IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 May 2023, accepted 30 June 2023, date of publication 10 July 2023, date of current version 18 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3293646

RESEARCH ARTICLE

Sequential Vision to Language as Story: A Storytelling Dataset and Benchmarking

ZAINY M. MALAKAN[1,2], SAEED ANWAR[3], GHULAM MUBASHAR HASSAN[1], (Senior Member, IEEE), AND AJMAL MIAN[1], (Senior Member, IEEE)

[1]Department of Computer Science and Software Engineering, The University of Western Australia, Crawley, WA 6009, Australia
[2]Department of Information Science, College of Computer and Information Systems, Umm Al-Qura University, Makkah 24382, Saudi Arabia
[3]Department of Information and Computer Science, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia

Corresponding author: Zainy M. Malakan (zmmalakan@uqu.edu.sa)

Professor Ajmal Mian is the recipient of an Australian Research Council Future Fellowship Award (project number FT210100268) funded by the Australian Government.

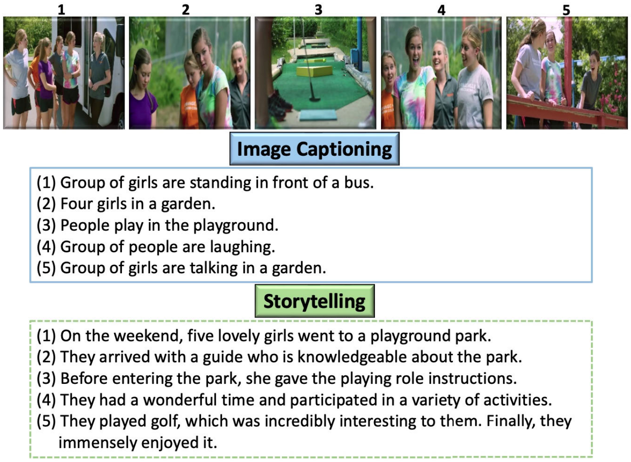
ABSTRACT Storytelling is a remarkable human skill that plays a significant role in learning and experiencing everyday life. Developing narratives is central to human mental health development, simultaneously encapsulating broad details such as psychology, morality and common sense. Contemporary deep-learning algorithms require similar skills to be able to tell a story from a visual perspective. However, most algorithms function at a superficial or factual level, aligning descriptive text with images in a one-to-one manner without considering the temporal relation. Stories are more expressive in style, language and content, involving imaginary concepts not explicit in the images. An ideal deep learning system should learn and develop cohesive, meaningful, and causal stories. Unfortunately, most existing storytelling methods are trained and evaluated on a single dataset, i.e., the VIsual STorytelling (VIST) dataset. Multiple datasets are essential to test the generalization ability of algorithms. We bridge the gap and present a new dataset for expressive and coherent story creation. We present the Sequential Storytelling Image Dataset (SSID, http://ieee-dataport.org/documents/sequential-storytelling-image-dataset-ssid) consisting of open-source video frames accompanied by story-like annotations. We provide four annotations (stories) for each set of five images. The image sets are collected manually from publicly available videos in three domains: documentaries, lifestyle, and movies, and then annotated manually using Amazon Mechanical Turk. We perform a detailed analysis and benchmarking of the current VIST dataset and our new SSID dataset and show that both datasets exhibit high variance within their multiple ground truth stories corresponding to the same image set. Moreover, our dataset achieves lower mean average scores across all metrics, meaning that the ground truth stories of our dataset are more diverse. Finally, we train and evaluate existing state-of-the-art rhetorical storytelling methods on both datasets and show that our dataset is more challenging and requires sophisticated techniques to accurately detect a significant variety of events.

INDEX TERMS Storytelling, visual understanding dataset, image and video captioning, computer vision, sequential storytelling image dataset (SSID).

I. INTRODUCTION

Visual storytelling refers to the manner of describing a set of images rather than a single image, also known as "multi-image captioning" [1], [2], [3]. Visual Storytelling Task (VST) takes a set of images as input and aims to generate a coherent story relevant to the input images. VST has a wide range of potential applications in our daily life, such as assisting visually disabled people in better comprehending everyday events and the contents of photos found on the Internet. It also exemplifies the sophisticated creativity that an artificial intelligence system can achieve.

The associate editor coordinating the review of this manuscript and approving it for publication was Antonio J. R. Neves.



(1) Group of girls are standing in front of a bus.
(2) Four girls in a garden.
(3) People play in the playground.
(4) Group of people are laughing.
(5) Group of girls are talking in a garden.

(1) On the weekend, five lovely girls went to a playground park.
(2) They arrived with a guide who is knowledgeable about the park.
(3) Before entering the park, she gave the playing role instructions.
(4) They had a wonderful time and participated in a variety of activities.
(5) They played golf, which was incredibly interesting to them. Finally, they immensely enjoyed it.

FIGURE 1. Illustrating the differences between the typical visual captioning/description and storytelling approaches. Each image is captioned with a single description in the first block. The second block represents the narrative sentences as a story for the same set of images.

The most comparable research problems to visual storytelling are image and video captioning problems. The availability of numerous public datasets for these two problems has contributed significantly to the rapid development of related methodologies over the past years. Image captioning datasets cannot be directly used for storytelling since, in these datasets, each image is described individually, and there is no logical connection between the sentences. Visual storytelling relies more on narrative structure than textual captioning from the linguistic perspective. In addition, maintaining coherency in the manner in which the story is described is challenging due to the fact that the images have significant perceptual variations when gathered collectively. Therefore, collecting and annotating datasets for visual storytelling methods accessible to the public will positively impact the development of solutions for this type of problem. Figure 1 illustrates the differences between describing a set of images in two ways. The traditional image captioning approach is shown in the blue box, indicating that the sentences are unrelated and not composed in a story-like fashion. On the other hand, the green-highlighted box demonstrates the storytelling method, which explains a set of images so that all of the sentences are written in a manner that is consistent, relevant and written in the style of a story.

The vast majority of visual storytelling models are trained and evaluated on the VIST dataset released by Microsoft Research [4]. Hence, there is a limitation of datasets that are accessible to the public for advancing research in this direction. It is not possible to test the generality of solutions when only one dataset is available for testing. In addition, the VIST dataset images were collected using the Flickr albums API and reorganized into a set of images of previously shot photos. Hence, the VIST dataset image sets inherently lack coherence.

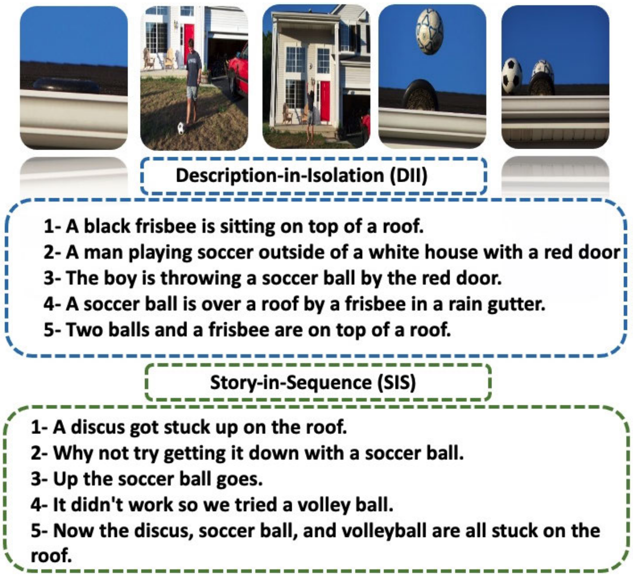
In this paper, we are motivated to develop a dataset for visual storytelling produced from images collected from open-source videos. Compared to VIST, our dataset images are collected as a set of five images from open-source videos where continuous scenes inherently have logical coherence. To construct our storytelling dataset, two fundamental processes need to be performed. The first process is to gather the images, and there should be five images of a sequence event in a set. In this dataset, we focused on three video categories: narrative movies, lifestyle documentaries, and media appearances from everyday life events. It is noteworthy to highlight that VIST dataset comprises two distinct categories of data, namely, descriptions in isolation and stories in a sequence. Indeed, the dataset employs identical images with varying annotation styles. In contrast, we used a different collection of images in SSID dataset that followed an identical annotation style, namely a story presented in a sequential format. In SSID dataset, we took five screenshots/frames and sorted them to form a story. The second process in creating the dataset is the annotation process. The annotations must be written using a story writing style. For this purpose, we used Amazon Mechanical Turk (AMT). We uploaded all the collected sets of images to the Amazon server so AMT workers could write a story for each captured set of images. Finally, we checked whether each written story was coherent and relevant to the images in the set. Incoherent or irrelevant stories were re-written or removed from the dataset.

The following is a summary of our contributions:

- We propose a novel Sequential Storytelling Images Dataset (SSID)[1] consisting of sets of images collected from open-source videos along with the corresponding descriptions in the form of a coherent story. Each set consists of five images accompanied by five connected sentences structured as a story.
- To demonstrate that the proposed dataset is beneficial in resolving issues related to visual storytelling, we benchmark various storytelling models recently published on our new dataset. We present the results of all automatic evaluation metrics, such as BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, and METEOR, for our SSID dataset and the existing VIST dataset.
- We comprehensively analyze the ground truth annotations accompanied by both storytelling datasets. For this, each annotated narrative is retrieved and contrasted with other associated narratives for the same set of images. This is possible because multiple annotators annotate each set of images with a story. We report the average of means and standard deviations to show the significant, problematic variations that exist in the ground truth annotations. Section VI-B includes additional details.
- We showcase the quality of the proposed dataset by incorporating human evaluation measures and comparing it to the existing VIST dataset, which is the only publicly available dataset.

