information

MDPI

Article

Story Generation from Visual Inputs: Techniques, Related Tasks, and Challenges

Daniel A. P. Oliveira [1,2,*], Eugénio Ribeiro [1,3] and David Martins de Matos [1,2]

[1] INESC-ID, 1000-029 Lisbon, Portugal; eugenio.ribeiro@inesc-id.pt (E.R.); david.matos@inesc-id.pt (D.M.d.M.)
[2] Instituto Superior Técnico, Universidade de Lisboa, 1649-004 Lisbon, Portugal
[3] Department of Information Science and Technology (ISTA), Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisbon, Portugal
* Correspondence: daniel.oliveira@inesc-id.pt

Abstract

Creating engaging narratives from visual data is crucial for automated digital media consumption, assistive technologies, and interactive entertainment. This survey covers methodologies used in the generation of these narratives, focusing on their principles, strengths, and limitations. The survey also covers tasks related to automatic story generation, such as image and video captioning, and Visual Question Answering. These tasks share common challenges with Visual Story Generation (VSG) and have served as inspiration for the techniques used in the field. We analyze the main datasets and evaluation metrics, providing a critical perspective on their limitations.

Keywords: Visual Story Generation; image captioning; visual question answering; storytelling

Academic Editor: Arkaitz Zubiaga

Received: 23 August 2025
Revised: 10 September 2025
Accepted: 16 September 2025
Published: 18 September 2025

Citation: Oliveira, D.A.P.; Ribeiro, E.; Martins de Matos, D. Story Generation from Visual Inputs: Techniques, Related Tasks, and Challenges. *Information* **2025**, *16*, 812. https://doi.org/10.3390/info16090812

1. Introduction

In this document, we survey the field of story generation from visual inputs, covering techniques, related tasks, and challenges. Stories are fundamental to human experience, serving as a bridge between imagination and reality [1]. They represent the art of telling tales, whether real or imagined, and present a sequence of events that can captivate, inform, and provoke thought. Storytelling has been an essential part of human culture for millennia, serving as a means of communication, entertainment, education, and moral guidance. The art of storytelling has evolved over time, beginning with the practice of narrating stories through spoken words, advancing to written literature, and, more recently, transitioning to digital media formats. Stories have played a crucial role in shaping societies, passing down knowledge and values from one generation to another. Whether told around a campfire, written in books, or shared through videos and interactive platforms, storytelling continues to captivate and connect people, transcending time and technology [2]. The field of story generation has gained attention, fueled by advances in Natural Language Processing (NLP) [3] and computer vision [4]. The goal is to create systems capable of generating novel and engaging stories, similar to those produced by humans [5]. With the increasing availability of images and videos, there is a growing interest in generating stories based on this visual data [6]. We refer to this task as VSG.

1.1. Definition and Challenges

VSG is the task of automatically creating coherent and engaging stories grounded on visual inputs, such as images or videos. It extends beyond simple image captioning or

video description, requiring an understanding of complex relationships, temporal progression, and implicit context within the visual data. The generated narrative must not only accurately represent the visual content, but also encapsulate the essence of the underlying story, thereby crafting an engaging experience for the audience. VSG serves as a robust indicator of machine comprehension of visual content and poses several challenges, some of which are as follows:

- Understanding visual content: The recognition of objects and scenes in the visual data, as well as understanding the relationships between them, is crucial for generating coherent and meaningful narratives [7].
- Temporal coherence: A story is inherently sequential and maintains a temporal order. Therefore, generated narratives should respect time progression, even if not in a linear manner (e.g., they may involve flashbacks, foreshadowing, or other such mechanisms), and ensure that events are narrated in a logically consistent manner.
- Contextual understanding: Often, elements of a story can be implicit or inferred from the context. Effective generation approaches must be capable of understanding and incorporating implicit information in the narrative.
- Engagement and creativity: One of the main goals of a story is to engage the audience. Therefore, the generated narrative should not only be accurate and coherent, but also creative and engaging.
- Evaluation: Determining the effectiveness of a VSG system is a challenge in itself. Metrics originally developed for tasks such as machine translation and image captioning have been adapted for VSG. However, these metrics are not appropriate to capture the creativity and engagement aspects of a story.

1.2. Background and Motivation

The motivations for VSG are numerous. For example, VSG can provide an interactive component to media consumption. Users could potentially influence the direction of the story by choosing different sequences of images or videos, creating a unique, personalized narrative. This could be particularly beneficial in areas like interactive gaming and movies, educational platforms, or even in digital marketing, where personalized narratives can greatly enhance user engagement and satisfaction. VSG can also be used in journalism: it can be used to summarize large amounts of visual data, giving insights into the underlying patterns and relationships, and assisting in content creation.

The growing interest in this area is also motivated by its potential to contribute to the understanding of human cognition and creativity. By developing models that emulate human-like storytelling abilities, researchers can gain insights into the cognitive processes involved in story creation, as well as explore the underlying mechanisms of creativity. These approaches can further lead to novel applications, such as more engaging virtual experiences [8,9].

These use cases represent only a fraction of the potential applications of VSG. As research in this field progresses, we anticipate that more applications will emerge, spanning an even broader range of domains.

1.3. Scope and Structure of the Survey

We start by defining the core elements of a story in Section 2, as they are essential for understanding the structure and dynamics of a story. Section 2 presents these elements and discusses their significance in storytelling. It also clarifies the distinctions between story, narrative, and plot, illustrating how a single story can be narrated in multiple ways to yield diverse experiences.

In Section 3, we then explore the main datasets used in story generation, and the various metrics used to assess their performance. We address the limitations that current evaluation metrics face in such an open-ended task and discuss potential directions for future research in this regard.

The connection of VSG with computer vision tasks such as Visual Question Answering (VQA) and image/video captioning is also acknowledged, since they share core challenges, such as visual understanding and narrative generation. Thus, they have contributed with useful techniques to the field. Section 4 explores this convergence between computer vision and NLP.

Concerning VSG proper, we aim to present a balanced perspective on current methodologies, emphasizing opportunities and challenges for future research. The inclusion of a detailed analysis of deep learning methodologies with a broader review of additional and earlier techniques ensures broad coverage of the field. Section 5 presents a broad perspective on the state of the art of VSG.

Section 6 examines the real-world applications of VSG. Finally, Section 7 presents the key findings of this survey and proposes directions for future research.

2. The Elements of Stories

Stories connect events and experiences through narration, ranging from personal tales to cultural epics. While the terms story, narrative, and plot are often used interchangeably, they have distinct meanings: a story is the chronological sequence of events (the "what"), the narrative is how these events are presented (the "how"), and the plot is the specific sequence as experienced by the audience [10].

A single story can be told through multiple narratives, each offering different perspectives and experiences. Narratives may follow linear structures with chronological progression, or nonlinear approaches using flashbacks, flash-forwards, or parallel storylines that create complexity and require audience interpretation [11,12].

Story elements—character, conflict, theme, setting, plot, and mode—form the foundation for crafting and analyzing narratives. Understanding these elements is crucial for both human storytellers and automated story generation systems. In the following subsections, we detail each element and its significance in VSG.

2.1. Character

Characters are the entities through which stories unfold, with their decisions and actions propelling the plot forward [13]. Through actions and speech, characters convey motives and emotions that shape audience perception and emotional connection. Initial character impressions influence audience expectations about future behavior. Characters who deviate from established patterns can create surprise or confusion, challenging audience understanding. Stories center around protagonists who encounter central conflicts, undergo personal growth, or acquire new knowledge. Antagonists add complexity and tension, often opposing the protagonist's goals. Traditional narratives feature heroes battling for moral causes against villains who perpetrate evil. Character development and evolution throughout the narrative engages audiences and drives story progression. The range of character types contributes to story depth and dynamics.

2.2. Conflict

Conflict generates tension and drives character actions, shaping narrative trajectory and audience engagement [2]. How conflicts are introduced, developed, and resolved determines story progression and emotional depth. Conflict resolution provides closure to the central problem and marks the culmination of narrative tension. Conflict encompasses

any tension influencing characters, from the protagonist's main challenge to secondary internal struggles like anxiety or indecisiveness. These layers add complexity to character journeys and story depth. Longer narratives involve multiple conflicts that occur alongside the primary tension. Common conflict types include the following: Character vs. Character (opposition), Character vs. Nature (natural forces), Character vs. Society (social norms), Character vs. Unavoidable Circumstances (fate), and Character vs. Self (internal dilemmas).

2.3. Theme

Themes are the major underlying ideas of a story, often abstract and open to interpretation [14]. Unlike concrete story elements, themes invite discussion and allow for multiple interpretations and conclusions. They function as exploratory concepts that encourage audience reflection on deeper meanings.

Themes can evolve as narratives unfold, revealing new dimensions and allowing audiences to develop personal interpretations of the story's messages. Different audience members may reach varying conclusions that differ from or exceed the creator's original intent. Understanding themes in narrative generation creates dynamic, thought-provoking storytelling experiences.

2.4. Setting

Setting encompasses the time, place, and context in which the story unfolds [15]. It includes physical and temporal surroundings, as well as social or cultural conventions that influence character actions and decisions. Setting can act dynamically, possessing specific traits and undergoing changes that affect the plot. This allows it to evolve throughout the story, reflecting and impacting narrative progression. Hogwarts School in the "Harry Potter" series exemplifies an evolving setting that influences plot and character development [16].

