encoder is passed on to a pre-trained GPT-2 [23] for story generation. Prior to this downstream task-specific fine-tuning, for a pre-determined number of epochs, TAPM comprises an adaptation step in which the language decoder is frozen and the visual encoder parameters are adapted based on the outputs of the frozen decoder. The authors argue that this step harmonizes the various pre-trained components of the model and facilitates semantic alignment between visual and textual representations. Recently, it has been shown that replacing GPT-2 using the Llama 2 [24] language model—TAPM (+Llama 2)—improves model performance across different datasets including VWP [12].

We note that Surikuchi et al. [12] used the VWP dataset version `v1.0.0` to train and test their models.[2] However, through the VGSG challenge authors released VWP `v2.1` which provided anonymized stories and included details pertaining to the training, validation, and test splits.[3] Therefore, we trained the TAPM (+Llama 2) model from scratch using VWP `v2.1` by following the procedure described in Surikuchi et al. [12].

Off-the-shelf VLMs. Large Language and Vision Assistant (LLaVA) [25] is a large vision-language foundation model pre-trained for various general-purpose tasks such as image captioning and visual question answering. We use LLaVA`v1.6` in a zero-shot manner by prompting it under the *visual context* setting proposed in Surikuchi et al. [12]. Specifically, we provide the model with the entire image sequence—sequence of images combined horizontally into a single composite image—as input and prompt it to generate a story with `[num-images-in-the-sequence]` sentences. To ensure that the generated stories are not sensitive to the prompt, we use three variations of the prompts and report the average of the resulting scores during evaluation. Besides LLaVA`v1.6`, we use two recent similarly sized general-purpose VLMs—Qwen`2.5`-VL [26] and DeepSeek-VL [27]—that have demonstrated strong performance on various vision-language benchmarks. Using the technical reports and the open source information, we verified that the VLMs were not directly pre-trained using any of the visual storytelling datasets, including VWP. Additional details concerning the models are provided in Appendix A and the inference procedure including the prompts are provided in Appendix B.

3.2. Closed Track

The `closed` track was a controlled setting in which visual features for VWP images were extracted using the pre-trained SwinTransformer [28] and provided as part of the VGSG challenge data. Participants of this track were instructed to not use any additional visual feature extractors and to focus on the components that map the vision and language modalities. For this track, we modified the TAPM (+Llama

[2] https://github.com/vwprompt/vwp/releases/tag/v1.0.0
[3] https://huggingface.co/datasets/tonyhong/vwp

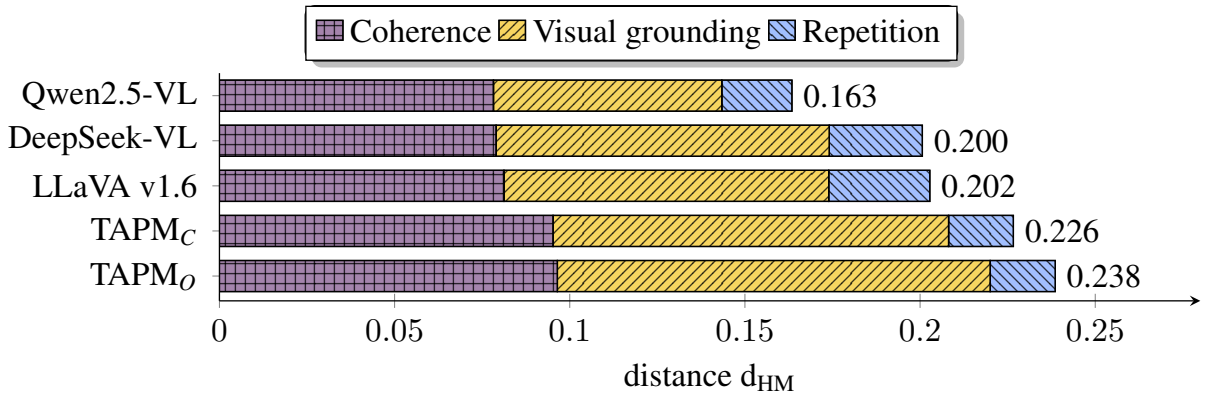

Figure 1: $\mathbf{d}_{HM}$ distances (the lower the better) for all models on the VWP validation set.