The subsequent sections of this study are organized as follows. Section II provides detailed background on visual

[1] https://dx.doi.org/10.21227/dbr9-dq51



Description-in-Isolation (DII)

1- A black frisbee is sitting on top of a roof.
2- A man playing soccer outside of a white house with a red door
3- The boy is throwing a soccer ball by the red door.
4- A soccer ball is over a roof by a frisbee in a rain gutter.
5- Two balls and a frisbee are on top of a roof.

Story-in-Sequence (SIS)

1- A discus got stuck up on the roof.
2- Why not try getting it down with a soccer ball.
3- Up the soccer ball goes.
4- It didn't work so we tried a volley ball.
5- Now the discus, soccer ball, and volleyball are all stuck on the roof.

FIGURE 2. An example of the VIST dataset, which consists of two types of image sets: Description In Isolation (DII), in which each image is described in isolation, and Description In Sequence (SIS), in which each image set is described in a narrative fashion [4].

description techniques, including the limitations of image and video captioning datasets and methods in describing a sequence of images as a story. Section III presents the proposed SSID dataset construction, including image collection domain, methodology, and image annotation, which serves as a valuable resource for future researchers seeking to construct similar datasets. Section IV illustrates all existing state-of-the-art models and experimental settings used to evaluate our proposed dataset. Section V presents the comparison of results of state-of-the-art methods. Section VI includes detailed experimental discussions, including time complexity, ground truth variance, and qualitative analysis. Finally, Section VII explains the limitations of the study and our future works, and Section VIII concludes the paper.

II. VISUAL DESCRIPTION LITERATURE REVIEW

This section includes a literature review of recent trends in publicly accessible datasets for visual description techniques. The effectiveness of a visual description dataset may indeed contribute to the algorithm's overall effectiveness. In addition, we investigate the approaches most similar to visual storytelling and why they fail to explain a set of images, including image captioning and video captioning. Finally, the most recent approaches to visual storytelling are highlighted and discussed.

A. IMAGE CAPTIONING

1) DATASETS

Image captioning datasets cover multiple subjects such as daily scene images [5], [6], [7], human activities [8], [9], [10], etc. The most common dataset, i.e., the daily scene annotated images, is split into training, validation and testing. The MSCOCO [5], Multi30K-CLID [6], and AIC [7] datasets comprise a total of five human-written annotations for each image. Similarly, the human activities datasets include Flickr8k [8], Flickr30k [9] and PASCAL 1K [10]. All these datasets have a set of five annotations associated with each image. Moreover, the image captioning methods employ the datasets that have topics of news [11], still natural [12], blind view [13], novel objects [14] and fashion items [15]. Unfortunately, all available image captioning datasets are constructed as unconnected sentences, limiting their use in storytelling techniques.

2) TECHNIQUES

A single frame or an image paired with a single sentence is an example of image captioning. Image captioning algorithms can be classified further into rule-based methods [16] and deep learning-based methods [17], [18]. Rule-based methods are classical approaches that use template-based algorithms to recognize a predefined and limited collection of patterns, actions, and factors inside an image and describe them in natural language. Advanced techniques rely on deep learning and other advanced concepts, such as reinforcement learning [19], semantic attribute integration [20], attention [21], and subject and object modelling [22]. Due to advancements in deep learning and the availability of larger datasets [23], the aforementioned image captioning techniques show superior performance. However, when designing a story for a set of images, none of these techniques seems particularly effective.

B. VIDEO CAPTIONING

1) DATASETS

Standard video captioning datasets for video captioning algorithms are organized in a multi-frame manner, equivalent to datasets for visual storytelling covering a variety of topics, including cooking [24], [25], [26], movies [27], humans [28] and social media [29]. The accessibility of annotated datasets to be utilized in video captioning has been the primary impetus behind the rapid development of this field of study. Only a minority of the mentioned datasets contain various phrases or paragraphs for each video sample, while the majority only provide a single description of each video. Although video captioning datasets help train a model to describe multiple frames in a sequence, visual storytelling algorithms require a set of images with corresponding connected sentences structured to generate a description more in the form of a story than a general description. This is because visual storytelling algorithms are expected to generate a story-like description rather than a general one.

2) TECHNIQUES

Video captioning is an extension of the image captioning field that may explain consecutive keyframes (i.e., a video) in a single sentence. Standard video captioning approaches utilize encoder-decoder architecture, similar to visual storytelling techniques. In order to extract visual features from

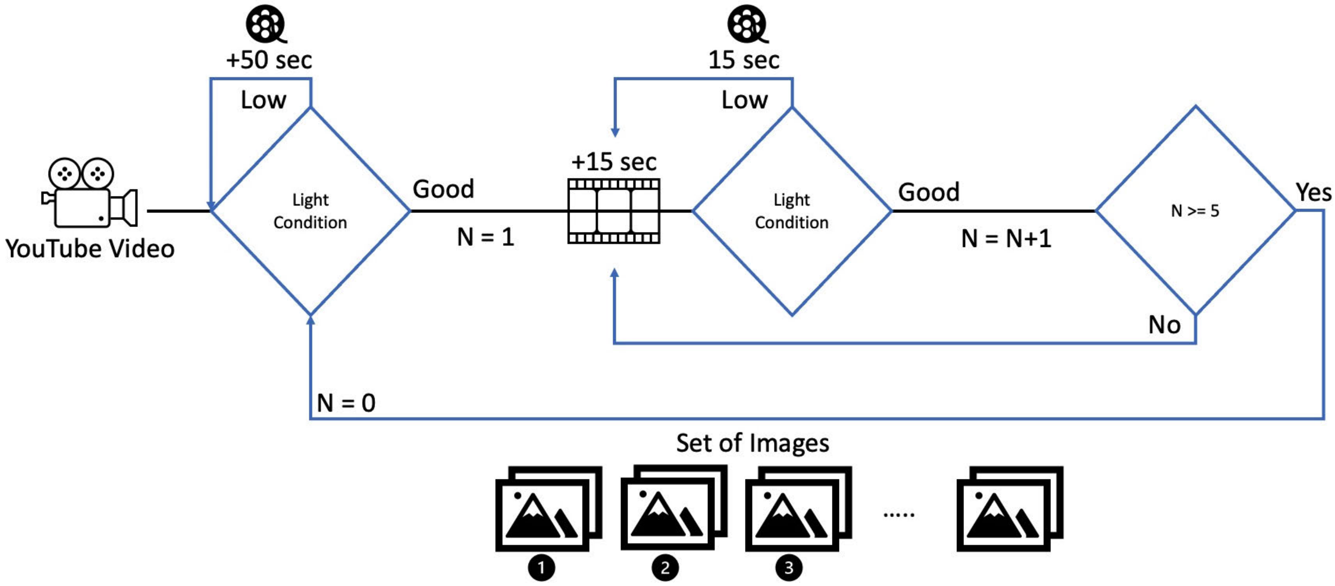


FIGURE 3. The overall process of collecting the open-source images for our proposed dataset. We require three conditions, lighting being an essential one for our image collection. If the light is low, we keep moving the video clip fifteen seconds forward until the lighting condition becomes standard; at this point, we select a frame and repeat the process until we collect five images. This procedure is performed manually. We repeat the process using videos from documentaries, movies and lifestyles to generate a large dataset of 5 sequential image sets.

the continuous stream of input images, an encoder using a 2D/3D CNN is utilized. A decoder, typically known as a language model based on a recurrent neural network [30], [31] or a transformer [32], captures these characteristics and transforms them into sentences in the chosen language. Standard video captioning methods use different strategies to boost the accuracy of the video captioning framework, such as object and action modelling [32], [33], [34], Fourier transform [35], attention mechanism [36], [37], and semantic attribute learning [38], [39].

C. STORYTELLING

1) DATASETS

On the Visual storytelling side, shifting from a single image to multi-images in context enables us to construct artificial intelligence (AI) that can logically deduce knowledge about a visual moment based on what it has previously seen. The first illustrations of utilizing a description that included multiple images were discovered in blog posts. While tourists visit destinations such as New York City and Disneyland, they capture several photos to memorialize the special moments. Researchers formulated multi-image analysis to learn the relationship between image streams and text sequences [40], [41]. To the best of our knowledge, visual storytelling (VIST) [4] is the first publicly available dataset that allows the training of models for generating stories from a set of images. The VIST dataset includes two categories: a description in isolation (DII) and a story in sequence (SIS). Both collections have five images and five corresponding sentences. However, the DII-type is constructed as a single image caption, meaning that all sentences are unrelated to a story. In comparison, the SIS-type is structured as a five-sentence story corresponding to a set of images as shown in Figure 2. Since 2016, all published research articles have been trained and evaluated only on the VIST dataset. Therefore, the ability of the published methods to generalize to other data has not been tested.

