

# Summary of the Visually Grounded Story Generation Challenge

Xudong Hong<sup>13</sup>, Asad Sayeed<sup>2</sup>, Vera Demberg<sup>13</sup>

<sup>1</sup>Dept. of Language Science and Technology and Dept. of Computer Science, Saarland University

<sup>2</sup>Dept. of Philosophy, Linguistics, and Theory of Science, University of Gothenburg

<sup>3</sup>Saarland Informatics Campus, Saarbrücken

{xhong, vera}@lst.uni-saarland.de  
asad.sayeed@gu.se

## Abstract

Recent advancements in vision-and-language models have opened new possibilities for natural language generation, particularly in generating creative stories from visual input. We thus host an open-sourced shared task, Visually Grounded Story Generation (VGSG), to explore whether these models can create coherent, diverse, and visually grounded narratives. This task challenges participants to generate coherent stories based on sequences of images, where characters and events must be grounded in the images provided. The task is structured into two tracks: the Closed track with constraints on fixed visual features and the Open track which allows all kinds of models. We propose the first two-stage model using GPT-4o as the baseline for the Open track that first generates descriptions for the images and then creates a story based on those descriptions. Human and automatic evaluations indicate that: 1) Retrieval augmentation helps generate more human-like stories, and 2) Large-scale pre-trained LLM improves story quality by a large margin; 3) Traditional automatic metrics can not capture the overall quality.<sup>1</sup>

## 1 Introduction

Vision-based language generation (VLG) is the generation of text from visual input and is an important task in natural language generation and artificial intelligence. Recently, large pre-trained vision-and-language models (VLMs), such as GPT-4 (OpenAI, 2023) and Gemini (Reid et al., 2024), have achieved remarkable performance across several multimodal tasks, including image captioning (Vinyals et al., 2016), visual question answering (Goyal et al., 2017), and visual dialogue generation (Das et al., 2017).

Although these advancements are notable, most of the current tasks involve predicting labels or

generating short pieces of text (typically under 30 words). It remains uncertain whether the latest VLMs can create longer, coherent texts consisting of multiple sentences based on visual input. The evaluation of long stories is still challenging (Min et al., 2023). On the other hand, humans can easily generate extended and logically connected text from visual stimuli. To further assess VLMs, a task more aligned with human capabilities is necessary (Bubeck et al., 2023).

Previous tasks have been designed to evaluate the ability of VLMs to produce more extended outputs, 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 primarily focus on literal descriptions, where sentences remain independent rather than forming a coherent whole. Coherence, especially local coherence—defined as the relationships between entities in a given context—is fundamental to human language comprehension and production. In vision and language research, local coherence is crucial for several reasons: **1.** Improved models of local coherence can enhance the performance of vision-language tasks, such as text-to-image retrieval (Park and Kim, 2015). **2.** Accurately modeling coherence is essential for developing event knowledge, as events revolve around entities. Stronger event modeling enhances vision-language pre-training (Zellers et al., 2021, 2022).

Story generation is a widely researched task in natural language generation and is frequently used to assess whether large pretrained models can track entities (Paperno et al., 2016) and produce locally coherent texts. Unlike image captions, stories involve multiple characters and events, with recurring entities interacting with one another and their surroundings. Moreover, the importance of characters and relevant content is central to successful story creation (Goldfarb-Tarrant et al., 2020). We contend that story generation is an appropriate bench-

<sup>1</sup>Source code and pre-trained models are available at <https://vgsg2024.github.io/>