2) approach outlined in Section 3.1 to leverage the image-level and object-level SwinTransformer features. In the subsequent text, we refer to this model as TAPM$_C$ and the one used for the open track as TAPM$_O$.

3.3. Evaluation

In this work, we aim to understand the degree to which model-generated stories comply with stories produced by humans regarding three different aspects essential for visual story generation—*coherence*, *visual grounding*, and *repetition*. For these aspects, we leverage the definitions and metrics operationalized by Surikuchi et al. [12]. Specifically, visual grounding is assessed using the GROOViST [11] metric, which measures the degree of alignment between noun phrases in the story and the bounding boxes in the images of the sequence. Coherence is operationalized using the RoViST-C [10] metric, which measures the average probability with which each sentence follows the preceding sentences. For repetition, the RoViST-NR [10] metric is used, which measures 'non-redundancy' in terms of the number of inter- and intra-sentence co-occurring words. We note that despite being referred to using similar terms, these three aspects are distinct in terms of their definitions from those considered for the VGSG shared task evaluation (described in Section 2).

Using the three metrics, we first obtain coherence, visual grounding, and repetition scores for both model-generated stories and corresponding human-annotated stories, independently. We then compute the absolute differences between the human stories and the model-generated ones to measure metric-level deviations (d^C_{HM}, d^G_{HM}, d^R_{HM}). Finally, to quantify the degree of 'closeness' between model- and human-stories, we compute the aggregate distance $\mathbf{d}_{HM}$ as the average of metric-level differences:

$$\begin{aligned} \mathrm{d}^C_{HM} &= |C_H - C_M|, \mathrm{d}^G_{HM} = |G_H - G_M|, \mathrm{d}^R_{HM} = |R_H - R_M|, \\ \mathbf{d}_{HM} &= (\mathrm{d}^C_{HM} + \mathrm{d}^G_{HM} + \mathrm{d}^R_{HM})/3 \end{aligned} \tag{1}$$

4. Results

Figure 1 shows the overall $\mathbf{d}_{HM}$ distances along with individual metric-level deviations for all the models. Model-generated stories with lower $\mathbf{d}_{HM}$ values are expected to be closer to human levels of storytelling in terms of the three dimensions—*coherence*, *visual grounding*, and *repetition*. Firstly, we observe that compared to the TAPM models, all off-the-shelf VLMs achieve better overall distances. This pattern differs from the results reported in recent work [12] for the earlier version of the VWP dataset (`v1.0.0`). However, the TAPM models achieve the lowest distance in terms of repetition, which aligns with the findings in Surikuchi et al. [12]. Among the foundational VLMs, `Qwen2.5-VL` achieves the lowest overall distance and the difference compared to other models is particularly significant along the dimension of visual grounding. We hypothesize that `Qwen2.5-VL`—which is a relatively recent VLM—benefits from the extensive pre-training on long video understanding datasets and interleaved