2) TECHNIQUES

Narration is one of the first behaviors humans have engaged in, and very recently, due to technological advances in computers and algorithms. It has also been the subject of a considerable proportion of empirical research. Visual storytelling approaches can assist in interpreting the activities shown in a set of images and describing these images in a single paragraph of multiple sentences [42], leveraged with ranking and retrieval networks, which concatenates sequence image features to generate their brief description [41]. The Coherent Recurrent Convolutional Network (CRCN) [43] was established to improve the fluency with which various phrases crossed a set of images. Critical components of this deep learning network include the Long Short Term Memory (LSTM) networks [31]. In addition, the reward functions method emphasizes the fundamental role that rewards perform in behaviourism [44]. Encoder-decoder architecture-based frameworks enhance visual characteristics and generate a story of multiple sequences [45], [46]. Furthermore, utilizing Recurrent Neural Networks (RNNs) incorporated with the Mogrifier technique to improve modulation has increased the relevance, coherence, and impressiveness of the constructed story [47], [48]. All current methods successfully generate grammatically correct stories from a set of images. However, these generated stories are not as good as human-written stories. The primary purpose of this study is to investigate and analyze the overall state-of-the-art achievements in storytelling methodologies in general, as well as to benchmark recent techniques on a new storytelling dataset.

III. DATASET CONSTRUCTION

This section covers the strategy behind the construction of our Sequential Storytelling Images Dataset (SSID). To initiate, we manually screen capture images (video frames) from open-source videos. Afterward, each of these sets of images was annotated through the Amazon Mechanical Turk (AMT) service. Details of the dataset organization are given below.

A. DATASET IMAGE DOMAIN

To generate multiple story-based sentences from a set of images, the first step is to construct a collection of images with story-based ground truths that can be used to train and evaluate a machine-learning model. YouTube is an excellent source for this purpose as it is the world's largest collection of videos. Within our proposed sequential images, we narrowed the emphasis of our video search to three categories: documentaries, lifestyle videos, and movies.

1) DOCUMENTARIES

Documentary videos are typically non-fictional motion pictures meant to depict reality, generally for instructing, educating or maintaining a historical record. Documentary videos can also be used as historical records. Such videos are often filmed in an informal style, which allows us to capture a better glimpse of naturalistic environments. In addition, we considered that most of the videos are shot during the day, ensuring that the screenshots we take have good lighting, as shown in Figure 4 (Documentary Image Sample).

2) LIFESTYLE VIDEOS

Lifestyle videos reflect everyday life and are intended to be photorealistic scenes. These videos show creators filming themselves going about their daily lives with their belongings, friends, and family members. We chose movies of people living different lifestyles, such as in a home, traveling and/or shopping. Figure 4 shows a sample of family trip images.

3) MOVIES

In this particular category, the essential factor is choosing a film that reflects a true story; hence, we did not include any science fiction, horror, fantasy, or cartoon videos. Therefore, silent and non-fiction films were chosen because they feature scenes with real people performing real-life activities. Figure 4 illustrates an example set from movies.

B. IMAGE CAPTURING PROCESS

Figure 3 illustrates the process of collecting the open-source images. First, video frames are played as

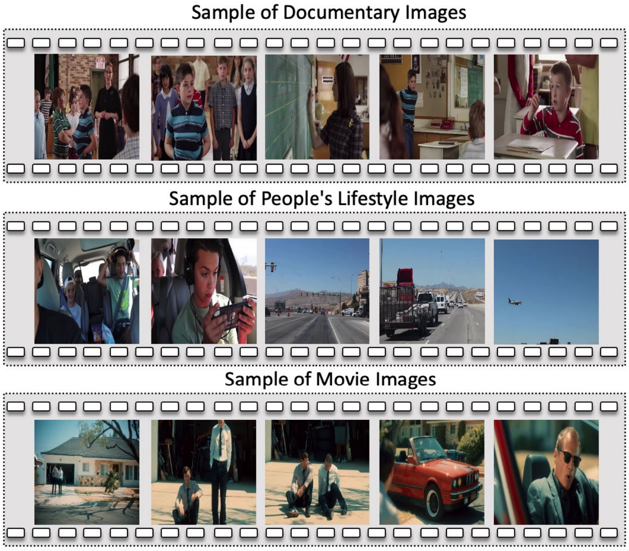

Sample of Documentary Images
Sample of People's Lifestyle Images
Sample of Movie Images

FIGURE 4. Our proposed dataset comprises image sets containing sequences of five images collected from three distinct video categories, namely, documentaries, lifestyle videos, and movies.

$$I_N = (I_1, I_2, \dots, I_N), \text{ where}$$
$$I = [I_1, I_2, \dots, I_N] \;\; s.t. \;\; I_N \in \mathbb{R}^{H\times W\times C}. \tag{1}$$

In the above equation, the term "I" represents the selected frames, i.e., (image) every fifteen seconds from the open-source video, where $N \in \{1, 2, 3, 4, 5\}$ is a collection of five images with H×W×C (Height × Width × Channels) shape that provides a distinctive representation of our proposed dataset. We use H×W×C = 224 × 224×3. During the collection of the dataset's images, the lighting conditions of the movie are essential. In the beginning, we manually check the lighting condition of the first image, and if it is below a certain threshold, we fast forward 50 seconds until we have N = 1 image with acceptable lighting. Next, we manually fast forward 15 seconds and check the lighting condition of the next image. If the lighting is above the threshold, we select that image, and if it is below the threshold, we fast forward 15 seconds again to ensure that the subsequent scene transition in the video accurately conveys the intended narrative, and it is imperative to present it in a manner that effectively portrays an image story. We repeat this process until our set has five images. Our final representation is a set of five images, which conveys stream-specific information and sequences of related visual activities. The procedures above are performed through human effort while carefully watching all video clips, resulting in a significant expenditure of time. Based on our personal experiences, it has been determined that the ideal duration for transitioning between scenes in a video and displaying the subsequent image is approximately 15 seconds.

C. IMAGE ANNOTATION PROCESS

The next stage was data annotation. We presented the sets of five images to crowd workers through AMT, asking them to write appropriate stories relevant to the images. Details are given below:

1) AMAZON MECHANICAL TURK (AMT)

AMT is a paid online tool that assists companies and researchers in annotating data by distributing it to workers around the world through the Internet. AMT has become very popular for generating labeled data for training and validating machine learning models. We used the image annotation service to create image descriptions as narration for our proposed dataset.

2) IMAGE REPRESENTATION

At AMT, the images of our proposed dataset were displayed in a web form as a succession of five images followed by five blank text boxes. Figure 4 is an example of ordering the dataset images, where each set of images represents a single movie scene. For the crowd workers to be able to complete the form, we required them to write their descriptions in a manner that is relevant, cohesive, informative, and story-like in their sentence structure. Each of the five sentences should be connected to reflect the images' subject matter. Additionally, we generated four ground truth stories for each set of five images. We randomized the order of the forms to ensure that multiple workers could create the narratives for a single set of images.

3) PRUNING STORIES

We read all responses (i.e., stories) carefully and accepted them only if they fulfilled our criteria outlined in Part III-C2. Otherwise, we rejected the story and sent it to other workers in the queue for rewriting. Overall, the workers on AMT spent around five to eleven minutes finishing each story. The complete annotation work was accomplished in 1,852 hours.

D. DATASET ORGANIZATION AND SAMPLES

The images obtained from videos on YouTube were saved in a folder with a unique ID for each image. For instance, images 1 through 5 provide the collection of the first set of images, while images 6 through 10 present the following set of images in a similar fashion. A JSON file organized all the images while including the ground truth corresponding to each image. The structure of the JSON file is as follows:

```
{
"annotations":
[[{"storylet_id": Int,
    "storytext": str,
    "youtube_image_id": int,
    "album_id": int,
    "story_id": int,
    "image_order": int }]
}
```

where *storylet_id* is the unique story identifier, and *storytext* is a string value consisting of the story's first sentence. *youtube_image_id*, *album_id*, and *story_id* are integer values containing the ID for each complete story. Finally, *image_order* represents the image's position within the set of images.

The SSID dataset is comprised of 17,365 images, which resulted in a total of 3,473 unique sets of five images. Each set of images is associated with four ground truths, resulting in a total of 13,892 unique ground truths (i.e., written stories). And each ground truth is composed of five connected sentences written in the form of a story. Table 1 summarizes the collection number of images in each split, including the total count of sentences in each division.

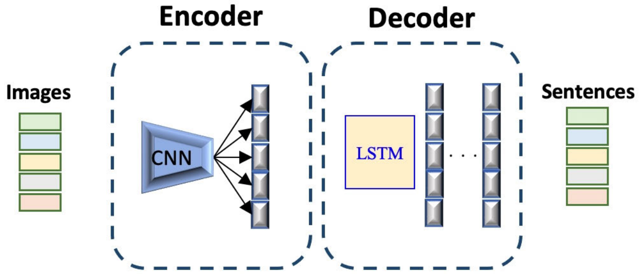


FIGURE 5. Most existing methods are encoder-decoder-based. The encoder uses a pre-trained CNN to extract visual information from the input images. The decoder component contains the language model, e.g., a Long Short-term Memory (LSTM) unit, where the story is formed.

IV. DATASET EXPERIMENTS SETTING

This section provides the most recent storytelling models in the literature, which we utilized to evaluate our proposed dataset. Furthermore, the training method and hardware are discussed, followed by different evaluation metrics.