2.5. Plot

Plot is the sequence of events that propels the story from beginning to end [17]. Each plot point represents a moment of change that influences character understanding, decisions, and actions. This progression typically involves protagonists encountering conflict and attempting resolution, creating emotional stakes and suspense for both characters and audience. Plot advancement deepens audience investment in character fate. Stories can have multiple plots representing different event sequences or perspectives. Plot structure follows cause and effect, where character actions and events lead to reactions and consequences, forming a chain that moves the story forward.

2.6. Mode

Mode encompasses the choices and techniques authors use to frame narratives [18]. A single story can be narrated using different modes, each offering distinct audience perspectives and experiences. Mode includes information scope, language style, medium, and the extent of narrative exposition and commentary. A key component is point of view or perspective. First-person narratives use personal pronouns like "I" and "me," directly engaging audiences with the character's perspective. Third-person narratives avoid personal pronouns, offering more detached viewpoints. These perspectives exemplify narration techniques that explicitly tell stories through spoken or written commentary. While narration is common in written narratives, it can be optional in visual or interactive media.

3. Benchmark Datasets and Evaluation Metrics

This section presents and discusses benchmark datasets used for story generation and evaluation metrics that assess the performance of story generation models. Table 1 provides a compilation of the datasets documented in the literature for VSG. Among these datasets,

VIST [19] is prominently utilized by the research community. This dataset serves as a primary benchmark, and most publications in this field report their experimental results on it. The VIST-Edit dataset [20] provides human-edited versions of machine-generated stories, offering insights into human preferences for story quality. The Visual Writing Prompts dataset [21] is noteworthy for its focus on stories that are visually grounded on the characters and objects present in the images. These datasets allow for the exploration of more complex and imaginative narrative structures, serving as a resource to train models to generate fictional stories.

Table 1. Visual Story Generation datasets, key works, and methods.

Dataset	No. of Stories	Key Works	Notes
NY City [22]	11,863	Park et al. [22] (VGG + BiRNN)	11,863 blog posts/78,467 images
Disneyland [22]	7717	Park et al. [22] (VGG + BiRNN)	7717 blog posts/60,545 images
LSMDC [23]	118,114	Rohrbach et al. [23] (CNN + LSTM), Yu et al. [24] (Faster R-CNN + GPT-2)	202 movies, video captioning
VIST [19]	50,000	Huang et al. [19] (AlexNet + GRU), Wang et al. [25] (ResNet-152 + BiGRU), Yu et al. [24] (Faster R-CNN + GPT-2), Zheng et al. [26] (Mask R-CNN + BERT + BART)	210,819 unique photos, primary VSG benchmark
VIST-Edit [20]	14,905	Hsu et al. [20] (CNN + RNN variants)	Human-edited machine-generated stories
Visual Writing Prompts [21]	12,000	Hong et al. [21] (Swin Transformer + GPT-2)	5 to 10 images/story, character-grounded
PororoQA [27]	27,328	Kim et al. [27] (CNN + LSTM)	Cartoon videos with scene/dialogue/QA pairs
YouCook2 [28]	2000	Zhou et al. [28] (3D-CNN + LSTM)	Average length of 5.26 min, cooking videos

The evaluation metrics, used to assess the performance of VSG models, provide quantitative measures of the quality of generated narratives. Most of these metrics were originally developed for machine translation or summarization but have been adopted by the community for other tasks such as image captioning and story generation: instead of comparing a generated text with a reference translation or summary, the generated text is compared with a reference caption in the case of image captioning, or with a reference story in the case of story generation. These metrics include Bilingual Evaluation Understudy (BLEU) [29], Self-BLEU [30], Metric for Evaluation of Translation with Explicit ORdering (METEOR) [31], Consensus-based Image Description Evaluation (CIDEr) [32], Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [33], Perplexity [3], and Bert Score [34]. These metrics guide the development and refinement of VSG models by offering insights into their linguistic fluency, textual diversity, and mostly lexical overlap with reference texts. They focus on specific linguistic and content-related aspects, providing quantitative measures for precision, recall, and consensus. However, they are limited by their focus on surface-level textual similarities rather than deep narrative alignment: since a story can be retold in various ways, employing different perspectives while preserving the original message, the corresponding generated text might deviate significantly from the reference text in terms of specific wording.

Huang et al. [19] conducted an analysis to evaluate how well BLEU and METEOR correlate with human judgment in the context of VSG according to different metrics. The

results in Table 2 show that METEOR, which places a strong emphasis on paraphrasing, has higher correlation with human judgment than BLEU. Nevertheless, in spite of METEOR showing the highest correlation with human judgment, this value is still low, suggesting that a large gap still exists between automatic metrics and human judgment.

Table 2. Correlations of automatic scores against human judgements, *p*-values in parentheses [19].

Metric	METEOR	BLEU
Pearson	0.22 (2.8×10^{-28})	0.08 (1.0×10^{-06})
Spearman	0.20 (3.0×10^{-31})	0.08 (8.9×10^{-06})
Kendall	0.14 (1.0×10^{-33})	0.06 (8.7×10^{-08})

To evaluate VSG models, additional metrics that capture the creative aspects of storytelling may be necessary. These metrics might assess the uniqueness, novelty, and diversity of generated narratives, ensuring that models produce stories that go beyond mere data memorization. Coherence metrics could delve deeper into the logical flow of events, ensuring that generated narratives maintain a consistent and plausible storyline.

The proliferation of diverse Large Language Models (LLMs), such as the GPT [35–38], Llama [39,40], Mistral [41,42], and Claude [43–46] families, has led to the development of systematic scoring systems, aiming for objective assessments across a wide array of language understanding tasks. These scoring systems include Multi-turn Benchmark (MT-Bench) [47] and Chatbot Arena [47], briefly described below. While they were initially developed for the evaluation of chatbots by human evaluators, they have been adapted to operate using a LLM as the evaluator achieving results close to those obtained by human evaluators.

As LLMs continue to improve, they show potential in replacing human annotators in many tasks [48]. Metrics to evaluate stories could be developed based on LLMs [47], possibly following the underlying principles discussed above.

MT-Bench is a benchmark tailored to test the multi-turn conversation and instruction-following capabilities of LLM-based chat assistants. It includes 80 high-quality multi-turn questions distributed across diverse categories such as writing, roleplay, reasoning, and knowledge domains. In the MT-Bench evaluation process, the user is presented with two distinct conversations, each generated by a different LLM-based chat assistant and is then tasked with deciding which assistant, A, B, or indicating a tie, better followed the instructions and answered the questions. For VSG evaluation, this framework could be adapted by using an LLM as an automated judge to compare stories generated by different models, evaluating them based on criteria such as narrative coherence, visual grounding, and creativity.

Chatbot Arena introduces a crowd-sourcing approach, fostering anonymous battles between LLM-based chatbot models. Users interact with two anonymous models simultaneously, posing identical questions to both and subsequently voting for the model that provides the preferred response. The Elo rating system [47] is employed to rank the models based on their performance in these battles. This rating system assigns a numerical score to each model based on their battle outcomes, adjusting these scores higher or lower after each encounter to reflect their relative ability to satisfy user queries effectively compared to their opponents. A similar automated evaluation system for VSG could employ an LLM judge to assess pairs of generated stories given the same image sequence, ranking models based on story quality across multiple dimensions without requiring human annotators. For VSG adaptation, MT-Bench-style prompts could evaluate stories on visual grounding, narrative coherence, and creativity dimensions. A VSG Chatbot Arena could present story pairs from identical image sequences for comparative ranking.

In summary, while current metrics provide insights into specific dimensions of VSG, the development of new metrics may be necessary to holistically evaluate creativity, coherence, and emotional engagement, ensuring a more comprehensive understanding of the capabilities and limitations of these models. While LLMs show promise as automated evaluators, their use requires addressing several challenges. Prompt sensitivity can affect consistency, model updates may change evaluation criteria, and training data biases may favor certain narrative styles. Subjective dimensions like creativity remain difficult to calibrate. Rigorous validation is needed, including correlation studies with human judgments, inter-annotator reliability testing, and robustness evaluation. Despite these limitations, LLMs offer potential for scalable evaluation frameworks if properly validated.

4. From NLP to Computer Vision

This section provides an overview of areas at the intersection of computer vision and Natural Language Processing (NLP) that are relevant to the topic of VSG, namely image and video captioning and Visual Question Answering (VQA). These tasks share a common challenge: they require the understanding of visual content and the ability to generate text based on that visual content. Some of the techniques used in these tasks have been adapted to VSG.

4.1. Image and Video Captioning

The transition from image and video captioning to VSG represents a shift from describing individual images or video segments to narrating image sequences or video segments. While the former focus on *what* is in images or videos, the latter also considers *why* and *how*, capturing the underlying narrative that links a series of images or video frames.

VSG models need to understand the temporal progression of events, causal relationships, and the ability to construct cohesive and engaging stories. VSG can also speculate on the motivations and emotions of characters since such interpretations do not contradict the visual content. Therefore, they build upon the foundational principles of image and video captioning, incorporating both the interpretation of visual data and the generation of coherent, contextually appropriate narratives.

The tasks of image and video captioning consist in generating descriptive, human-readable sentences that represent the content of an image or video, identifying the main components in a visual scene and their relationships, alongside the ability to express these details in natural language.

One notable contribution in this domain was the application of encoder–decoder architectures, known as the 'show and tell' approach [49–51]. This technique employs Convolutional Neural Networks (CNNs) as encoders to extract visual features. Specifically, 2D-CNNs are employed for image processing, where the model interprets two dimensions: height and width. On the other hand, for video processing, 3D-CNNs are typically used, where an additional dimension, time or depth, is taken into account to perceive the temporal dynamics in video sequences. Then, Recurrent Neural Networks (RNNs) or their improved variants, Long Short-Term Memory networks (LSTMs) [52] and Gated Recurrent Units (GRUs) [53], known for their ability to effectively model sequential data, are commonly used as decoders in image and video captioning tasks, enabling the generation of coherent and contextually meaningful textual captions and descriptions from the visual features extracted by the CNNs. This combination of CNNs and RNNs effectively bridges the gap between vision and language.