### Visual Writing Prompts (Ours)

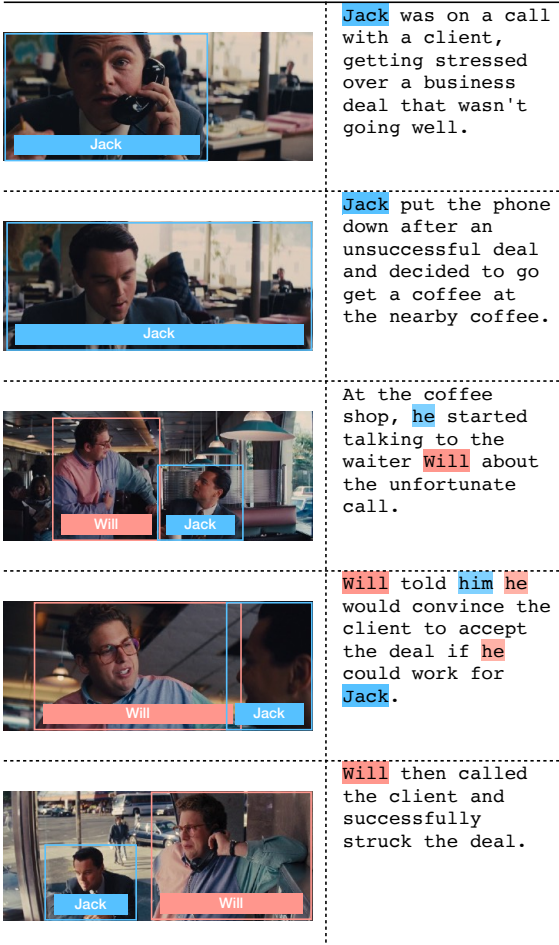


Figure 1: Example of Visual Grounded Story Generation on Visual Writing Prompts dataset. The dataset has recurring characters across all five images and sub-stories. Each occurrence of a character in a sub-story has a bounding box in the corresponding image, which grounds the textual appearance to visual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

mark for testing the ability of VLMs to generate coherent text.

In response, we introduce a new shared task called Visually Grounded Story Generation (VGSG), which challenges VLMs to generate coherent, diverse, and visually grounded stories. This task presents two primary challenges: **1.** The characters in the stories must be grounded in the images, meaning their actions and descriptions should align with the visual information provided. **2.** The generated stories must be coherent, with a clear beginning, middle, and end, and maintain a logical progression from one sentence to the next. Our goal is to identify the pros and cons of the current VLMs and automatic metrics on this task.

We conduct both automatic and human evaluations. For automatic evaluations, we mainly em-

ploy traditional metrics, including 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), to set up an efficient standard evaluation pipeline for this task. We also follow Hong et al. (2023b) to create a solid human evaluation across properties for good stories including Coherence, Diversity, Grammaticality, Visual Grounding, and Overall quality.

Our major findings are 1) Retrieval augmentation based on visual input similarities aids in generating more human-like stories; 2) Large-scale pre-trained language models significantly enhance story quality in that proprietary models with large-scale pre-training are still difficult to outperform using smaller models; and 3) Traditional automated metrics are inadequate in assessing overall quality because they do not correlate with human judgments.

Through this shared task, we hereby call for further research on visually grounded story generation, especially on the evaluation of the excessively long output from the models with large-scale pre-training.

## 2 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.

### 2.1 Datasets

We use four datasets for evaluation, two of which provide grounding annotations for characters. One of these is our own Visual Writing Prompts dataset: **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.<sup>2</sup> Each image corresponds to a part of a story. Instances of each protagonist are annotated with the character’s name (see Figure 1).

We follow Hong et al. (2023b) to use the default data split, that is 11778 for train, 849 for validation,

<sup>2</sup><https://vwprompt.github.io/>

and 586 for test<sup>3</sup>.

**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 VIST dataset (Huang et al., 2016), along with an importance rating of all characters in any story.<sup>4</sup> We only use it for evaluation.

We also evaluate on these datasets:

**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.

## 2.2 Tracks

We ran two evaluation tracks for this task:

**Closed Track** focuses on exploring Language and Vision Mapping methods and Language Generation models through a controlled experiment where the visual encoder is fixed. We provide extracted visual features from a pre-trained vision model. Participants must use these features as input (instead of raw images) to train their models on the provided dataset.

**Open Track** aims to test the state-of-the-art on 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.

## 3 Evaluation and Results

In this section, we describe our designs for both automatic and human evaluations for the submissions. The scripts for all automatic metrics be provided after the submission system is open; human evaluation be conducted after all submissions have been received. We release the annotator instructions and source code of all metrics after the shared task.

### 3.1 Automatic Evaluation

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

<sup>3</sup>Please contact the authors for details on the other datasets and how they are applied during the evaluation.

<sup>4</sup><https://github.com/iz2late/VIST-Character>

**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). In our initial proposal, we planned to use BERTScore (BS; Zhang et al., 2020) which is effective in text summarization. Unfortunately, we did not have enough resources to run it by ourselves, because it requires usage of a large amount of GPU time.