A. EXISTING FRAMEWORK VARIANTS

The general architectural overview of most existing storytelling techniques is presented in Figure 5. The Encoder-Decoder technique drives the majority of these methodological approaches. First, a Convolutional Neural Network (CNN) is utilized to extract the features vector from the image. After that, a language model serves as a decoder to generate grammatically correct sentences. Overall, we choose the five most recent approaches to test our proposed dataset, which are as follows:

1) GLACNet

GLocal Attention Cascading Networks for Multi-image Cued Story Generation (GLACNet) [49] is the pioneering approach evaluated on the VIST dataset. The encoder-decoder approach is the key to this method's success. The encoder is composed of a ResNet-152 network, and its purpose is to extract a deep feature vector from a set of images. After that, the extracted features are successively input into the bi-LSTM to ensure that the image context is re-reflected accurately across the whole narration. Finally, a LSTM decoder implemented as a language model generates

TABLE 1. The proposed SSID dataset contains a total of 17,365 images collected from open-source videos. Additionally, each set of images (i.e., Album) is accompanied by at least five sentences of annotations. In addition, each set of images has four ground truth stories, resulting in 13892 unique stories.

Partition	distribution	# Image	# Album	# Sentence	# Story
Training Images	90%	15,625	3,125	62,500	12,500
Validation Images	5%	870	174	3,480	696
Testing Images	5%	870	174	3,480	696
Total	**100%**	**17,365**	**3,473**	**69,460**	**13,892**

a story consisting of five sentences representing one of the five input images.

2) CAMT

Encoder-decoder strategy, such as the one outlined for the GLACNet model in Section IV-A1, are also utilized by Contextualise, Attend, Modulate and Tell (CAMT) [47]. In addition, CAMT improves the story generated by combining the language model with a Mogrifier LSTM. This helped to obtain state-of-art results on the VIST dataset at the time of publication.

3) SAES

The theory of object detection techniques contributed to the overall improvement of the story that Semantic Attribute Enriched Storytelling (SAES) [48] generated from a sequence of images. That method employs YOLOv5 output represented as a multi-hot vector containing object detection information as well as noun attribute recognition. After that, these feature vectors are concatenated with the image features using ResNet-152. Using such as approach, the model has a more detailed representation of the images. As a direct result, the generated stories are improved in relevance, coherence and informational content.

4) ViT

Malakan et al., [50] replaced the standard CNN with Vision Transformer, which divides each image into 16 × 16 patches. The overall architecture includes the Vision Transformer (ViT) as the visual encoder, the Bidirectional-LSTM as a decoder and the standard LSTM unit enhanced by the Mogrifier LSTM. These modules work harmoniously to generate a story from a set of images and improve performance on several automatic evaluation metrics, including BLEU-2, BLEU-3, ROUGE-L, and METEOR.

B. EXPERIMENTAL SETUP DETAILS

We trained the models discussed in Section IV-A from scratch on our proposed dataset. The dataset provided a vocabulary size of 3,890 after being processed through a threshold of 8 for the minimal number of words. In addition, we tokenized each piece of extracted vocabulary with the assistance of the Natural Language Toolkit (NLTK) package in Python. During the training process, we utilized all the parameters proposed in [50] and configured them as: image feature size = 1024; dimension of word embedding = 265; the number of layers in LSTM = 2; and the number of Mogrifier steps in LSTM = 5. Finally, the learning rate is set at 10^{-3}, and the weight decay is set at e^{-5} to prevent overfitting during the training stage. Table 2 persents a summary of all parameters applied in these experiments.

TABLE 2. Experimental setup and parameters' details for each model trained on SSID dataset.

Parameter	Value	Parameter	Value
Vocab size	3,890	Image size	224
Image size	224×224	Image feature size	1024
Word embedding	256	LSTM layers	2
LSTM size	1024	Mogrification	5
Batch size	32	Learning rate	0.001
Weight decay	1e-4	Optimizer	Adam
Epochs	40	Log step	5

In order to ensure equitable experiments across all models, SSID dataset was randomly divided into three discrete subsets: a training set consisting of 90% of the data, a validation set consisting of 5%, and a testing set consisting of 5% of the data as detailed in Table 1. Each model was subjected to the same experimental setup during the training procedure. The same testing set was ultimately used to evaluate all results presented in Table 4.

C. HARDWARE USED

The training was performed on a desktop computer equipped with an Intel i9 3.60GHz processor with 16 cores and an RTX 2080 Ti NVIDIA GeForce graphics card with 12 GB of memory. We trained each model individually for over 40 epochs, using a 32-patch size, ensuring that we efficiently utilized the available RAM.

D. AUTOMATIC EVALUATION METRICS

Human judgment is the most reliable evaluation method for determining the quality of the stories that are machine-generated because of the challenging nature of the storytelling problem. However, human evaluation is time-consuming. Hence, automatic evaluation metrics, although problematic [54], are still beneficial for efficiently benchmarking the progress in this direction. In this study, we have selected the most popular automatic evaluation metrics as follows:

1) BLEU

a: BILINGUAL EVALUATION UNDERSTUDY

(BLEU) [55] evaluates the machine language models by utilizing n-grams to compare a set of reference texts to machine-generated text. It is considered the most effective

TABLE 3. A comparison of recently published methods on the Visual Storytelling Dataset (VIST). Quantitative results were obtained using seven different automated measures of evaluation. "-" indicates that the authors of the corresponding study did not publish the results. Higher scores represent higher accuracy, and the results in bold represent the best scores [50].

Model	Year	B-1	B-2	B-3	B-4	CIDEr	ROUGE-L	METEOR
AREL [44]	2018	0.536	0.315	0.173	0.099	0.038	0.286	0.352
GLACNet [49]	2018	0.56	0.321	0.171	0.091	0.041	0.264	0.306
HCBNet [45]	2019	0.59	0.348	0.191	0.105	0.051	0.274	0.34
HCBNet(w/o prev. sent. attention) [45]	2019	0.59	0.338	0.180	0.097	0.057	0.271	0.332
HCBNet(w/o description attention) [45]	2019	0.58	0.345	0.194	0.108	0.043	0.271	0.337
HCBNet(VGG) [45]	2019	0.59	0.34	0.186	0.104	0.051	0.269	0.334
ReCo-RL [51]	2020	-	-	-	0.124	**0.086**	0.299	0.339
BLEU-RL [51]	2020	-	-	-	0.144	0.067	0.301	0.352
VS with MPJA [52]	2021	0.601	0.325	0.133	0.082	0.042	0.303	0.344
CAMT [47]	2021	0.64	0.361	0.201	**0.184**	0.042	0.303	0.335
Rand+RNN [53]	2021	-	-	0.133	0.061	0.022	0.272	0.311
SAES Encoder-Decoder OD [48]	2021	0.64	0.363	0.196	0.106	0.051	0.294	0.330
SAES Encoder-Decoder OD & Noun [48]	2021	0.63	0.357	0.195	0.109	0.048	0.299	0.331
SAES Encoder OD [48]	2021	**0.65**	0.372	0.204	0.12	0.054	0.303	0.335
ViT model [50]	2022	0.63	**0.375**	**0.215**	0.123	0.044	**0.310**	**0.354**

TABLE 4. A comparison of recently published methods on our proposed dataset (SSID). Quantitative results were obtained using six different automated measures of evaluation. "-" indicates that the authors of the corresponding study did not publish the results. The higher scores represent higher accuracy, and the results in bold represent the best scores.

Model	Year	B-1	B-2	B-3	B-4	ROUGE-L	METEOR
GLACNet [49]	2018	0.256	0.118	0.049	0.023	0.141	0.160
CAMT [47]	2021	0.256	0.122	0.051	0.023	0.141	0.159
SAES Encoder-Decoder OD & Noun [48]	2021	0.262	0.110	0.039	0.013	0.142	0.162
SAES Encoder OD [48]	2021	**0.298**	**0.143**	**0.062**	**0.028**	**0.155**	**0.181**
ViT model [50]	2022	0.267	0.129	0.055	0.022	0.148	0.170

method for measuring the efficacy of approaches consisting of a few sentences and various variants. BLEU has multiple evaluation metrics, including BLEU-1, BLEU-2, BLEU-3, and BLEU-4, all chosen to evaluate our proposed dataset.

2) CIDEr

a: CONSENSUS-BASED IMAGE DESCRIPTION EVALUATION

(CIDEr) [56] is developed to compare the similarity of several reference sets of sentences to the machine-generated one. Furthermore, it is a significant measure for image captioning techniques because it captures the characteristics of language similarity, grammaticality, importance, saliency, and accuracy score of precision and recall.

3) ROUGE

a: RECALL-ORIENTED UNDERSTUDY FOR GISTING EVALUATION

(ROUGE) [57] comprises of several types, including ROUGE-1, ROUGE-2, ROUGE-W, ROUGE-L and ROUGE-SU4. We have selected ROUGE-L as reported in [57] to efficiently evaluate machine-generated stories with multiple sentences.