In the realm of image captioning, Transformers [54] have been used as encoders to process visual features extracted from images. Leveraging the self-attention mechanism inherent in Transformers, these models are able to capture relationships between visual

elements, attending to relevant regions in the image, and highlighting visual cues and their interdependencies [55–58]. Transformer models also demonstrate the ability to generate precise and elaborate textual descriptions, significantly improving the overall quality of image captions.

Similarly, Transformers have made contributions to video captioning, addressing the challenges associated with understanding and describing temporal dynamics in videos. By treating video frames as a sequential series of images, Transformers can leverage their attention mechanism to capture long-range dependencies across frames [59,60] This enables the models to map the temporal progression of events, infer causal relationships, and grasp the underlying narrative structure within the video.

The integration of Transformers in image and video captioning has yielded improvements that surpass the performance of traditional encoder–decoder architectures [57,59]. Transformers not only generate semantically accurate captions but also exhibit enhanced coherence and contextual relevance. Their attention mechanism allows for a incorporation of local and global context, enabling a more holistic understanding of the visual content and facilitating caption generation.

The use of Transformers in image and video captioning holds great promise, but there are challenges to be addressed. Efficient fusion of visual features with textual information [56], handling long-term dependencies in video sequences [59], and ensuring temporal coherence in caption generation for longer videos are among the ongoing research topics.

4.2. Visual Question Answering

Visual Question Answering (VQA) combines computer vision and NLP to enable the extraction of information from visual content, identifying objects, scenes, actions, and comprehending context, and provide meaningful answers to questions about that content. Both VQA and VSG involve understanding visual content and generating meaningful responses and the developments in VQA can be directly applied in VSG tasks. Through the use of VQA techniques, VSG models can gain valuable insights into the content and context of images and videos, improving the overall quality of their visual narratives.

Encoder–classifier deep learning VQA models encode visual and textual inputs and select an answer from a predefined set using a classification process. Image encoders, typically pretrained CNNs, extract visual features from the input images. Question encoders, often RNNs, processes the input questions, converting them into fixed-length latent representations. Fusion techniques combine the visual and textual features, creating a fused representation that captures their interactions that is then classified using Fully Connected layers, mapping them to possible answer classes. Traditional VQA models are trained using supervised learning with large-scale datasets, where image–question–answer triplets are provided as training examples. The parameters are optimized to minimize the loss between predicted answers and ground truth answers during training. Some examples are BUTD [51], that uses an attention-based model that combines bottom-up and top-down mechanisms to improve performance in image captioning and Visual Question Answering tasks, and "Show, Ask, Attend, and Answer" [61], that combines LSTMs with multiple attention layers.

The transition from encoder–classifier methods to Transformer methods in VQA was driven by the Transformer's ability to capture long-range dependencies and global contextual information. Transformers offered the potential to more efficiently capture the interactions between visual and textual modalities, enabling a more comprehensive understanding of the relationship between images and questions. In these VQA models, images and questions are encoded using Transformer encoders and self-attention mechanisms capture the dependencies between words and the visual features, allowing the model to

focus on relevant image regions while processing the question. The fused encoded representations are then fed into classification layers to generate answers [62–65]. Recent advances in VQA also explore new models that approach the task as a generation problem rather than a classification one, allowing for a more flexible and diverse range of answers [66].

5. Visual Story Generation

VSG combines computer vision and NLP to create coherent narratives from sequences of images or videos. This task extends beyond simple image captioning by requiring models to understand temporal relationships, infer causal connections between visual scenes, and generate engaging stories that capture the underlying narrative flow. Unlike video captioning, which focuses on objective description of visual content, VSG can also speculate on the motivations and emotions of characters since such interpretations do not contradict the visual content.

The field has evolved from early encoder–decoder architectures using CNN and RNN to Transformer-based models that leverage large-scale pretraining. These approaches must address challenges including visual understanding, temporal coherence, and narrative creativity while maintaining grounding in the provided visual content. The following subsections present the key methodologies and their contributions to advancing VSG capabilities.

5.1. Visual Storytelling with Convolutional and Recurrent Neural Networks

Huang et al. [19] proposed a dataset of sequential vision-to-language mappings, SIND v.1 (later renamed to VIST), with 81,743 images in 20,211 sequences, with both descriptive (captions) and story language, and aiming at facilitating research, from understanding individual images to comprehending image sequences that narrate events over time.

This work also proposed models for generating stories based on image sequences, using a sequence-to-sequence RNN approach, which extended the image captioning technique. It extracts features from each image using a pretrain CNN called AlexNET [67]. It then encodes image sequences by using a GRU over the features of each individual image. The story decoder model consists of another GRU and generates narratives word by word. Four different decoding methods, shown in Table 3, were considered: beam search, greedy search, greedy search with removal of duplicates, and greedy search with removal of duplicates and incorporation of grounded words.

Table 3. Results of each decoding approach using evaluated using METEOR scores on the VIST [19] dataset. Red indicates the lowest score, green the highest.

Method	METEOR Score
Beam = 10	23.13
Greedy	27.76
Greedy − Duplicates	30.11
Greedy − Duplicates + Grounded	31.42

Results indicate that beam search alone does not perform well in generating high-quality stories. This contrasts with the results for image captioning, where beam search is more effective than greedy search. The reason is that generating stories with beam search results in generic high-level descriptions such as “This is a picture of a dog” originating from the label bias problem. On the other hand, by using a beam size of 1 (greedy search) there is an improvement in the quality of the generated stories.

Since stories generated using a greedy search may sometimes contain repeated words or phrases that can adversely affect textual quality, a duplicate-removal heuristic was

added to ensure that content words cannot be produced more than once on a given story. Finally, to further improve quality, the previous approach was extended to include "visually grounded" words. These are words that are authorized for use by the story model only if they are licensed by the caption model: for a word to be included in the story, it must have appeared with some frequency in the captions generated by the caption model for the sequence of images. Figure 1 shows an example of the generated stories for an image sequence.

Figure 1. Sequence of images used to generate stories using different decoding methods. Beam search: This is a picture of a family. This is a picture of a cake. This is a picture of a dog. This is a picture of a beach. This is a picture of a beach. Greedy: The family gathered together for a meal. The food was delicious. The dog was excited to be there. The dog was enjoying the water. The dog was happy to be in the water. Greedy − duplicates: The family gathered together for a meal. The food was delicious. The dog was excited to be there. The kids were playing in the water. The boat was a little too much to drink. Greedy − duplicates + grounded words: The family got together for a cookout. They had a lot of delicious food. The dog was happy to be there. They had a great time on the beach. They even had a swim in the water. Adapted from [19].

5.2. Visual Storytelling with Convolutional and Bidirectional Recurrent Networks Networks

Park et al. [22] proposed Coherent Recurrent Convolutional Network (CRCN), integrating various neural network components to bridge the gap between visual content and textual narrative: VGGNet [68] as image encoder and Bidirectional Recurrent Networks (BRNN) to capture contextual information from both the forward and backward directions within a given text sequence. CRCN also introduces a local coherence model that focuses on maintaining coherence and context within the generated sentence sequences. It also utilizes parse trees to detect grammatical roles and structural attributes, ensuring that the generated sentences integrate with the broader narrative context. During training, correctly aligned image–sentence pairs are given higher scores than misaligned pairs.

The evaluation of CRCN considered several automatic metrics and several baseline works, including one that excludes the entity coherence model. Table 4 shows the results for the NY City [22] and Disneyland [22] datasets. Both datasets were introduced by Park et al. [22] and consist of image sequences with the corresponding blog posts. In these datasets, results were superior to all baselines. The exception, where it was a close second, was with the baseline without entity coherence and in only certain metrics. The work was further validated by user studies conducted via Amazon Mechanical Turk (AMT), showing that human evaluators prefer CRCN-generated sequences over several baselines.

Table 4. Performance metrics of CRCN on New York City and Disneyland Datasets. Green values represent state-of-the-art results at the time. B stands for BLEU, C for CIDEr, M for METEOR, and R for Recall [22].

Metric	B-1	B-2	B-3	B-4	C	M	R@1	R@5	R@10
NY City	26.83	5.37	2.57	2.08	30.9	7.69	11.67	31.19	43.57
Disneyland	28.40	6.88	4.11	3.49	52.7	8.78	14.29	31.29	43.20

Figure 2 shows an example of a sequence of images and the corresponding generated sentences.

Figure 2. Sequence of images used by CRCN in a story. (**1**) One of the hallway arches inside of the library. (**2**) As we walked through the library I noticed an exhibit called lunch hour nyc it captured my attention as I had also taken a tour of nyc food carts during my trip. (**3**) Here is the top of the Chrysler building everyone's favorite skyscraper in new york. (**4**) After leaving the nypl we walked along 42nd st. (**5**) We walked down fifth avenue from rockefeller centre checking out the windows in saks the designer stores and eventually making our way to the impressive new york public library. Adapted from [22].

In addition to the work by Park et al. [22], which was the first to use a bidirectional RNN for visual storytelling, other works have also employed bidirectional RNNs architectures. XE [25] and AREL [25] have proposed models for visual storytelling that use bidirectional RNNs on the VIST dataset.