Human Story	It was a normal evening for [female0] when she entered her room. She saw her friend [female1] sitting in the bed. [female0] didn't know what was happening. [female1] looks like she has been crying. [female0] was confused by [female1]'s state. She asked [female1] what was happening as she didn't know why [female1] is crying like that. Then [female1] explained that she finally got a job after all the struggles she faced. Eve started to open her bag. [female0] was so happy for her friend as she finally able to live her life as she wishes. Eve took few money from her bag saying that it is her advanced payment for the job. They went to eat out to celebrate their happiness.
Qwen2.5-VL	In the cozy bedroom, Sarah **noticed her friend**, Emily, **looking distressed**. Emily sat on the bed, her head in her hands, **clearly upset**. Sarah approached her, concerned, and asked what was wrong. Emily, still in shock, told Sarah about a recent betrayal she had discovered. Sarah **listened intently, offering comfort and support**. As they talked, Sarah noticed a **stack of money** in Emily's **handbag**. Curious, she asked about it. Emily explained that the money belonged to someone else but had been left behind. Sarah suggested they return it. They decided to go to **Pink's Chili Dogs, a local favorite**.
LLaVAv1.6	1. In a bustling city, a young woman finds herself lost in a maze of **narrow streets**. 2. She stops to ask for directions, her eyes wide with confusion. 3. **The man she approaches is kind and offers to guide her.** 4. As they walk, he tells her stories of the city's history and culture. 5. They come across a small, hidden bookstore, its shelves filled with ancient tomes. 6. The woman is captivated by the sight of a rare, first edition book. 7. **The man helps her purchase it**, and they continue their journey together. 8. As they walk, they share their dreams and aspirations, finding a connection in their shared love of literature. 9. They reach a beautiful garden, where they sit and read together under the stars. 10. As the night wears on, they realize they have found something special in each other's company.
DeepSeek-VL	In a cozy library, a young woman in a vibrant red sweater engrossed in a book, her fingers gently turning the pages. Nearby, **an older man** in a warm, brown coat, his eyes scanning the shelves, searching for a specific title. The soft glow of the library lights illuminates their focused expressions, as they delve into the world of knowledge. **Outside, a child's laughter** pierces the air, drawing the attention of the patrons to a **colorful** scene of **joy and play**. The sequence continues, capturing the essence of a day in the life of this library, where stories unfold in the quiet corners and on the open shelves.
TAPM$_O$	**He** walked into his office, but it was dark and empty. The woman in pink asks the woman in brown where the earring might be. The woman in brown says that the man probably hid the earring inside a book. The woman in pink thinks about this. Then the woman in pink asks the woman in brown if she is certain about this. The woman in brown insists that the man would have hidden the **earring** inside a **book**. But she was feeling bad about it. [female0] cried, but there was nothing she could do. In one of the presents it is full of stacks of money. They were ready to take off in their **helicopters** and go to **battle**.
TAPM$_C$	Her **husband [male0] was looking at her while she was on the floor**. The woman in pink asks the woman in brown where the earring might be. The woman in brown says that the man probably hid the earring inside a book. The woman in pink thinks about this. Then the woman in pink asks the woman in brown if she is certain about this. The woman in brown insists that the man would have hidden the **earring** inside a book. The woman checks the book by the **computer**, and sure enough, she finds her earring. She called her friend just to check on her. She kept thinking about how she would get out of this job. **It was a large and grand house with grand gardens**.

Figure 2: Stories generated by various models for an image sequence from the VWP validation set. The red colored text reflects the inconsistencies in stories with regards to entities and emotions, and the green colored text highlights aspects of stories that are well-grounded in the input visual sequence.

image-text data crawled from the web, that might be enabling the model to accurately ground stories across the temporal positions in image sequences.

To understand if the $\mathbf{d}_{HM}$ method and the individual metrics reflect the quality of stories generated by models, we qualitatively inspected a random selection of the generated stories. Along the three dimensions considered, we observed that the current evaluation metrics reflect the overall quality of model-generated stories. However, we find that these metrics do not fully capture all the aspects relevant to visual storytelling. For instance, we find that the inconsistencies in stories regarding the overarching topic, the characters, and their emotions (see Figure 2) are not completely accounted for by the evaluation metrics.[4] Moreover, with the current set of evaluation approaches it is unclear how to accurately differentiate creative expressions that are visually grounded from implausible hallucinations. We believe these findings add evidence to the claims made by Surikuchi et al. [12] and emphasize the need for improving evaluation methods.

[4]Appendix C provides more examples.

5. Conclusion

In this work, we studied the visual storytelling task using the VWP dataset and underlined the various challenges concerning evaluation of model-generated stories. We compared various models using existing evaluation frameworks and showed that a general-purpose foundation model, Qwen2.5-VL, achieves the best overall scores along the dimensions of visual grounding and coherence. Qualitatively, we verify the generated stories and find that along the three dimensions considered, the evaluation methods reflect the quality of the models. However, in line with the latest studies, we also observed that the current automatic evaluation methods do not fully capture all the aspects essential for visual storytelling. Our findings support the need for research efforts toward automatic evaluation methods that approach the problem in a comprehensive manner for accurately assessing the quality of stories.