4) METEOR

a: METRIC FOR EVALUATION OF TRANSLATION WITH EXPLICIT ORDERING

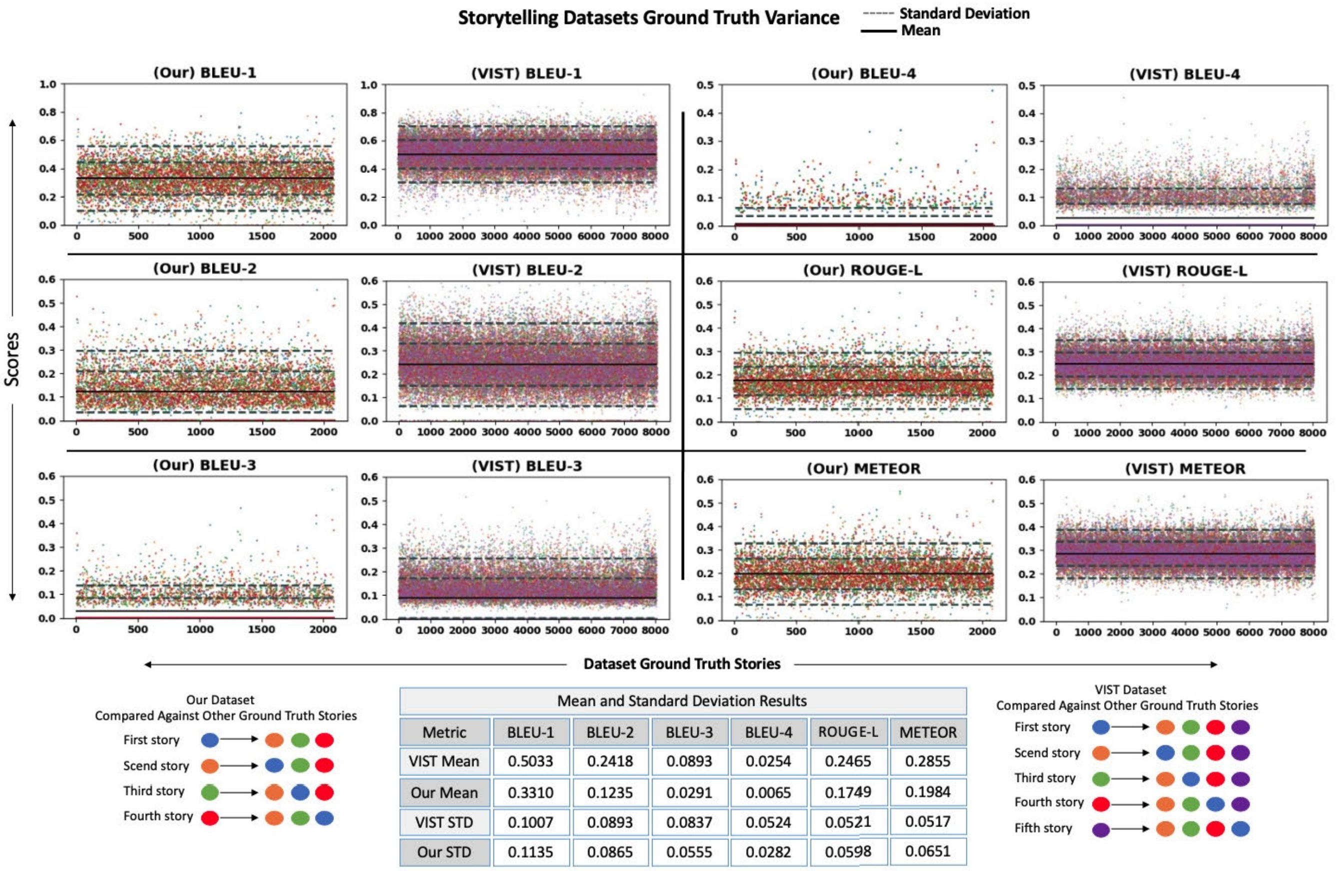
(METEOR) [58] assigns ratings to hypotheses for machine translation by aligning them with one or more reference translations. The alignment of words and phrases is determined by exact, stem, synonym, and paraphrased matches. Segment and system-level metric scores are computed based on the alignments between hypothesis-reference pairings. It is most suitable for assessing efficiency at the sentence level of the generated story since it analyses the words' synonyms and their matchings with the text references.

V. RESULT COMPARISON

We perform extensive experiments to evaluate our storytelling dataset. First, we analyze the datasets' annotations, including the average means and standard deviations for VIST and our proposed SSID datasets, a measurement of the datasets' human evaluations, and a comparison of a ground truth sample from VIST and our proposed SSID datasets, and conclude by illustrating state-of-the-art model results over the mentioned datasets.

A. HUMAN EVALUATION

Given the inherent flaws in the automatic evaluation metrics, we also conducted human evaluations to assess the quality of the ground truth stories of the two datasets, i.e., VIST and SSID. First, we used a random selection process to extract twenty-five ground truth stories from the VIST dataset and our proposed dataset. Using Google Form service, we conducted a survey in which all of these stories and the associated image sets were presented (one set at a time) to human evaluators. Based on the three criteria: coherence, relevance, and informativeness, fifty participants were tasked with ranking and evaluating each story on a scale from one to five (worst



Mean and Standard Deviation Results

Metric	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
VIST Mean	0.5033	0.2418	0.0893	0.0254	0.2465	0.2855
Our Mean	0.3310	0.1235	0.0291	0.0065	0.1749	0.1984
VIST STD	0.1007	0.0893	0.0837	0.0524	0.0521	0.0517
Our STD	0.1135	0.0865	0.0555	0.0282	0.0598	0.0651

FIGURE 6. The within ground truth variance in the VIST and our datasets. In VIST, each image set has five ground truth stories, whereas our dataset contains only four. We compare each ground truth story against the others. The means and standard deviations are also listed in the Table. The mean scores for our dataset are lower, indicating that our dataset has more diverse stories. Both datasets have high variance in the stories.

TABLE 5. Comparative analysis of the time complexity is performed in the Encoder-Decoder components of the investigated storytelling model using the floating-point operations per second (FLOPs). For each module, FLOPs is calculated separately and then the time complexity of each module is summed together to determine the total time complexity. "-" indicates that the authors of the corresponding study did not include the module in their methodology. The aforementioned terms denote the numerical values of B for Billion, M for Million, and K for Thousand.

Storytelling Model	Time Complexity in Encoder				Time Complexity in Dncoder		Total Time Complexity
	ResNet-152	ViT	Dectction	Bi-LSTM	Mogrifier-LSTM	LSTM Unit	
GLACNet [49]	11.3 B	-	-	2.15 B	-	4 K	13.45 B
CAMT [47]	11.3 B	-	-	2.15 B	4.19 B	4 K	17.64 B
SAES [48]	11.3 B	-	27 M	2.15 B	4.19 B	4 K	17.7 B
ViT model [50]	-	725 M	-	2.15 B	4.19 B	4 K	**6.35 B**

TABLE 6. Human evaluation survey investigations of Fifty ground truth stories extracted from the VIST dataset and our proposed dataset. Participants ranked each story from 1 to 5 (worst to best). Our proposed dataset surpasses the VIST dataset in every category, including relevance, coherence, and informative story.

Storytelling Dataset	Rank 1-5 (worst-best)		
	Relevance	Coherence	Informative
VIST [4]	3.73	3.81	3.87
SSID (ours)	**3.75**	**3.82**	**3.89**

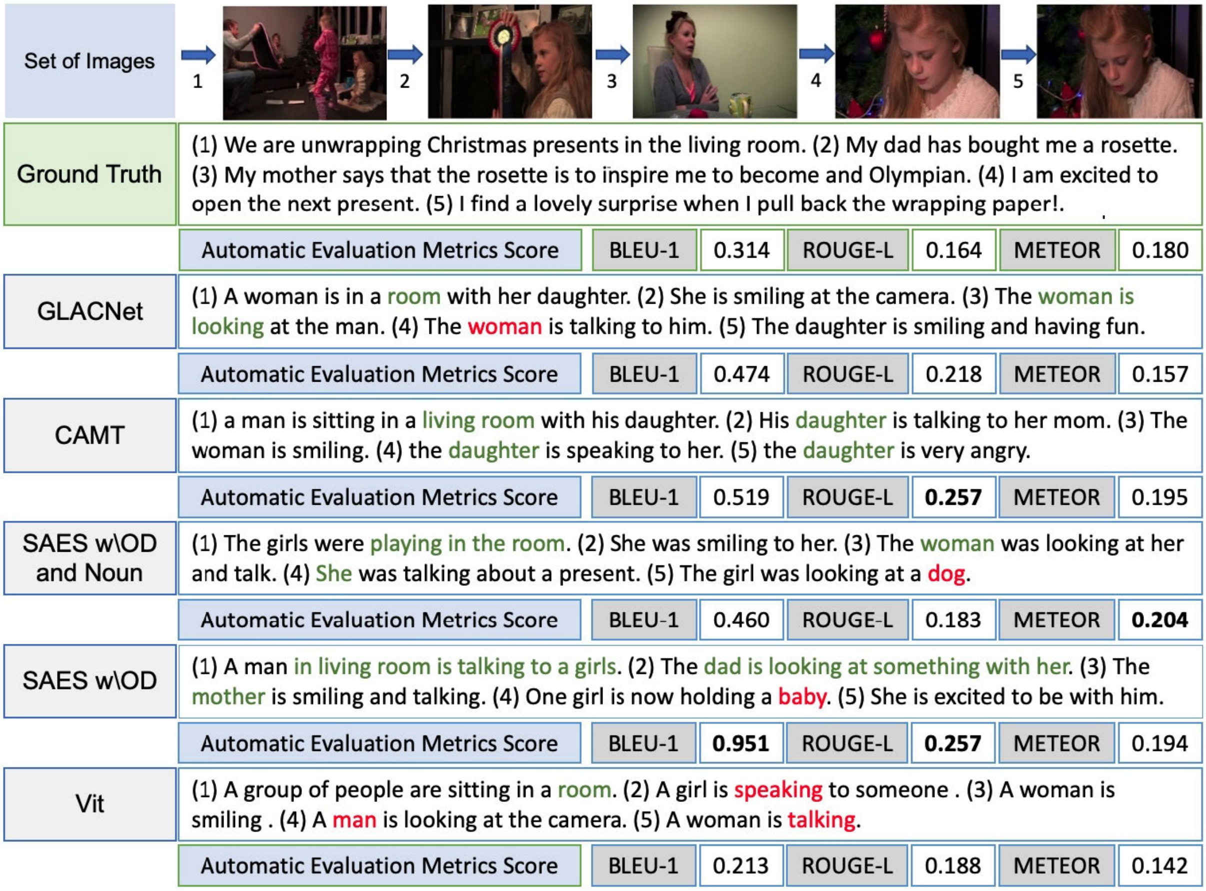
to best). The survey results are summarized in Table 6. These results demonstrate that both datasets received almost the same rating. However, our proposed dataset outweighs the VIST dataset in all three evaluation criteria.