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

**Diversity** We use metrics used by Hong et al., 2023b 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 use the generative Entity Grid model to calculate the log-likelihood based on entity transitions in system outputs.

### 3.2 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 also conducted a human evaluation for the submissions, focused 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 encouraged participants to perform their own human evaluation and include the results in their reports.

### 3.3 Baselines

We employ two models as baselines for each track.

**EntityGrid** (Hong et al., 2023b) is the baseline for Closed track. It is a Transformer-based model that adapts the visual features with pre-trained GPT-2.

Team	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
baseline	EntityGrid	37.12	13.86	7.33	3.96	34.27	14.78	0.65
team-DMG	LLaVA-S	35.03	14.08	7.90	4.07	34.02	12.16	0.88

Table 1: Performance comparison of different teams on Closed track. All numbers are the higher the better.

Team	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
baseline	GPT-4o	20.71	1.52	0.07	0.00	14.21	10.88	1.21
HTWK	GPT4-RA	19.39	1.47	0.03	0.00	12.53	10.70	0.92
team-DMG	LLaVA-O	22.28	2.56	0.14	0.00	18.09	13.51	1.64

Table 2: Performance comparison of different teams on Open track. †We observed extremely low numbers on BLEU-4. All numbers are the higher the better.

**GPT-4-GPT-4o** (OpenAI, 2023) is the baseline for Open track.

### 3.4 Teams and Models

There are two teams that participated in our tasks. One team participated in the Open track only, and the other team participated in both.

**HTWK** is a team from Leipzig University of Applied Sciences, Germany. They only participated in the Open track. They employ two similarity retrievers to find semantically closest samples from the training set, which serve as examples for the multimodal generative model. First, an image similarity retriever identifies the most similar images for each image in the input sequence. A prompt is then constructed using the retrieved images along with their descriptions, which are provided as examples for the model to generate descriptions for each image. Next, the method concatenates all the generated descriptions and uses a textual similarity retriever to find the most semantically related story. This story serves as the example in the prompt, guiding the model to generate a coherent and reasonable narrative for the input sequence of images.

**team-DMG** is a team from the University of Amsterdam, Netherlands. They participated in both tracks. For the Closed track, they proposed an updated version of the TAPM model (**LLaVA-S**). To enhance TAPM’s performance while maintaining a lower parameter count, they replaced the original language model with LLaVA, a state-of-the-art large language model, and adapted the visual encoder accordingly. They utilized a 4-bit quantized version of LLaVA and fine-tuned it using the LoRA approach, focusing on the multi-head self-attention blocks. Additionally, they improved the vision component by supplementing ResNet-101 features with representations extracted from a pre-trained

Vision Transformer (ViTbase) model.

For the Open track, they use a fine-tuned LLaVA model (**LLaVA-O**), which is a general-purpose multimodal foundation model similar to BLIP-2. However, instead of focusing on model architecture, LLaVA emphasizes training data and procedure. It is notable for extending instruction-tuning to the language-image multimodal space by training on vision-language instruction-following data. This data is constructed by querying GPT-4 with various in-context-learning prompts to generate <image, caption> pairs from existing datasets like COCO. LLaVA connects visual features with language embeddings using a single linear layer, unlike BLIP-2, which uses Q-Former. The team uses LLaVA to generate stories in a zero-shot manner under different linguistic context settings.

### 3.5 Automatic Metrics

Here we summarize the results for the Closed and Open tracks in the tables above.

In the Closed track (Table 1), team-DMG’s LLaVA-S model outperforms the baseline EntityGrid model in terms of CIDEr (0.88 vs. 0.65) and BLEU scores, with notable improvements in BLEU-2, BLEU-3, and BLEU-4, although both models perform similarly in METEOR and ROUGE-L. While team-DMG’s submission shows competitive performance, the overall improvement in BLEU and CIDEr suggests that the submissions are gradually advancing beyond the baseline’s entity-based approach.