References

[1] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, M. Mitchell, Visual Storytelling, in: K. Knight, A. Nenkova, O. Rambow (Eds.), Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1233–1239. URL: https://aclanthology.org/N16-1147. doi:10.18653/v1/N16-1147.

[2] T. Kim, M.-O. Heo, S. Son, K.-W. Park, B.-T. Zhang, GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation, CoRR abs/1805.10973 (2018).

[3] X. Wang, W. Chen, Y.-F. Wang, W. Y. Wang, No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 899–909. URL: https://aclanthology.org/P18-1083. doi:10.18653/v1/P18-1083.

[4] C.-C. Hsu, Z.-Y. Chen, C.-Y. Hsu, C.-C. Li, T.-Y. Lin, T.-H. Huang, L.-W. Ku, Knowledge-Enriched Visual Storytelling, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 7952–7960. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6303. doi:10.1609/aaai.v34i05.6303.

[5] C.-y. Hsu, Y.-W. Chu, T.-H. Huang, L.-W. Ku, Plot and Rework: Modeling Storylines for Visual Storytelling, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 4443–4453. URL: https://aclanthology.org/2021.findings-acl.390. doi:10.18653/v1/2021.findings-acl.390.

[6] H. Chen, Y. Huang, H. Takamura, H. Nakayama, Commonsense Knowledge Aware Concept Selection For Diverse and Informative Visual Storytelling, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 999–1008. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16184. doi:10.1609/aaai.v35i2.16184.

[7] Y. Yu, J. Chung, H. Yun, J. Kim, G. Kim, Transitional Adaptation of Pretrained Models for Visual Storytelling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12658–12668.

[8] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[9] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or

Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909.

[10] E. Wang, C. Han, J. Poon, RoViST: Learning Robust Metrics for Visual Storytelling, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 2691–2702. URL: https://aclanthology.org/2022.findings-naacl.206. doi:10.18653/v1/2022.findings-naacl.206.

[11] A. K. Surikuchi, S. Pezzelle, R. Fernández, GROOViST: A metric for grounding objects in visual storytelling, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 3331–3339. URL: https://aclanthology.org/2023.emnlp-main.202/. doi:10.18653/v1/2023.emnlp-main.202.

[12] A. K. Surikuchi, R. Fernández, S. Pezzelle, Not (yet) the whole story: Evaluating visual storytelling requires more than measuring coherence, grounding, and repetition, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 11597–11611. URL: https://aclanthology.org/2024.findings-emnlp.679/. doi:10.18653/v1/2024.findings-emnlp.679.

[13] X. Hong, A. Sayeed, K. Mehra, V. Demberg, B. Schiele, Visual Writing Prompts: Character-Grounded Story Generation with Curated Image Sequences, Transactions of the Association for Computational Linguistics 11 (2023) 565–581. URL: https://aclanthology.org/2023.tacl-1.33. doi:10.1162/tacl_a_00553.

[14] X. Hong, K. Mehra, A. Sayeed, V. Demberg, Visually Grounded Story Generation Challenge, in: S. Mille (Ed.), Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges, Association for Computational Linguistics, Prague, Czechia, 2023, pp. 17–22. URL: https://aclanthology.org/2023.inlg-genchal.3.

[15] Q. Huang, Y. Xiong, A. Rao, J. Wang, D. Lin, MovieNet: A Holistic Dataset for Movie Understanding, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 709–727.

[16] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[17] R. Vedantam, C. Lawrence Zitnick, D. Parikh, CIDEr: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[18] K. S. Smith, W. Aziz, L. Specia, Cohere: A toolkit for local coherence, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 4111–4114. URL: https://aclanthology.org/L16-1649/.

[19] X. Hong, V. Demberg, A. Sayeed, Q. Zheng, B. Schiele, Visual coherence loss for coherent and visually grounded story generation, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 9456–9470. URL: https://aclanthology.org/2023.findings-acl.603/. doi:10.18653/v1/2023.findings-acl.603.