B. EXISTING MODELS COMPARISON

To demonstrate the efficacy of our proposed dataset, we retrain and evaluate a wide range of existing state-of-the-art storytelling models discussed in section IV-A on SSID. The following are details of our experiments:

1) MACHINE GENERATED-STORY SAMPLES

Figure 7 presents predicted stories based on a set of images from our proposed dataset, as well as the ground truth story. It can be observed that the existing models can predict grammatically correct phrases containing a variety of narratives or concepts. The emphasized shades of green indicates that the generated sentence part is relevant to the image content. For instance, the first predicted sentence in each scenario contains


Set of Images
1
2
3
4
5

Ground Truth	(1) We are unwrapping Christmas presents in the living room. (2) My dad has bought me a rosette. (3) My mother says that the rosette is to inspire me to become and Olympian. (4) I am excited to open the next present. (5) I find a lovely surprise when I pull back the wrapping paper!.						
	Automatic Evaluation Metrics Score	BLEU-1	0.314	ROUGE-L	0.164	METEOR	0.180
GLACNet	(1) A woman is in a room with her daughter. (2) She is smiling at the camera. (3) The woman is looking at the man. (4) The woman is talking to him. (5) The daughter is smiling and having fun.						
	Automatic Evaluation Metrics Score	BLEU-1	0.474	ROUGE-L	0.218	METEOR	0.157
CAMT	(1) a man is sitting in a living room with his daughter. (2) His daughter is talking to her mom. (3) The woman is smiling. (4) the daughter is speaking to her. (5) the daughter is very angry.						
	Automatic Evaluation Metrics Score	BLEU-1	0.519	ROUGE-L	**0.257**	METEOR	0.195
SAES w\OD and Noun	(1) The girls were playing in the room. (2) She was smiling to her. (3) The woman was looking at her and talk. (4) She was talking about a present. (5) The girl was looking at a dog.						
	Automatic Evaluation Metrics Score	BLEU-1	0.460	ROUGE-L	0.183	METEOR	**0.204**
SAES w\OD	(1) A man in living room is talking to a girls. (2) The dad is looking at something with her. (3) The mother is smiling and talking. (4) One girl is now holding a baby. (5) She is excited to be with him.						
	Automatic Evaluation Metrics Score	BLEU-1	**0.951**	ROUGE-L	**0.257**	METEOR	0.194
Vit	(1) A group of people are sitting in a room. (2) A girl is speaking to someone . (3) A woman is smiling . (4) A man is looking at the camera. (5) A woman is talking.						
	Automatic Evaluation Metrics Score	BLEU-1	0.213	ROUGE-L	0.188	METEOR	0.142

FIGURE 7. Variations in the stories generated by the most prevalent storytelling techniques for the same set of images after training them on our proposed dataset. The highlighted green shows that these words are relevant to the associated image. The highlighted red indicates words that do not apply to the corresponding image or are commonly predicted words. Automatic evaluation metrics BLEU-1, ROUGE-L, and METEOR are provided to compare the relative performance.

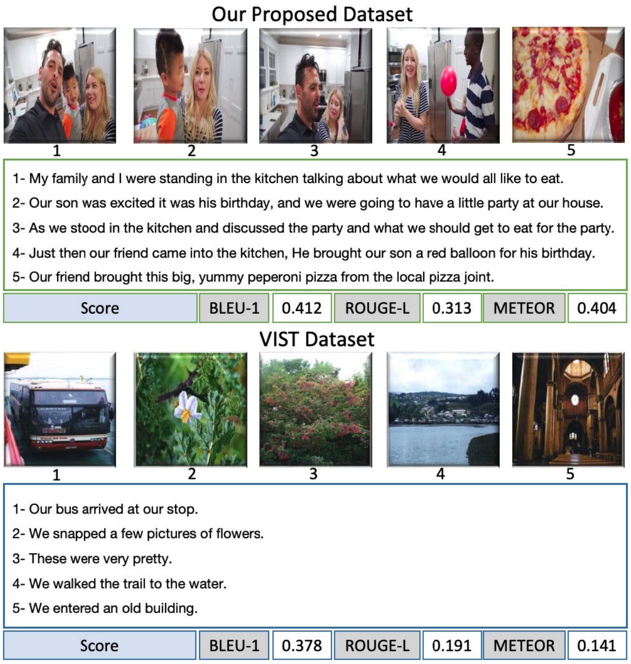
the word "*room*", which accurately reflects the first image. Similarly, SAES [48], which has improved with the object detection model, can correctly relate the second sentence to the story by showing that a female is associated with an object.

The emphasized red indicates that the generated sentence part does not correlate with the corresponding image. For instance, SAES [48] with object detection and a noun attribute model predicts the word "*dog*" in the fifth sentence, despite the reality that the fifth image is not related to the word "*dog*". Similarly, ViT's model predicts the word "*speaking*" in the second sentence, the word "*man*" in the fourth sentence, and the word "*talking*" in the fifth sentence, which results in an incoherent story. Due to the difficulties above, it becomes evident that our proposed dataset is more challenging and necessitates a model that can detect numerous events accurately.

2) MACHINE GENERATED-STORY SCORES

For quantitative comprehension analysis, Figure 7 demonstrates the performance of each story generated by various approaches in addition to the ground truth. First, we compared the automatic evaluation metrics, BLEU-1, ROUGE-L, and METEOR, to each generated story. The SAES w/OD model outperforms all other models, scoring 0.951 points in BLEU-1 and 0.257 points in ROUGE-L. The CAMT model achieves a score of 0.257 in the ROUGE-L metric, which is identical to the performance of the SAES w/OD model. However, the CAMT model does not perform sufficiently in the other metrics. In addition, SAES improved via object detection and noun attribute learning, obtaining 0.204 points, which allowed it to succeed in the METEOR metric and surpass all other models.

Each image set in the proposed dataset has four ground truth annotations (Human Generated Story or HGS). For illustrative and comparative analysis, one of these HGSs is showcased in Figure 7 (Green box), which is obtained by comparing it to other HGSs of the considered image set. Compared to the other stories generated by state-of-the-art techniques, the HGS story received the worst performance across all of the specified automatic evaluation metrics. This indicates that all models can learn and fully comprehend the ground truths provided during models' training. As a summary of the analysis, a wide variety of HGS, also known as high variance stories, would assist fundamental algorithms such as sequence encoder-decoders in learning and



1- My family and I were standing in the kitchen talking about what we would all like to eat.
2- Our son was excited it was his birthday, and we were going to have a little party at our house.
3- As we stood in the kitchen and discussed the party and what we should get to eat for the party.
4- Just then our friend came into the kitchen, He brought our son a red balloon for his birthday.
5- Our friend brought this big, yummy peperoni pizza from the local pizza joint.

Score	BLEU-1	0.412	ROUGE-L	0.313	METEOR	0.404

1- Our bus arrived at our stop.
2- We snapped a few pictures of flowers.
3- These were very pretty.
4- We walked the trail to the water.
5- We entered an old building.

Score	BLEU-1	0.378	ROUGE-L	0.191	METEOR	0.141

FIGURE 8. A comparative case between the VIST dataset and our proposed dataset. Automatic evaluation metrics such as BLEU-1, ROUGE-L, and METEOR are presented for both samples.

comprehending more effectively to generate better stories based on sequential vision.

3) STATE-OF-THE-ART RESULTS COMPARISON

Table 3 presents the state-of-the-art storytelling scores on the VIST dataset. It shows that ViT outperforms other storytelling approaches for ROUGE-L, METEOR, BLEU-2 and BLEU-3. In comparison, CAMT continues to perform superior in BLEU-4, scoring 0.184 points. Moreover, ReCo-EL improved through multiple Recurrent neural networks and achieved the highest score of 0.086 for CIDEr.