5.3. Transitional Adaptation of Pretrained Models for Visual Storytelling

Yu et al. [24] proposed Transitional Adaptation of Pretrained Model (TAPM), aimed at refining the generation of textual descriptions for visual content, particularly in the context of VSG tasks. It aims at bridging the gap between pretrained language models and visual encoders. In contrast with previous models for vision-to-language generation tasks, which typically pretrain a visual encoder and a language generator separately and then jointly fine-tune them for the target task, TAPM proposes a transitional adaptation task to harmonize the visual encoder and language model for downstream tasks like VSG, mitigating discord between visual specificity and language fluency arising from separate training on large corpora of visual and text data. TAPM introduces an approach that adapts multimodal modules through a simpler alignment task focusing solely on visual inputs, eliminating the need for text labels.

TAPM components aim at enhancing the quality of textual descriptions for visual content in storytelling tasks. They are a visual encoder, a language generator, adaptation training, sequential coherence loss, training with adaptation loss, and a fine-tuning and inference process. The visual encoder, a pretrained model, extracts features from images or videos. In TAPM, it becomes integral during the adaptation phase, where it integrates with the language generator to fuse visual and textual information. The pretrained language generator model is responsible for converting visual information into textual descriptions. During the adaptation phase, it generates video and text embeddings, aligning textual representations with the corresponding visual features, based on a sequential coherence loss function.

The loss function divides sequential coherence into three components: past, current, and future matching losses. The past matching loss uses a Fully Connected (FC) layer f_p to project the text representation b_{s_i} of video i, drawing it nearer to the visual representation $v_{b_{i-1}}$ of the preceding video $i-1$ and distancing it from those of non-sequential videos. The future matching loss projects b_{s_i} via a distinct FC layer f_f, aligns with the subsequent visual representation $v_{b_{i+1}}$. The current matching loss then aligns the current visual representation v_{b_i} with b_{s_i} through another FC layer f_c. These components are unified by the FC layer projections in their respective visual spaces, pulling the embeddings of correct matches closer and pushing incorrect matches further apart. Margin ranking losses are utilized to implement this concept, contrasting correct matches with incorrect ones. The final

sequential coherence loss for a given video i is formulated as shown in Equation (1), where cos is the cosine similarity, and j represents indices of incorrect matches.

$$\begin{aligned} L_i = & \sum_{j \neq i-1} \max(0, 1 + \cos(v_{b_j}, f_p(b_{s_i})) - \cos(v_{b_{i-1}}, f_p(b_{s_i}))) \\ & + \sum_{j \neq i} \max(0, 1 + \cos(v_{b_j}, f_c(b_{s_i})) - \cos(v_{b_i}, f_c(b_{s_i}))) \\ & + \sum_{j \neq i+1} \max(0, 1 + \cos(v_{b_j}, f_f(b_{s_i})) - \cos(v_{b_{i+1}}, f_f(b_{s_i}))), \end{aligned} \tag{1}$$

TAPM uses a split-training strategy to optimize model performance. Initially, the visual encoder undergoes adaptation training with the adaptation loss, while the text encoder and language generator remain fixed. Subsequently, all components are jointly updated with the generation loss, allowing the model to optimize the adaptation task before addressing the more challenging generation objective. After the adaptation and split-training phases, TAPM undergoes fine-tuning. The model is then ready for the inference phase, generating captions for unseen visual inputs.

Tables 5 and 6 show results for TAPM against selected baselines on the LSMDC 2019 and VIST datasets. Table 7 shows human evaluation results in which TAPM surpasses adversarial baselines on LSMDC 2019. On VIST, it surpasses the XE [25] and AREL [25] baselines in relevance, expressiveness, and concreteness, as shown in Table 8. These results highlight the strengths in word choice and contextual accuracy, showcasing its ability to capture causal relationships between images. However, the score is still far from human performance, indicating that there is still room for improvement.

Table 5. Quantitative results on the VIST test set. C stands for CIDEr, M for METEOR, and R for ROUGE-L. Green indicates the highest score, red the lowest.

Models	C	M	R
Huang et al. [19]	-	31.4	-
AREL [25]	9.4	35.0	29.5
StoryAnchor [69]	9.9	35.5	30.0
HSRL [70]	10.7	35.2	30.8
INet [71]	10.0	35.6	29.7
TAPM	13.8	37.2	33.1

Table 6. Quantitative results on LSMDC 2019 public and blind test set. C stands for CIDEr and R for ROUGE. Green indicates the highest score, red the lowest. Adapted from [24].

Models	Public Test		Blind Test	
	C	R	C	R
Baseline [72]	7.0	12.0	6.9	11.9
XE [25]	7.2	11.5	-	-
AREL [25]	7.3	11.4	-	-
TAPM	10.0	12.3	8.8	12.4

Table 7. Human evaluation results on the LSMDC 2019 blind test set according to a Likert scale from 5 (worst) to 1 (best), where lower is better. Green indicates the highest score, red the lowest. Adapted from [24].

Models	Scores
Human	1.085
Official Baseline [72]	4.015
TAPM	3.670

Table 8. Human evaluation results on VIST. Higher is better. Green indicates the highest score, red the lowest. Adapted from [24]

	TAPM vs. XE			TAPM vs. AREL		
Choice (%)	**TAPM**	**XE**	**Tie**	**TAPM**	**AREL**	**Tie**
Relevance	59.9	34.1	6.0	61.3	32.8	5.9
Expressiveness	57.3	32.3	10.4	57.3	34.0	8.7
Concreteness	59.1	30.3	10.7	59.6	30.4	10.0

5.4. Interactive and Creative Visual Storytelling

Building upon the foundational CNN and RNN approaches, interactive and creative visual storytelling represents an important evolution in the field, introducing intermediate approaches that bridge visual comprehension with narrative creativity. These methods combine elements of image captioning and VQA while introducing narrative structure and creative interpretation, addressing limitations of purely technical approaches.

Lukin et al. [73] introduced a three-module pipeline for creative visual storytelling that systematically approaches the challenge of generating narratives from visual content. Their pipeline consists of object identification, single-image inferencing, and multi-image narration, that serve as a preliminary design for building a creative visual storyteller. The approach defines computational creative visual storytelling as one with the ability to alter the telling of a story along three aspects: to speak about different environments, to produce variations based on narrative goals, and to adapt the narrative to the audience. This modular approach demonstrates how the foundational capabilities of image captioning and VQA can be systematically extended and combined to support more complex narrative generation tasks. The pipeline explicitly separates visual understanding from narrative construction, allowing for more systematic development and evaluation of each component while maintaining the overall coherence required for storytelling.

Interactive visual storytelling has also explored user engagement and personalization in narrative generation. Halperin and Lukin [74] examined how creative visual storytelling can serve as an anthology for narrative intelligence, investigating the intersection of human creativity and automated story generation through an analysis of 100 visual stories from authors who participated in a systematic creative process of improvised story building based on image sequences. Their work on surreal visual storytelling [75] further explores how visual narrative systems can handle ambiguous or dreamlike imagery, investigating AI "hallucination" by stress-testing a visual storytelling algorithm with different visual and textual inputs designed to probe dream logic inspired by cinematic surrealism. These approaches highlight the importance of modularity in visual storytelling systems, where different components can be optimized independently while contributing to the overall narrative coherence. The integration of creative elements with systematic visual analysis provides a foundation for more sophisticated VSG systems that can balance factual visual description with imaginative narrative construction.

5.5. Knowledge-Enhanced Visual Storytelling

A challenge in VSG is the tendency of end-to-end approaches to produce monotonous stories with repetitive text and limited lexical diversity. This limitation arises because existing approaches are constrained by the vocabulary and knowledge available in single training datasets. To address this challenge, researchers have explored knowledge enhanced approaches that leverage external resources to enrich the story generation process.

Hsu et al. [76] introduced KG-Story, a three-stage framework that allows VSG systems to take advantage of external knowledge graphs to produce more diverse stories. The framework implements a distill–enrich–generate approach: first distilling a set of representative words from input prompts, then enriching the word set using external knowledge graphs, and finally generating stories based on the enriched word set. This framework allows the use of external resources not only for the enrichment phase, but also for the distillation and generation phases. The KG-Story framework operates through three distinct stages. In stage 1, an image-to-term model distills representative terms from each input image, creating conceptual representations that capture the essential elements of the visual content. In stage 2, external knowledge graphs are utilized to identify possible connections between the extracted term sets from different images, generating enriched term paths that capture relationships and associations not explicitly visible in the images. In stage 3, a Transformer architecture transforms these term paths into coherent stories, incorporating techniques such as length difference positional encoding and repetition penalties to improve narrative quality. Evaluation results demonstrate that stories generated by KG-Story are on average ranked better than previous state-of-the-art systems in human ranking evaluations. The approach successfully addresses the vocabulary limitations of traditional end-to-end methods while maintaining narrative coherence.

Interactive approaches to knowledge-enhanced storytelling have also been explored. Hsu et al. [77] introduced Dixit, an interactive visual storytelling system that allows users to iteratively compose stories through term manipulation. The system extracts text terms describing objects and actions from photos, then allows users to add new terms or remove existing ones before generating stories based on these modified term sets. Behind the scenes, Dixit uses an LSTM model trained on image caption data to distill terms from images and utilizes a Transformer decoder to compose context coherent stories. This approach opens up possibilities for interpretable and controllable visual storytelling, allowing users to understand the story formation rationale and to intervene in the generation process.