For the Open track (Table 2), team-DMG’s LLaVA-O model also surpasses the baseline GPT-4o model, achieving the highest scores in BLEU-1, BLEU-2, and ROUGE-L, as well as a significantly better CIDEr score (1.64 vs. 1.21). In comparison, HTWK’s GPT4-RA performs slightly lower

	Model	Coherence	Diversity	Grammaticality	Grounding	Overall
Closed	baseline (EntityGrid)	1.53	2.30	3.13	2.17	1.47
	team-DMG (LLaVA-S)	1.72	2.85	2.98	1.74	1.47
Open	baseline (GPT-4o)	4.35	3.76	4.90	4.31	3.65
	HTWK (GPT4-RA)	4.04	3.65	4.94	3.94	3.29
	team-DMG (LLaVA-O)	1.41	2.67	3.08	1.47	1.33

Table 3: Human evaluation of teams in both tracks. Higher numbers are better for all measures.

across all metrics, trailing behind both the baseline and team-DMG in key metrics such as METEOR and CIDEr. Notably, despite the improvements, all systems in both tracks still perform poorly in higher-level BLEU metrics (BLEU-3, BLEU-4), indicating challenges in producing more refined n-gram matches.

Overall, while team-DMG’s models consistently improve over the baselines in both tracks, there remains room for further advancements, particularly in terms of the more nuanced and detailed metrics like higher-order BLEU scores. Additional analysis may be needed to explore why these improvements are not more pronounced across all metrics.

### 3.6 Human Evaluations

The human evaluation results for both the Closed and Open tracks reveal notable differences in system performance across various metrics, including Coherence, Diversity, Grammaticality, Grounding, and Overall scores.

Table 3 presents a comparison of different models evaluated on several criteria under both Closed and Open settings. For the Closed setting, team-DMG (LLaVA-S) achieves a slight improvement in terms of Coherence (1.72) and Diversity (2.85) compared to the baseline (EntityGrid), although both models achieve the same overall score (1.47). Grounding scores remain relatively low for both models in this setting, with team-DMG scoring 1.74 and EntityGrid slightly higher at 2.17.

In the Open setting, the baseline model (GPT-4o) outperforms all other models in nearly every category, with a Coherence score of 4.35, Grammaticality of 4.90, and Grounding of 4.31. HTWK (GPT4-RA) follows closely with slightly lower Coherence (4.04) and Grounding (3.94), but surpasses GPT-4o in Grammaticality (4.94). In contrast, team-DMG (LLaVA-O) shows lower scores across all metrics, particularly in Coherence (1.41) and Grounding (1.47), resulting in the lowest overall score of 1.33.

These results highlight that while team-DMG

demonstrates some advantages in Diversity and Coherence under Closed conditions, the Open setting models show a clear dominance of GPT-4o and GPT4-RA, particularly in Grammaticality and Grounding. Overall, the baseline models perform better in terms of general language quality, while team-DMG struggles to match their performance, especially in the Open setting.

### 3.7 Case Study

We also conduct case study to inspect the generated stories. The results suggest that models like GPT-4o and GPT4-RA are more adept at balancing narrative coherence, character interaction, and environmental immersion, making them suitable for tasks that require rich storytelling and visual grounding. GPT-4o generates stories that are visually grounded on characters while GPT4-RA generates stories that are more grounded on environments. In contrast, models like the baseline EntityGrid and team-DMG LLaVA-O face challenges with redundancy, coherence, and character development. Future improvements could focus on refining character relations and enhancing logical event progression to further bridge the gap between generated narratives and human storytelling.

## 4 Related Work

**VLG with Coherence.** One relevant VLG 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 Storytelling is a task that requires local coherence, where the input is a sequence of images and the output is a coherent story (Huang et al., 2016). However, it is not specif-