To determine how effectively the storytelling frameworks will perform on our proposed dataset, we retrained and evaluated the models that are publicly available. Table 4 shows that the SAES model achieved superior results in all of the automatic evaluation metrics that were investigated, including BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L and METEOR. In general, the scores of SAES reduced on our SSID dataset (Table 4) compared to the VIST dataset (Table 3). This shows that our proposed dataset is more challenging and necessitates a more advanced model that can distinguish between each visual object, activity, and event and then connect all of these components into a single coherent story.

VI. DETAILED DISCUSSION

This section of the paper has undertaken a comprehensive analysis of the performance of the existing storytelling models that were experimentally evaluated. To better understand the computational requirements of processing visual description, the time complexity of each model was meticulously examined. Furthermore, both storytelling datasets were subjected to a rigorous analysis, utilizing standard statistical measures such as the mean and standard deviation. Finally, a qualitative analysis of the ground truth samples from both datasets was conducted to gain further insights into the quality and accuracy of the datasets.

A. TIME COMPLEXITY OF EXISTING TECHNIQUE

In this study, we compare the time complexity of the Encoder-Decoder parts of each storytelling model using floating-point operations per second (FLOPs). The initial stage of visual representation involves the extraction of features from a set of images. Specifically, ResNet-152 [59] exhibits a time complexity of approximately 11.3 billion FLOPs per image when employing a 224 × 224 input size. For Vision Transformer (ViT) [60], the time complexity of the linear projection FLOPs is calculated as (p^2d), where p denotes the dimensionality of the resized image patch (in this case, $p = 16$), and d refers to the dimensionality of the patch embeddings (in this case, $d = 768$). The total FLOPs for all image patches equal (np^2d), where n is the number of patches. The final time complexity of image size of 224 × 224 with a patch size of 16 × 16 is approximately 724,775,936 FLOPs. The object detection component employs the YOLOv5 algorithm [61], characterized by a computational complexity of 27 million FLOPs per image. Given the batch size of one embodiment, the total computational effort required for processing a single batch of images using YOLOv5 can be expressed as 27 million FLOPs per batch. The last part of the encoder is the bi-LSTM. According to the experimental sittings mentioned in Section IV-B, the total time complexity of sequence length of 64 is $(2 \times 1024)^2 \times 64 \times 32) = (2, 147, 483, 648)FLOPs$. The last component is the decoder utilizing Mogrifier-LSTM [62] with 5 rounds of mogrifications. The time complexity with a sequence length of 64 can be approximated as $(N^2 \times D^2)$, where N is the number of time steps, and D is the hidden size. This can be calculated as follows $(64^2 \times 1024^2) = (4, 194, 304, 000)FLOPs$. The last part of the decoder is the standard LSTM [63] unit to generate the story. The time complexity can be donated as (N^2), where N is each time step that requires two matrix multiplications (one for the input and one for the previous state) of size $N \times N$. As the sequence length is 64, the time complexity is calculated as $(64^2) = (4096)FLOPs$.

Table 5 presents a comprehensive comparison of the time complexity of the experimented models. Notably, SAES model demonstrates the highest FLOPs of 17.7 billion, suggesting a significantly increased computational complexity level. Conversely, ViT model exhibits a considerably lower time complexity of 6.35 billion FLOPs which is approximately one-third of SAES model. Despite this difference, the evaluation results in Table 4 indicate that ViT model performs comparable to the SAES model. It is worth noting that SAES model is an intricate architecture that incorporates both object detection and noun attribute. It exhibits high time complexity and achieves the highest evaluation results among the models compared in Table 4.

B. STORYTELLING DATASETS ANALYSIS

Due to the brilliance of the human mind, one set of images can inspire multiple narratives. As a result, a high-variance image description can be produced by constructing a storytelling dataset consisting of five ground truth stories. In this experiment, we developed a loop method to extract the multiple ground truth stories for each set of images from the VIST and our proposed dataset to analyze the variations within the ground truth stories. We automatically compared one story from each group to the remaining ones. For instance, for each set of images, the first ground truth story is taken and compared to the rest of the ground truth stories for each dataset. This is then repeated for the second story, and so on. Figure 6 illustrates the results of this experiment. We report all the automatic evaluation metrics scores (discussed in Section IV-D) for each set of stories. These scores include BLEU-1 to 4, ROUGE-L, and METEOR. From these scores, the average means and standard deviations are also summarized in a table in Figure 6.

Plots in Figure 6 present the ground truth variances for both VIST and our proposed datasets for each automatic evaluation metric. In both datasets, the conceptual ground truth appears to have a significant variability according to all the metrics. To quantitatively analyze the VIST dataset and our proposed dataset, the Table in Figure 6 depicts each automatic evaluation parameter's mean and standard deviation. Based on the findings, we observe that the VIST dataset's ground truths achieved the highest average of means across all evaluated metrics, including BLEU-1 to 4, ROUGE-L, and METEOR. Our proposed dataset receives a lower average of means, indicating that it contains a broader diversity of ground truths. Based on the average of standard deviations, the consistency of both datasets is significantly close to each other; however, our proposed dataset obtains less in the average of standard deviations across BLEU-2, BLEU-3, and BLEU-4, indicating that the majority of ground truths are annotated slightly different. In particular, for the most contributed indicator at the sentence level [58], METEOR, VIST dataset showed higher averages in both average of means and standard deviations than our dataset indicates higher structured sentences. In summary, these results indicate that our proposed dataset contains scenarios (i.e., ground truth stories) with a high variation due to the natural manner of story authoring. In addition, the conducted experiment indicates that our proposed dataset is more challenging and, as a result, requires more effective storytelling strategies.

C. DATASETS GROUND TRUTH ANALYSIS

Figure 8 is an illustration of human-written ground truth extracted from a set of images from our proposed dataset and VIST dataset. Both sets of images feature diverse visual content, allowing for the creation of distinct stories. However, our ground truth is more informative and challenging. For instance, image two in our story indicates that there will be a birthday party, which the image itself does not convey but is expected by image four, which depicts a person holding a balloon. Finally, all ground truths are evaluated alongside linked references to demonstrate their relevance and consistency. Our proposed dataset's ground truth scores around 10 points higher than BLEU and ROUGE-L measures and approximately 26 points higher than VIST in METEOR.

VII. LIMITATION AND FUTURE OUTLOOK

It is important to mention that this study has its limitations. The most prominent of which pertains to the time-consuming nature of the image construction process. The selection of images for each visual sequence story required careful consideration during the capture phase from YouTube videos. Each image set was manually selected to ensure that it accurately represented a cohesive and meaningful story. Furthermore, the size of the dataset is limited, which may have implications for the overall performance of machine learning models trained on it. Despite these limitations, the present study serves as an essential starting point for researchers seeking to improve the sequential vision description dataset. This work may involve expanding the size of the storytelling dataset, which could enhance the accuracy and effectiveness of machine learning models trained on the dataset.

In the future, one of our goals is to increase the size of our dataset from YouTube by including additional images to cover more visual content. These images will be annotated by individuals with higher proficiency in English or rephrased using a large language model such as GPT-4 [64] so that the sentences are well connected and the story is coherent in the ground truth annotations. In addition, approaches to storytelling that can establish various events and activities with a stream of visual scenes are expected to be investigated.

VIII. CONCLUSION

This research introduced a new storytelling dataset contributing to the multi-image description technique. Our storytelling dataset consists of image sets manually taken from Youtube videos in specific areas, such as documentaries, people's Lifestyles and movies. We leveraged Amazon Mechanical Turk (AMT) service to annotate these image sets as story descriptions. After exhaustive studies on VIST and our proposed dataset, we discovered a correlation with significant variance between both datasets. In addition, existing cutting-edge algorithms are trained and assessed using our proposed dataset. In conclusion, the exhaustive analysis and experiment reveal that our dataset is more challenging and necessitates advanced storytelling approaches to portray the relationship between a set of images.

REFERENCES

[1] Y. Li, Y. Pan, T. Yao, and T. Mei, "Comprehending and ordering semantics for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17969–17978.

[2] T. do Carmo Nogueira, C. D. N. Vinhal, G. da Cruz Júnior, and M. R. D. Ullmann, "Reference-based model using multimodal gated recurrent units for image captioning," *Multimedia Tools Appl.*, vol. 79, nos. 41–42, pp. 30615–30635, Nov. 2020.

[3] Z. Fei, X. Yan, S. Wang, and Q. Tian, "DeeCap: Dynamic early exiting for efficient image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12206–12216.

[4] T.-H. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell, "Visual storytelling," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1233–1239.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[6] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30K: Multilingual English-German image descriptions," 2016, *arXiv:1605.00459*.

[7] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, Y. Wang, and Y. Wang, "Large-scale datasets for going deeper in image understanding," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1480–1485.

[8] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Dec. 2014.

[9] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2641–2649.

[10] M. Everingham et al., "The 2005 PASCAL visual object classes challenge," in *Proc. Mach. Learn. Challenges Workshop*. Berlin, Germany: Springer, 2005, pp. 117–176.