These knowledge-enhanced approaches represent an important advancement in VSG, demonstrating how external knowledge resources can be systematically integrated to overcome the limitations of purely data-driven methods. By explicitly modeling the relationship between visual content and broader conceptual knowledge, these systems can generate more diverse, engaging, and contextually rich narratives while maintaining controllability and interpretability.

5.6. Iterative and Planning-Based Approaches

Traditional VSG models typically generate stories in a single forward pass. However, creative writers use their knowledge and worldview to put disjointed elements together to form a coherent storyline, and work and rework iteratively toward perfection. VSG models, however, make poor use of external knowledge and iterative generation when attempting to create stories. This observation has motivated the development of iterative and planning-based approaches that more closely mirror human creative writing processes. Hsu et al. [78] introduced PR-VIST (Plot and Rework Visual Storytelling), a framework that represents the input image sequence as a story graph in which it finds the best path to form a storyline. PR-VIST then takes this path and learns to generate the final story

via a re-evaluating training process. The framework draws inspiration from the human creative writing process, which involves plotting (planning the overall narrative structure) and reworking (iteratively refining the story content).

The PR-VIST framework operates through two main stages that correspond to the plotting and reworking phases of creative writing. In stage 1 (story plotting), the system constructs a story graph from the input image sequence using external knowledge graphs, including VIST and Visual Genome knowledge graphs. A storyline predictor model identifies the best path through this graph to form a coherent storyline, creating a structured narrative plan that connects the images through meaningful conceptual relationships. This plotting stage essentially creates a roadmap for the story that will be generated. In stage 2 (story reworking), the framework generates the actual story text based on the predicted storyline path. The approach uses a length-controlled Transformer that is first pretrained on the ROC Story dataset and then fine-tuned on the VIST dataset with a discriminator component to encourage higher quality story generation. The re-evaluating training process allows the model to iteratively improve the story generation by learning from feedback about narrative quality and coherence.

Evaluation results demonstrate that this framework produces stories that are superior in terms of diversity, coherence, and humanness, per both automatic and human evaluations. An ablation study shows that both plotting and reworking contribute to the model's superiority. The explicit separation of planning and generation phases allows for better control over narrative structure while maintaining the flexibility needed for creative story generation. This planning-based approach represents an important advancement toward more human-like story generation, demonstrating how computational systems can benefit from explicitly modeling the structured creative processes that human writers naturally employ.

5.7. Combining LLMs (GPT-2) and Character Features

Hong et al. [21] proposed CharGrid (Character-Grid Transformer), a Transformer model that integrates diverse input features, from global image attributes to character-specific details. It takes a sequence of input tokens, including global image features obtained using Swin Transformer's [57], character features extracted by cropping character instances, a character grid for coherence assessment, and text features introduced incrementally during processing. The trainable image and character encoders process the image feature inputs. The character grid is flattened and fed to a FC layer. The Transformer module then processes these inputs, generating an output as a probability distribution over potential tokens using a pretrained Generative Pretrained Transformer (GPT)-2 tokenizer. The model architecture is shown in Figure 3.

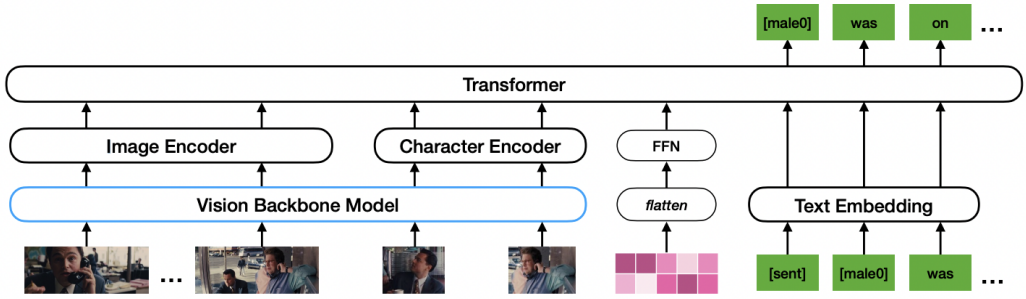

Figure 3. CharGrid architecture: the blue boxes are pretrained components where the parameters are fixed. Adapted from [21].

During training, model parameters, excluding the vision backbone, are randomly initialized. The training objective involves maximizing the likelihood of image sequence–

story pairs through backpropagation. Nucleus sampling [79] is employed for validation, and the METEOR score is used for evaluation.

CharGrid has two additional variants: ObjGrid (Object-Grid Transformer) and EntiGrid (Entity-Grid Transformer). ObjGrid replaces character features with object features. This model incorporates global features and object features, with an object grid representing coherence based on the similarity between global image features and object features. It is similar to CharGrid but includes both character and object features. EntiGrid incorporates global, character, and object features, with an entity grid representing coherence through the similarity between global image features and both character and object features. These variants explore the impact of different input feature combinations on model performance. CharGrid, with its emphasis on character coherence, serves as the primary model, while ObjGrid and EntiGrid address the contributions of object-related features.

The results in Table 9 show that CharGrid outperforms TAPM with character features and GPT-2 with character features on various metrics, emphasizing the effectiveness of character grid representations for coherence in VSG. A crowd-sourcing experiment was conducted with 28 workers to obtain binary judgments on grammaticality, coherence, diversity, and visual groundedness of the generated stories. The results in Table 10 show that TAPM with character features excels in visual groundedness over plain TAPM, while CharGrid surpasses TAPM with character features across all metrics. Two-sided binomial tests support that the character grid representation yields superior stories, affirming the reference-based metrics findings.

Table 9. Results using different input features on the test set of Visual Writing Prompts (VWP). Superscripts ^{1}, ^{2}, and ^{3} indicate 1/2/3 standard deviations away from the CharGrid mean. Global features are always included. "+ obj" and "+ char" represent that object and character features are also included, respectively. BLEU (B), METEOR (M), ROUGE-L (R-L), and CIDEr (C) values are averages of three runs with different random seeds. Green indicates the highest score, red the lowest. Adapted from [21].

Model	B-1	B-2	B-3	B-4	M	R-L	C
GPT-2	38.65 ^{3}	20.28 ^{3}	9.78 ^{3}	4.68 ^{2}	31.64 ^{3}	24.24 ^{1}	1.66 ^{3}
GPT-2 + obj	40.65 ^{3}	21.35 ^{3}	10.2 ^{3}	4.87 ^{2}	31.69 ^{3}	24.05 ^{1}	1.85 ^{3}
GPT-2 + char	39.95 ^{3}	21.04 ^{3}	10.11 ^{3}	4.92 ^{1}	31.85 ^{2}	24.19 ^{1}	1.57 ^{3}
GPT-2 + obj,char	40.41 ^{3}	21.44 ^{3}	10.56 ^{3}	5.06 ^{1}	32.03 ^{2}	24.38	1.87 ^{3}
TAPM	39.85 ^{3}	21.7 ^{3}	10.72 ^{3}	5.19	32.38 ^{1}	25.09	1.48 ^{3}
TAPM + obj	40.86 ^{3}	22.13 ^{3}	10.83 ^{3}	5.25	32.34 ^{1}	24.91	1.82 ^{3}
TAPM + char	40.03 ^{3}	21.68 ^{3}	10.66 ^{3}	5.18	32.42 ^{1}	24.88	1.4 ^{3}
TAPM + obj,char	40.87 ^{3}	21.99 ^{3}	10.72 ^{3}	5.06 ^{1}	32.48 ^{1}	24.87	1.59 ^{3}
ObjGrid + obj	47.66	25.26	11.95	5.42	32.83	24.42	4.68
EntityGrid + obj,char	45.83	24.85	12.11	5.7	32.68	24.89	3.53 ^{1}
CharGrid + char	47.71	25.33	11.95	5.42	33.03	25.01	4.83

Table 10. Human binary judgments in percentage of generated stories between TAPM and TAPM with character features (TAPM + char), TAPM + char, and CharGrid on the test set of VWP across four criteria: Grammaticality (G), Coherence (C), Visual Groundedness (VG), and Diversity (D). The numbers are percentages. * means p-value < 0.05. ** means p-value < 0.01. Adapted from [21].

Model	G	C	VG	D
TAPM + char vs. TAPM	+2.45	+1.99	+3.99 *	+1.69
CharGrid vs. TAPM + char	+6.49 **	+8.41 **	+6.25 *	+11.06 **

5.8. Visual Story Generation using Graphs for Event Ratiocination

Zheng et al. [26] introduced Hypergraph-Enhanced Graph Reasoning (HEGR) for "visual event ratiocination", a task that involves generating interpretations and narratives for events that precede, coincide with, or follow a visual scene, as well as discerning the intents of depicted characters. The work addresses several downstream tasks, including VSG. The work introduces a framework to improve the integration and interpretation of visual and textual data, using hypergraphs to address challenges in multimodal interactions and temporal dynamics. Figure 4 shows this architecture.

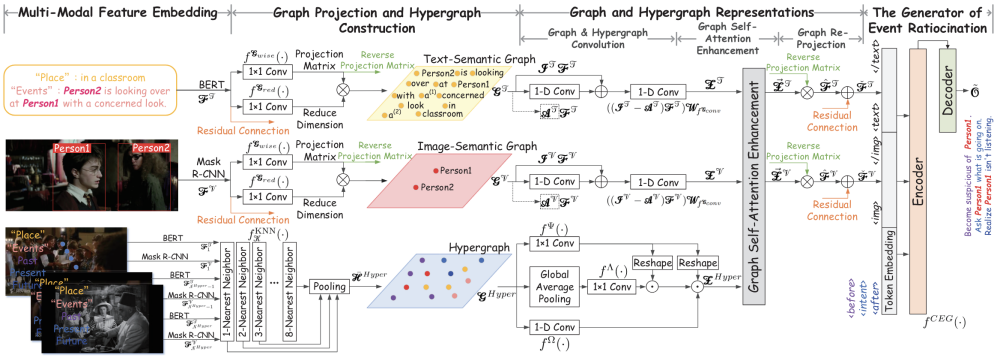

Figure 4. Architecture of HEGR by Zheng et al. Adapted from [26].