						R-L	C	D	G	V	O
<b>baseline-EntityGrid</b>	Liam was a soldier that was being held by a soldier . the enemy soldiers were shooting at the enemy soldiers . the soldiers are on top of a village trying to find the enemy . the enemy soldiers were hiding in the middle of the area trying to get them to get out . Oliver , Elijah and James are trying to get the best way to get the next move .					26.53	1.0	2.5	3	2.5	1
<b>team-DMG-LLaVA-S</b>	Liam was checking that the gunfire was over . one soldier started fired shots at the other . he knew they were in trouble . suddenly there was an explosion and people started screaming . he knew it was the right time to escape .					24.54	4.5	4.5	3.5	4	3.5
<b>baseline-GPT-4o</b>	The air was thick with smoke and the acrid scent of gunpowder as Sergeant Jameson led his squad through the war-torn streets. The once bustling city was now a desolate battlefield, with ruined structures and debris marking the path of destruction. Jameson, his helmet slightly askew, carried a rifle slung over his back, his eyes scanning the horizon for any sign of enemy movement. Behind a makeshift barricade of sandbags, Corporal Davis and Private Miller crouched, their faces smeared with dirt and sweat. Davis, gripping a rocket launcher, took aim at a distant tank that rumbled ominously through the haze. Miller, his rifle at the ready, kept a vigilant watch, his heart pounding in his chest. ...					12.01	4.0	4.5	5	4.5	4
<b>HTWK-GPT4-RA</b>	Amidst the chaos of a war-torn city, soldiers in combat gear surveyed the battlefield, shrouded in smoke and debris from recent explosions. Taking cover behind sandbags, they prepared for the next wave of combat, their eyes scanning the ruins for any signs of movement. The air was thick with tension as they navigated through the rubble, dodging flames and damaged military vehicles. Explosions echoed in the distance, urging them to stay alert. With every step, they braced for the unpredictable nature of battle, determined to survive the destruction that surrounded them and protect their comrades in this relentless conflict.					13.62	4.5	4.5	5	4.5	4.0
<b>team-DMG-LLaVA-O</b>	Liam and Noah are getting ready to leap a large helicopter from a field that is not already in the air . the soldiers were all aiming to the direction of Olivia . the soldiers were ready to land and go into battle . Noah and Liam look out over the soldiers who are dead . they have fighting to do however and do not have time to go and see him .					16.17	1.5	3	3.5	1.5	2.0

Figure 2: Case study of generated/human-written stories given an example image sequence. We present the ROUGE-L (R-L) for automatic evaluation. We also report the human evaluation scores of Coherence (C), Diversity (D), Grammaticality (G), Visual Grounding (V), and Overall (O) quality.

ically designed to ensure that there are a tellable event sequences depicted in the image sequences, because the dataset used in this task, VIST, is based on photo albums.

**Visual Story Generation.** Most of the previous tasks for visual story generation have major 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.

**Story Generation** There are several existing datasets for generating a story conditioned on a prompt such as previous context (Mostafazadeh et al., 2016), title (Fan et al., 2018), keyword (Yao et al., 2019), cue phrase (Xu et al., 2020), script (Pu et al., 2022), story plot (Rashkin et al., 2020), or detailed plots (Akoury et al., 2020). However, all these datasets relying on textual prompts suffer from the grounding problem that the meanings of textual stories are grounded on textual symbols

(Harnad, 1990).

## 5 Conclusions

We organized the Visually Grounded Story Generation task (VGSG) this year for the first time. Although Visual Language Models have made huge progress in the past couple of years, they are generally not specifically designed with the intention of producing narratively coherent and grounded stories. This task provided further impetus for development in this area. We obtained a couple of submissions, although private communications suggested that other potential participants instead decided to retain their results for publication in other venues. The training data and test platform were mounted on the web and on HuggingFace in order to enable further progress.

Consider the automatic measures: in the controlled (Closed) experiment, the use of a LLaVA model produced what appear to be modest improvements, although the significance of the result is not clear. However, in the Open experiment, it turned out to be difficult to beat GPT-4o with another GPT model, but using a LLaVA-based language model brought noticeable improvements.

However, it should be noted that automatic eval-

uations do not track the human evaluations. The submission made no difference in the Closed track to the overall human evaluation, and the best system in automatic evaluation had the worst outcome in human evaluation. This was reflected in our case study. This mismatch between automatic and human evaluation highlights the need for better automatic measures and for future work on this topic to "go the extra mile" and produce robust human evaluations.

## References

- Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. *Storium: A dataset and evaluation platform for machine-in-the-loop story generation*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484.
- 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.
- 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)*.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.
- 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.
- Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.
- 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*, 11:565–581.
- 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.
- 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.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. *FActScore: Fine-grained atomic evaluation of factual precision in long form text generation*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- 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.

- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. [A corpus and cloze evaluation for deeper understanding of commonsense stories](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.
- 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.
- Dongqi Pu, Xudong Hong, Pin-Jie Lin, Ernie Chang, and Vera Demberg. 2022. [Two-stage movie script summarization: An efficient method for low-resource long document summarization](#). In *Proceedings of The Workshop on Automatic Summarization for Creative Writing*, pages 57–66, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. [PlotMachines: Outline-conditioned generation with dynamic plot state tracking](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295, Online. Association for Computational Linguistics.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- 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.
- Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. [MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.
- Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.
- 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*.