[20] X. Hong, A. Sayeed, V. Demberg, Summary of the visually grounded story generation challenge, in: S. Mille, M.-A. Clinciu (Eds.), Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges, Association for Computational Linguistics, Tokyo, Japan, 2024, pp. 39–46. URL: https://aclanthology.org/2024.inlg-genchal.3/.

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[22] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama,

R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 28, Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.
[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners, OpenAI blog (2019). URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
[24] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. arXiv:2307.09288.
[25] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 34892–34916. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.
[26] Q. Team, Qwen2.5-vl, 2025. URL: https://qwenlm.github.io/blog/qwen2.5-vl/.
[27] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, Y. Sun, C. Deng, H. Xu, Z. Xie, C. Ruan, Deepseek-vl: Towards real-world vision-language understanding, 2024. arXiv:2403.05525.
[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022.
[29] A. K. Surikuchi, R. Fernández, S. Pezzelle, Natural language generation from visual sequences: Challenges and future directions, 2025. URL: https://arxiv.org/abs/2502.13034. arXiv:2502.13034.

A. Models

For off-the-shelf VLMs, we use LLaVa version 1.6, Qwen2.5-VL-7B-Instruct, and DeepSeek-vl-7b-chat models, and access them using the HuggingFace transformers library.[5] We note that the VLMs follow the model architecture presented in Surikuchi et al. [29]. The specific vision encoder and language decoder components of these VLMs are outlined in Table 1.

Table 1
Pre-trained vision encoders and language decoders of the VLMs.

	Vision Encoder	Language Decoder
LLaVAv1.6	CLIP-ViT-L-336px	Mistral-7B
Qwen2.5-VL	ViT	Qwen2.5
DeepSeek-VL	SigLIP, SAM-B	DeepSeek

B. Training and Inference

For generating stories using off-the-shelf VLMs, we used the following prompts:

[5] https://huggingface.co/docs/transformers/

```
P1 = 'Write a story using exactly [num-images-in-the-sequence] sentences for
    this image sequence. Do not use more than [num-images-in-the-sequence]
    sentences.'
P2 = 'Generate a story consisting of [num-images-in-the-sequence] sentences for
    this image sequence. Use only [num-images-in-the-sequence] sentences and not
    more.'
P3 = 'Output a story about this sequence of images using only
    [num-images-in-the-sequence] sentences. Make sure the story does not include
    more than [num-images-in-the-sequence] sentences.'
```

Number of images in the input sequences of the VWP dataset ranged between 5 to 10 images and the prompts were adapted accordingly during inference.

C. Generated Stories

Human Story	A man ducks for cover. There are bullet holes in the wall near him, and he covers his face in fear. He tries to get away, but another person grabs him from behind . He holds him in place as the man struggles to get away . Suddenly, the bullets start flying. Both men duck or dive for safety. Food and bits of plaster go flying as the bullets fill the room. Two men are hit, and they both go flying back.
Qwen2 . 5-VL	In the opulent room, a man in a tuxedo was playing the accordion, oblivious to the chaos unfolding around him. Suddenly, another man burst into the room, his face contorted with anger as he lunged at the first man. The second man's attack was met with a swift counter from the first, who managed to push him back. As they struggled, a cloud of dust filled the air, obscuring their movements. In the midst of the commotion, a third man appeared, his expression one of shock and fear as he watched the two men grapple.

Human Story	[female0] ran towards [male0] after she saw him fall down. She asked him if he was okay and if he could hear her. [male0] laid on the ground as the pool of blood surrounded him. He was fading fast due to the loss of blood. Finally the paramedics arrived to attend to him.
Qwen2 . 5-VL	The emergency responders hurriedly attended to the injured man, their faces reflecting the urgency of the situation. Nearby, a woman watched with a mix of fear and determination, her eyes wide as she tried to understand what was happening. The man lay unconscious on the ground, his body showing signs of severe trauma. As the paramedics worked quickly, another person knelt beside him, offering silent support and comfort. Despite the chaos, there was a sense of coordinated effort as they prepared to transport the injured man to safety.

Figure 3: Stories generated by the Qwen2 . 5-VL model for image sequences from VWP validation set.