The method begins with feature extraction: visual features are extracted using Mask R-CNN [80] to detect objects and their spatial characteristics within images. Textual features are extracted from associated descriptions or captions using BERT [81], capturing semantic information. In the second phase, semantic graphs are constructed for each modality: nodes represent detected features or semantic tokens, while edges depict relationships within the same modality. To capture dependencies between modalities and across various temporal segments, hypergraphs are employed, linking semantic graphs based on contextual and temporal relevance, forming a high-dimensional interaction map. In the third phase, graph convolution is used to update node representations within each graph, allowing contextual information aggregation through learned adjacency matrices. Hypergraph convolution extends this to include higher-order relationships across multiple graphs and temporal instances. In the fourth phase, a self-attention mechanism refines graph representations by dynamically adjusting the significance of nodes within and across modalities. This process focuses on the most relevant features for narrative generation. Finally, the event ratiocination generator uses the enriched feature representations to generate narratives using BART [82].

The method was evaluated on the VIST dataset, achieving state-of-the-art performance in terms of BLEU-4, ROUGE-L, and METEOR scores achieving 16.4, 37.4, and 37.8, respectively.

5.9. Discussion

The works presented in this section have made substantial contributions to the field of VSG, establishing robust benchmarks for the assessment of emerging models and methodologies.

Specifically, the first three techniques, presented in Sections 5.1–5.3, employed a CNN encoder for visual feature extraction and a RNN decoder for language generation. The methods detailed in Sections 5.1 and 5.3 offered a comparative analyses on the VIST dataset, with METEOR scores of 31.4 and 37.2, respectively. Notably, the latter technique surpasses the former, which can be attributed to its transitional training approach.

The study outlined in Section 5.7, CharGrid, does not present findings on the VIST dataset. However, this method is benchmarked against the method detailed in Section 5.3, TAPM, on the VWP dataset, where it demonstrates superior performance across all eval-

uated metrics. This comparison is shown in Table 9, specifically within the rows labeled "TAPM + obj,char" and "CharGrid". This method differentiates itself from the previous by focusing on the extraction of object-level features from visual inputs rather than employing a CNN to encode the entire visual input. Furthermore, it leverages the capabilities of an LLM, using its extensive world knowledge to enhance language modeling, illustrating a novel integration of visual and linguistic elements in story generation.

Finally, the method discussed in Section 5.8, HEGR, introduces a different approach that uses hypergraphs to model the contextual and temporal relationships between visual and textual data. This approach is explicitly able to model some story elements, namely the characters and the setting.

Table 11 summarizes the performance of the methods described above, along with other methods from the literature that contributed to the field of VSG. The listed models range from earlier methods using CNNs and RNNs to more recent advances that integrate object detection models like Faster R-CNN [83] and Mask-RCNN [80], and LLMs like GPT-2 [84] and BART [82]. The first two methods by Park et al. (2015) [22] and Park et al. (2019) [72] are not comparable to the other methods as they do not use the VIST dataset. The remaining methods are evaluated on the VIST dataset, with the exception of CharGrid (2023) [21] that is evaluated on the VWP dataset. The authors of CharGrid trained and evaluated the performance of TAPM on the VWP [21] dataset, making the two methods directly comparable. It is, however, not possible to know if CharGrid would perform better than HEGR on the VIST dataset without further experiments.

Table 11. Analysis of Visual Story Generation works. The columns "B-4", "R-L", and "M" represent the BLEU-4, ROUGE-L, and METEOR scores, respectively. Green indicates the highest score and red indicates the lowest score for each metric within each dataset. All works, except Huang et al. (2015) [19] and CST (2018) [85], use attention mechanisms. All works use images as input, except ActivityNet [86] (it uses videos). The mark [1] indicates works that used extra training data (direct comparison may not be fair). The mark [2] indicates works that explicitly model some story elements, namely the characters by the detection of objects in the case of CharGrid (2023) [21] and HEGR (2021) [26] and the setting in the case of HEGR (2021) [26].

Model	Encoding	Decoding	Dataset	B-4	R-L	M
Park et al. (2015) [22]	CNN, Bi-RNN	RNN	Blog Posts	3.49	-	8.78
Park et al. (2019) [72]	Resnet-101/152, LSTM	LSTM	ActivityNet	9.91	-	16.48
Huang et al. (2015) [19]	CNN, GRU	GRU	VIST	-	-	31.4
Xu et al. (2015) [7]	VGGNet	LSTM	VIST	-	28.94	32.98
Yu et al. (2017) [87]	ResNet-101, Bi-GRU	GRU	VIST	-	29.53	34.12
CST (2018) [85]	CNN, LSTM	LSTM	VIST	12.7	29.2	34.4
AREL (2018) [25]	ResNet-152, Bi-GRU	GRU	VIST	14.1	29.5	35.0
HSRL (2019) [70]	ResNet-152	LSTM	VIST	12.32	30.84	35.23
GLACNet (2019) [88]	ResNet-152, Bi-LSTM	LSTM	VIST	-	-	30.63
StoryAnchor (2020) [69]	ResNet-152, Bi-GRU	GRU	VIST	14.0	30.0	35.5
INET (2020) [71]	CNN, GRU	GRU	VIST	14.7	29.7	35.5
AOG-LSTM (2023) [89]	RestNet-152	LSTM	VIST	12.9	30.1	36.0
CoVS (2023) [90]	RestNet-152, GRU	GRU	VIST	15.2	30.8	36.5
TAPM (2021) [24] [1]	Faster R-CNN, ResNet-152	GPT-2 s	VIST	-	33.1	37.2
HEGR (2021) [26] [1,2]	Mask-RCNN, BERT	BART	VIST	16.4	37.4	37.8
TAPM (2021) [21]	Faster R-CNN, ResNet-152	GPT-2	VWP	5.19	25.09	32.38
CharGrid (2023) [21] [2]	Mask-RCNN, SwinTransformer	GPT-2	VWP	5.42	25.01	33.03

6. Real-World Applications

As VSG technologies mature, they are increasingly being deployed in real-world scenarios. These applications demonstrate the practical potential of VSG systems and highlight new challenges that arise when deploying these technologies in dynamic, context-rich environments. This section explores emerging applications that integrate visual storytelling with contextual information, location awareness, and real-time data processing.

6.1. Location-Aware Visual Storytelling

Suwono et al. [91] introduced location-aware visual question generation with lightweight models, representing an important advancement in spatially grounded narrative applications. Their work addresses the challenge of generating contextually relevant questions and narratives that are informed by geographical location, opening new possibilities for travel documentation, educational applications, and location-based social media platforms. The location-aware approach integrates geographical metadata with visual content analysis to create narratives that are visually grounded and spatially contextualized. This integration enables the generation of stories that can reference specific landmarks, cultural contexts, and geographical features that would be meaningful to users familiar with particular locations. The use of lightweight models addresses practical deployment constraints, making the technology suitable for mobile applications and resource-constrained environments.

6.2. Automotive and Real-Time Context Integration

Belz et al. [92] explored story-driven approaches in automotive contexts, investigating how real-time contextual information can enhance automated storytelling systems. Their research represents a step toward practical deployment of VSG in everyday scenarios, particularly in automotive systems where visual storytelling could enhance the driving experience and provide meaningful documentation of journeys. The automotive application domain presents challenges and opportunities for visual storytelling. In-car camera systems continuously capture visual data, creating an opportunity for real-time narrative generation that could help drivers and passengers understand and remember their journeys. The integration of real-time context information—such as speed, location, weather conditions, and traffic patterns—enables the generation of richer, more informative narratives that extend beyond visual description. Their work demonstrates how dynamic contextual data can be incorporated into storytelling algorithms, showing measurable improvements in narrative relevance and user engagement when real-time information is provided. This approach could lead to applications in autonomous vehicles, where passengers might receive automatically generated travel narratives, or in fleet management systems where journey documentation becomes an automated process. Deployment challenges include processing continuous visual streams with limited computational resources, handling rapid scene changes, and maintaining narrative coherence across extended journeys.

7. Conclusions

This survey reviewed the field of VSG, explaining its development and current methods. It started by explaining basic story elements, datasets, and evaluation metrics used to judge the quality of the stories. The discussion pointed out the need for better metrics to assess complex tasks like VSG accurately. The document also covered related areas like image and video captioning and VQA, discussing how improvements in these fields have helped enhance VSG. The discussion then moved on to the evolution of VSG methods, starting with early approaches that used CNNs and RNNs to methods based on LLMs and hypergraphs. Additionally, we examined emerging real-world applications that demon-

strate the practical potential of these technologies in domains such as automotive systems and location-aware storytelling.

Despite the advances, challenges remain for VSG. In the following subsections, we discuss some of the key challenges and opportunities in the field that could guide future research directions. We start with the ones that are more concrete and could influence the field in the short term and end with the ones that are more abstract and could potentially influence the field in the long term.

7.1. Exploring Large Language Models

The field of VSG using LLMs is under-explored, with only a few works delving into this domain. To the best of our knowledge, the existing works predominantly rely on GPT-2, which is an outdated LLM when compared to models such as Qwen3-235B-A22B [93], which is the best performing open-source model on Chatbot Arena, or Gemini 2.5 Pro [94], which is the best performing proprietary model. The scarcity of the integration of LLMs with visual elements presents an opportunity for further research, exploring the potential of these models in generating more coherent and contextually relevant stories.