[11] A. F. Biten, L. Gomez, M. Rusiñol, and D. Karatzas, "Good news, everyone! Context driven entity-aware captioning for news images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12458–12467.

[12] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The IAPR TC-12 benchmark: A new evaluation resource for visual information systems," in *Proc. Int. Workshop ontoImage*, vol. 2, 2006, pp. 1–11.

[13] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning images taken by people who are blind," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 417–434.

[14] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, "Nocaps: Novel object captioning at scale," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8947–8956.

[15] X. Yang, H. Zhang, D. Jin, Y. Liu, C.-H. Wu, J. Tan, D. Xie, J. Wang, and X. Wang, "Fashion captioning: Towards generating accurate descriptions with semantic rewards," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 1–17.

[16] A. Mogadala, X. Shen, and D. Klakow, "Integrating image captioning with rule-based entity masking," 2020, *arXiv:2007.11690*.

[17] B. B. Phukan and A. R. Panda, "An efficient technique for image captioning using deep neural network," 2020, *arXiv:2009.02565*.

[18] L. Huang, W. Wang, J. Chen, and X. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4633–4642.

[19] X. Shen, B. Liu, Y. Zhou, J. Zhao, and M. Liu, "Remote sensing image captioning via variational autoencoder and reinforcement learning," *Knowl.-Based Syst.*, vol. 203, Sep. 2020, Art. no. 105920.

[20] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8927–8936.

[21] J. Wang, W. Wang, L. Wang, Z. Wang, D. D. Feng, and T. Tan, "Learning visual relationship and context-aware attention for image captioning," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107075.

[22] S. Ding, S. Qu, Y. Xi, A. K. Sangaiah, and S. Wan, "Image caption generation with high-level image features," *Pattern Recognit. Lett.*, vol. 123, pp. 89–95, May 2019.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[24] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1194–1201.

[25] P. Das, C. Xu, R. F. Doell, and J. J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2634–2641.

[26] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding action descriptions in videos," *Trans. Assoc. Comput. Linguistics*, vol. 1, pp. 25–36, Dec. 2013.

[27] A. Torabi, C. Pal, H. Larochelle, and A. Courville, "Using descriptive video services to create a large data source for video annotation research," 2015, *arXiv:1503.01070*.

[28] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 510–526.

[29] S. Gella, M. Lewis, and M. Rohrbach, "A dataset for telling the stories of social media videos," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 968–974.

[30] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[33] B. Pan, H. Cai, D. Huang, K. Lee, A. Gaidon, E. Adeli, and J. C. Niebles, "Spatio-temporal graph for video captioning with knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10867–10876.

[34] Q. Zheng, C. Wang, and D. Tao, "Syntax-aware action targeting for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13093–13102.

[35] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12479–12488.

[36] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, and Q. Dai, "STAT: Spatial-temporal attention mechanism for video captioning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 229–241, Jan. 2020.

[37] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4507–4515.

[38] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1141–1150.

[39] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 984–992.

[40] C. C. Park and G. Kim, "Expressing an image stream with a sequence of natural sentences," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[41] G. Kim, S. Moon, and L. Sigal, "Ranking and retrieval of image sequences from multiple paragraph queries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1993–2001.

[42] P. W. Wiessner, "Embers of society: Firelight talk among the Ju/'hoansi Bushmen," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 39, pp. 14027–14035, Sep. 2014.

[43] C. C. Park, Y. Kim, and G. Kim, "Retrieval of sentence sequences for an image stream via coherence recurrent convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 945–957, Apr. 2018.

[44] X. Wang, W. Chen, Y.-F. Wang, and W. Y. Wang, "No metrics are perfect: Adversarial reward learning for visual storytelling," 2018, *arXiv:1804.09160*.

[45] M. Nahian, S. Al, T. Tasrin, S. Gandhi, R. Gaines, and B. Harrison, "A hierarchical approach for visual storytelling using image description," in *Proc. Int. Conf. Interact. Digit. Storytelling*. Cham, Switzerland: Springer, 2019, pp. 304–317.

[46] Q. Huang, Z. Gan, A. Celikyilmaz, D. Wu, J. Wang, and X. He, "Hierarchically structured reinforcement learning for topically coherent visual story generation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8465–8472.

[47] Z. Malakan, N. Aafaq, G. Hassan, and A. Mian, "Contextualise, attend, modulate and tell: Visual storytelling," in *Proc. 16th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, vol. 5, 2021, pp. 196–205.

[48] Z. M. Malakan, G. M. Hassan, M. A. A. K. Jalwana, N. Aafaq, and A. Mian, "Semantic attribute enriched storytelling from a sequence of images," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2021, pp. 1–8.

[49] T. Kim, M.-O. Heo, S. Son, K.-W. Park, and B.-T. Zhang, "GLAC net: GLocal attention cascading networks for multi-image cued story generation," 2018, *arXiv:1805.10973*.

[50] Z. M. Malakan, G. M. Hassan, and A. Mian, "Vision transformer based model for describing a set of images as a story," in *Proc. Australas. Joint Conf. Artif. Intell.* Cham, Switzerland: Springer, 2022, pp. 15–28.

[51] J. Hu, Y. Cheng, Z. Gan, J. Liu, J. Gao, and G. Neubig, "What makes a good story? Designing composite rewards for visual storytelling," in *Proc. AAAI*, 2020, pp. 7969–7976.

[52] Y. Guo, H. Wu, and X. Zhang, "Steganographic visual story with mutual-perceived joint attention," *EURASIP J. Image Video Process.*, vol. 2021, no. 1, pp. 1–14, Dec. 2021.

[53] H. Chen, Y. Huang, H. Takamura, and H. Nakayama, "Commonsense knowledge aware concept selection for diverse and informative visual storytelling," 2021, *arXiv:2102.02963*.

[54] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, and M. Shah, "Video description: A survey of methods, datasets, and evaluation metrics," *ACM Comput. Surv.*, vol. 52, no. 6, pp. 1–37, Nov. 2020.

[55] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.

[56] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.

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

[58] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[61] G. Jocher et al., "ultralytics/yolov5: v7.0—YOLOv5 SOTA real-time instance segmentation," Zenodo, USA, Tech. Rep., 2022, doi: 10.5281/zenodo.7347926.

[62] G. Melis, T. Kočiskỳ, and P. Blunsom, "Mogrifier LSTM," 2019, *arXiv:1909.01792*.

[63] A. Graves and A. Graves, "Long short-term memory," in *Supervised Sequence Labelling With Recurrent Neural Networks*. Berlin, Germany: Springer, 2012, pp. 37–45.

[64] *GPT-4 Technical Report*, OpenAI, USA, 2023.

ZAINY M. MALAKAN received the B.S. degree in information science from Umm Al-Qura University (UQU), Makkah, Saudi Arabia, and the M.Sc. degree in information systems specializing in vision tracking from Monmouth University, West Long Branch, NJ, USA. He is currently pursuing the Ph.D. degree in computer science, specializing in computer vision, with The University of Western Australia. He was a Lecturer with the Department of Information Science, College of Computers and Information Systems, UQU. His research interests include sequence vision understanding, object detection, video analysis, data sciences, machine learning, scene recognition, and localization and tracking.

SAEED ANWAR received the master's degree (Hons.) from the Erasmus Mundus Vision and Robotics (Vibot), jointly offered by Heriot-Watt University, U.K., the University of Girona, Spain, and the University of Burgundy, France, and the Ph.D. degree from The Australian National University and National ICT Australia. He was with NICTA and CSIRO's Data61, Australia. He is currently an Assistant Professor with the King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia. He also holds honorary positions with The Australian National University (ANU), The University of Technology Sydney (UTS), and the University of Canberra, Australia. He has a strong teaching experience in many reputed universities and a substantial industry presence. He leads commercial projects and supervises Ph.D., M.S., and B.S. students. He has published in top-tier conferences and journals, including One Best Paper Nomination in CVPR and a Best Paper Honorable Mention in *Pattern Recognition* (PR).

GHULAM MUBASHAR HASSAN (Senior Member, IEEE) received the B.S. degree from the University of Engineering and Technology, Peshawar, Pakistan, the M.S. degree from Oklahoma State University, USA, and the Ph.D. degree from The University of Western Australia (UWA). He is currently a Faculty Member with the Department of Computer Science and Software Engineering, UWA. His research interests include artificial intelligence, machine learning, and their applications in multidisciplinary problems. He was a recipient of multiple teaching excellence and research awards.

AJMAL MIAN (Senior Member, IEEE) is currently a Professor of computer science with The University of Western Australia. His research interests include computer vision, deep learning, shape analysis, face recognition, human action recognition, and video analysis. He was a recipient of three prestigious national fellowships from the Australian Research Council and several awards, including the HBF Mid-Career Scientist of the Year 2022 Award, the West Australian Early Career Scientist of the Year 2012 Award, the Excellence in Research Supervision Award, the EH Thompson Award, the ASPIRE Professional Development Award, the Vice-Chancellors Mid-Career Award, the Outstanding Young Investigator Award, the Australasian Distinguished Dissertation Award, and various best paper awards. He served as the General Chair for DICTA 2019 and ACCV 2018. He is a Senior Editor of IEEE Transactions on Neural Networks and Learning Systems and an Associate Editor of IEEE Transactions on Image Processing and *Pattern Recognition*.