7.2. Automatic Evaluation Metrics

The need for robust and efficient automatic evaluation metrics is emphasized in our survey. While existing metrics like METEOR and CIDEr have been widely adopted, they do not fully capture the nuances of story quality. The metrics mostly focus on the similarity between the generated text and the reference text. A story can be retold in various ways, employing different perspectives while preserving the original message. Consequently, a generated text might deviate significantly from the reference text in terms of specific wording and still deliver the same story in a coherent way. Thus, while these metrics provide valuable insights into certain dimensions of text generation, they did not fully encompass the multifaceted nature of storytelling. As LLMs continue to improve, they show potential in replacing human annotators in many tasks [48] and so metrics to evaluate stories could be developed based on LLMs. We advance that these metrics could be based on the existing systematic scoring systems [47] that were developed to evaluate the performance of LLM-based chat bots.

7.3. Decomposition of Story Generation

Addressing the complexity of stories, we underscore the importance of decomposing the story generation process into modular components. Each module can focus on specific attributes such as author goals, character interactions, consistency, etc. This decomposition has already been explored with modeling characters in CharGrid [24] and the modeling of characters and setting in HEGR [26]. We believe that explicit modeling of other story elements and abstract concepts such as emotions and culture, possibly using hypergraphs or similar structures, could further enhance the quality of generated stories.

7.4. Hybrid Systems and Knowledge Integration

Building on the decomposition concept, hybrid systems, combining planning-based techniques and machine learning approaches, present an intriguing direction. Hypergraphs, as seen in HEGR [26], could possibly be a starting point as a deep learning approach to model the planning process replacing the traditional handcrafted planning systems. The integration of multiple knowledge sources can enrich the diversity and quality of generated stories, paving the way for more robust and context-aware narrative creation.

7.5. Real-World Deployment and Context Integration

The transition from academic research to practical applications represents a crucial next step for VSG technologies. Recent work in location-aware storytelling and automotive applications demonstrates the potential for deploying these systems in real-world scenarios. Future research should focus on developing robust systems that can integrate multiple contextual information sources including location, temporal data, and real-time sensor information while maintaining computational efficiency for deployment in resource-constrained environments such as mobile devices and embedded systems. The success of these real-world applications will ultimately determine the broader impact of VSG research on society.

7.6. Ethical Considerations and Societal Impact

The deployment of VSG systems in real-world applications such as journalism, autonomous vehicles, and personalized content generation raises important ethical considerations that must be addressed by the research community. Dataset biases present a concern, as many existing VSG datasets may lack demographic and cultural representativeness. For example, VIST images are predominantly North American/European family/leisure scenes, while LSMDC and PororoQA inherit scripted, gender-stereotyped roles. This limitation can lead to systems that perform poorly for underrepresented communities or perpetuate cultural stereotypes in generated narratives. The datasets listed in Table 1, while foundational to the field, primarily reflect Western cultural contexts and may not adequately represent global narrative traditions and perspectives. Such skews can surface in stories, for instance, defaulting to Western names. The risk of generating misleading or unverified content is relevant when VSG systems are applied in journalism or automated news generation. These systems may extrapolate beyond what is directly observable in visual content, potentially creating false narratives or misrepresenting events. Copyright issues surrounding the images used for training present legal and ethical challenges. Many datasets incorporate images from blogs, social media, and other sources without explicit permission for machine learning applications. Privacy concerns are inherent in datasets created from personal blogs and social media content. These sources may contain personally identifiable information (faces, locations, blog text) or private moments that were not intended for use in machine learning research, and may in some cases not comply with data protection regulations such as GDPR. The VIST dataset and similar collections derived from photo-sharing platforms raise questions about informed consent and data subject rights.

Author Contributions: Investigation, D.A.P.O.; writing—original draft preparation, D.A.P.O.; writing—review and editing, D.A.P.O., E.R. and D.M.d.M. supervision, E.R. and D.M.d.M.; funding acquisition, D.A.P.O. and D.M.d.M. All authors have read and agreed to the published version of the manuscript.

Funding: Daniel A. P. Oliveira is supported by a scholarship granted by Fundação para a Ciência e Tecnologia (FCT), with reference 2021.06750.BD. Additionally, this work was supported by Portuguese national funds through FCT, with reference UIDB/50021/2020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article as it is a survey paper reviewing existing literature.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Hühn, P.; Meister, J.; Pier, J.; Schmid, W. *Handbook of Narratology*; De Gruyter Handbook; De Gruyter: Berlin, Germany, 2014.
2. Abbott, H.P. *The Cambridge Introduction to Narrative*, 2nd ed.; Cambridge Introductions to Literature, Cambridge University Press: Cambridge, UK, 2008. [CrossRef]
3. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*; Pearson Prentice Hall: Upper Saddle River, NJ, USA, 2009.
4. Szeliski, R. *Computer Vision: Algorithms and Applications*; Texts in Computer Science; Springer: London, UK, 2010.
5. Gervas, P. Computational Approaches to Storytelling and Creativity. *AI Mag.* **2009**, *30*, 49. [CrossRef]
6. Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; Niebles, J.C. Dense-Captioning Events in Videos. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 706–715.
7. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.C.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015.
8. Fan, A.; Lewis, M.; Dauphin, Y. Hierarchical Neural Story Generation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018.
9. Peng, N.; Ghazvininejad, M.; May, J.; Knight, K. Towards Controllable Story Generation. In Proceedings of the First Workshop on Storytelling, New Orleans, LA, USA, 5 June 2018; pp. 43–49.
10. Simpson, J. *Oxford English Dictionary: Version 3.0: Upgrade Version*; Oxford University Press: Oxford, UK, 2002.
11. Lau, S.Y.; Chen, C.J. Designing a Virtual Reality (VR) Storytelling System for Educational Purposes. In *Technological Developments in Education and Automation*; Springer: Dordrecht, The Netherlands, 2008; pp. 135–138.
12. Mitchell, D. *Cloud Atlas: A Novel*; Random House Publishing Group: New York, NY, USA, 2008.
13. DiBattista, M. *Novel Characters: A Genealogy*; Wiley: Hoboken, NJ, USA, 2011; pp. 14–20.
14. Griffith, K. *Writing Essays About Literature*; Cengage Learning: Boston, MA, USA, 2010; p. 40.
15. Truby, J. *The Anatomy of Story: 22 Steps to Becoming a Master Storyteller*; Faber & Faber: London, UK, 2007; p. 145.
16. Rowling, J. *Harry Potter and the Sorcerer's Stone*; Harry Potter, Pottermore Publishing: London, UK, 2015.
17. Dibell, A. *Elements of Fiction Writing—Plot*; F+W Media: Cincinnati, OH, USA, 1999; pp. 5–6.
18. Pinault, D. *Story-Telling Techniques in the Arabian Nights*; Studies in Arabic Literature; Brill: Leiden, The Netherlands, 1992; Volume 15.
19. Huang, T.H.K.; Ferraro, F.; Mostafazadeh, N.; Misra, I.; Agrawal, A.; Devlin, J.; Girshick, R.; He, X.; Kohli, P.; Batra, D.; et al. Visual Storytelling. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1233–1239.
20. Hsu, T.Y.; Huang, C.Y.; Hsu, Y.C.; Huang, T.H. Visual Story Post-Editing. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Korhonen, A., Traum, D., Màrquez, L., Eds.; Association for Computational Linguistics: Florence, Italy, 2019; pp. 6581–6586.
21. Hong, X.; Sayeed, A.; Mehra, K.; Demberg, V.; Schiele, B. Visual Writing Prompts: Character-Grounded Story Generation with Curated Image Sequences. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 565–581. [CrossRef]
22. Park, C.C.; Kim, G. Expressing an Image Stream with a Sequence of Natural Sentences. In *Proceedings of the Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
23. Rohrbach, A.; Torabi, A.; Rohrbach, M.; Tandon, N.; Pal, C.J.; Larochelle, H.; Courville, A.C.; Schiele, B. Movie Description. *Int. J. Comput. Vis.* **2016**, *123*, 94–120. [CrossRef]
24. Yu, Y.; Chung, J.; Yun, H.; Kim, J.; Kim, G. Transitional Adaptation of Pretrained Models for Visual Storytelling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 15–19 June 2021; pp. 12658–12668.
25. Wang, X.; Chen, W.; Wang, Y.F.; Wang, W.Y. No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; Gurevych, I., Miyao, Y., Eds.; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 899–909.
26. Zheng, W.; Yan, L.; Gou, C.; Wang, F.Y. Two Heads are Better Than One: Hypergraph-Enhanced Graph Reasoning for Visual Event Ratiocination. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; Meila, M., Zhang, T., Eds.; Machine Learning Research; 2021; Volume 139, pp. 12747–12760.
27. Kim, K.M.; Heo, M.O.; Choi, S.H.; Zhang, B.T. DeepStory: Video story QA by deep embedded memory networks. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia 19–25 August 2017; pp. 2016–2022.
28. Das, P.; Xu, C.; Doell, R.; Corso, J. A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2634–2641.

29. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; Isabelle, P., Charniak, E., Lin, D., Eds., Association for Computational Linguistics: Philadelphia, PA, USA, 2002; pp. 311–318.
30. Zhu, Y.; Lu, S.; Zheng, L.; Guo, J.; Zhang, W.; Wang, J.; Yu, Y. Texygen: A Benchmarking Platform for Text Generation Models. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018.
31. Banerjee, S.; Lavie, A. 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*; Goldstein, J., Lavie, A., Lin, C.Y.,Voss, C., Eds.; Association for Computational Linguistics: Ann Arbor, MI, USA , 2005; pp. 65–72.
32. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2014; pp. 4566–4575.
33. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
34. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. *arXiv* **2019**, arXiv:1904.09675.
35. OpenAI. Hello GPT-4o. 2024. Available online: https://openai.com/index/hello-gpt-4o (accessed on 9 February 2025).
36. OpenAI. New Models and Developer Products Announced at DevDay. 2024. Available online: https://openai.com/blog/new-models-and-developer-products-announced-at-devday (accessed on 3 January 2024).
37. OpenAI. GPT-4 and GPT-4 Turbo: Models Documentation. 2024. Available online: https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo (accessed on 4 March 2024).
38. OpenAI. GPT-3.5: Models Documentation. 2024. Available online: https://platform.openai.com/docs/models/gpt-3-5 (accessed on 3 January 2024).
39. Touvron, H.; Martin, L.; Stone, K.R.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288. [CrossRef]
40. Meta. Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date. 2024. Available online: https://ai.meta.com/blog/meta-llama-3/ (accessed on 28 May 2024).
41. team, M.A. Mixtral of Experts. Mistral AI Continues Its Mission to Deliver Open Models to the Developer Community, Introducing Mixtral 8x7B, a High-Quality Sparse Mixture of Experts Model. 2023. Available online: https://mistral.ai/news/mixtral-of-experts/ (accessed on 10 September 2025).
42. Mistral AI team. Mistral NeMo: Our New Best Small Model. 2024. Available online: https://mistral.ai/news/mistral-nemo, (accessed on 27 July 2025).
43. Anthropic. Introducing Claude 4. 2025. Available online: https://www.anthropic.com/news/claude-4 (accessed on 6 August 2025).
44. Anthropic. Claude-2.1: Overview and Specifications. 2024. Available online: https://www.anthropic.com/index/claude-2-1 (accessed on 3 January 2024).
45. Anthropic. Claude-2: Overview and Specifications. 2024. Available online: https://www.anthropic.com/index/claude-2 (accessed on 3 January 2024).
46. Anthropic. Introducing Claude. 2024. Available online: https://www.anthropic.com/index/introducing-claude (accessed on 3 January 2024).
47. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv* **2023**, arXiv:2306.05685.
48. Gilardi, F.; Alizadeh, M.; Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2305016120. [CrossRef] [PubMed]
49. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2014; pp. 3156–3164.
50. Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; Courville, A. Describing Videos by Exploiting Temporal Structure. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Los Alamitos, CA, USA, 7–13 December 2015; pp. 4507–4515.
51. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2017; pp. 6077–6086.
52. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
53. Chung, J.; Gülçehre, Ç.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555. [CrossRef]
54. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.

55. Olimov, F.; Dubey, S.; Shrestha, L.; Tin, T.T.; Jeon, M. Image Captioning using Multiple Transformers for Self-Attention Mechanism. *arXiv* **2021**, arXiv:2103.05103. [CrossRef]
56. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021.
57. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
58. Yu, J.; Li, J.; Yu, Z.; Huang, Q. Multimodal Transformer with Multi-View Visual Representation for Image Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 4467–4480. [CrossRef]
59. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video Swin Transformer. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2021; pp. 3192–3201.
60. Lin, K.; Li, L.; Lin, C.C.; Ahmed, F.; Gan, Z.; Liu, Z.; Lu, Y.; Wang, L. SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2021; pp. 17928–17937.
61. Kazemi, V.; Elqursh, A. Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering. *arXiv* **2017**, arXiv:1704.03162. [CrossRef]
62. Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O.K.; Aggarwal, K.; Som, S.; Piao, S.; Wei, F. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. In *Proceedings of the Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 32897–32912.
63. Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O.K.; Singhal, S.; Som, S.; et al. Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 19175–19186.
64. Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; Yang, H. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022.
65. Wang, P.; Wang, S.; Lin, J.; Bai, S.; Zhou, X.; Zhou, J.; Wang, X.; Zhou, C. ONE-PEACE: Exploring One General Representation Model Toward Unlimited Modalities. *arXiv* **2023**, arXiv:2305.11172. [CrossRef]
66. Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.J.; Padlewski, P.; Salz, D.M.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. PaLI: A Jointly-Scaled Multilingual Language-Image Model. *arXiv* **2022**, arXiv:2209.06794.
67. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Pereira, F., Burges, C., Bottou, L., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2012; Volume 25.
68. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; Conference Track Proceedings; Bengio, Y., LeCun, Y., Eds.; ICLR: San Diego, CA, USA, 2015.
69. Modi, Y.; Parde, N. The Steep Road to Happily Ever after: An Analysis of Current Visual Storytelling Models. In Proceedings of the Second Workshop on Shortcomings in Vision and Language, Minneapolis, MN, USA, 6 June 2019; Bernardi, R., Fernandez, R., Gella, S., Kafle, K., Kanan, C., Lee, S., Nabi, M., Eds.; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 47–57.
70. Huang, Q.; Gan, Z.; Celikyilmaz, A.; Wu, D.O.; Wang, J.; He, X. Hierarchically Structured Reinforcement Learning for Topically Coherent Visual Story Generation. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
71. Jung, Y.; Kim, D.; Woo, S.; Kim, K.; Kim, S.; Kweon, I.S. Hide-and-Tell: Learning to Bridge Photo Streams for Visual Storytelling. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12February 2020.
72. Park, J.S.; Rohrbach, M.; Darrell, T.; Rohrbach, A. Adversarial Inference for Multi-Sentence Video Description. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2018; pp. 6591–6601.
73. Lukin, S.; Hobbs, R.; Voss, C. A Pipeline for Creative Visual Storytelling. In Proceedings of the First Workshop on Storytelling, New Orleans, LA, USA, 5 June 2018; Mitchell, M., Huang, T.H.K., Ferraro, F., Misra, I., Eds.; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 20–32.
74. Halperin, B.A.; Lukin, S.M. Envisioning Narrative Intelligence: A Creative Visual Storytelling Anthology. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 23–28 April 2023; CHI '23.

75. Halperin, B.A.; Lukin, S.M. Artificial Dreams: Surreal Visual Storytelling as Inquiry Into AI 'Hallucination'. In Proceedings of the 2024 ACM Designing Interactive Systems Conference, New York, NY, USA, 1–5 July 2024; pp. 619–637.
76. Hsu, C.C.; Chen, Z.Y.; Hsu, C.Y.; Li, C.C.; Lin, T.Y.; Huang, T.H.; Ku, L.W. Knowledge-Enriched Visual Storytelling. *AAAI Conf. Artif. Intell.* **2020**, *34*, 7952–7960. [CrossRef]
77. Hsu, C.C.; Chen, Y.H.; Chen, Z.Y.; Lin, H.Y.; Huang, T.H.K.; Ku, L.W. Dixit: Interactive Visual Storytelling via Term Manipulation. In Proceedings of the World Wide Web Conference, New York, NY, USA, 13–17 May 2019; pp. 3531–3535.
78. Hsu, C.y.; Chu, Y.W.; Huang, T.H.; Ku, L.W. Plot and Rework: Modeling Storylines for Visual Storytelling. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 4443–4453.
79. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The Curious Case of Neural Text Degeneration. *arXiv* **2019**, arXiv:1904.09751.
80. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. *arXiv* **2017**, arXiv:1703.06870.
81. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019.
82. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019.
83. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [CrossRef]
84. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. OpenAI Technical Report. 2019. Available online: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed on 15 September 2025).
85. Gonzalez-Rico, D.; Pineda, G.F. Contextualize, Show and Tell: A Neural Visual Storyteller. *arXiv* **2018**, arXiv:1806.00738. [CrossRef]
86. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Niebles, J.C. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970.
87. Yu, L.; Bansal, M.; Berg, T.L. Hierarchically-Attentive RNN for Album Summarization and Storytelling. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; Palmer, M., Hwa, R., Riedel, S., Eds.; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 966–971.
88. Kim, T.; Heo, M.O.; Son, S.; Park, K.W.; Zhang, B.T. GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation. *arXiv* **2018**, arXiv:1805.10973.
89. Liu, H.; Yang, J.; Chang, C.H.; Wang, W.; Zheng, H.T.; Jiang, Y.; Wang, H.; Xie, R.; Wu, W. AOG-LSTM: An adaptive attention neural network for visual storytelling. *Neurocomputing* **2023**, *552*, 126486. [CrossRef]
90. Gu, J.; Wang, H.; Fan, R. Coherent Visual Storytelling via Parallel Top-Down Visual and Topic Attention. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 257–268. [CrossRef]
91. Suwono, N.; Chen, J.; Hung, T.; Huang, T.H.; Liao, I.B.; Li, Y.H.; Ku, L.W.; Sun, S.H. Location-Aware Visual Question Generation with Lightweight Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Singapore, 2023; pp. 1415–1432.
92. Belz, J.H.; Weilke, L.M.; Winter, A.; Hallgarten, P.; Rukzio, E.; Grosse-Puppendahl, T. Story-Driven: Exploring the Impact of Providing Real-time Context Information on Automated Storytelling. In Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, New York, NY, USA, 13–16 October 2024; UIST '24.
93. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 Technical Report. *arXiv* **2025**, arXiv:2505.09388. [CrossRef]
94. Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv* **2025**, arXiv:2507.06261. [CrossRef]