Learning to Rank Visual Stories From Human Ranking Data

Anonymous ACL submission

Abstract

Visual storytelling (VIST) is a typical vision and language task that has seen extensive development in the natural language generation research domain. However, it remains unclear whether conventional automatic evaluation metrics for text generation are applicable on VIST. In this paper, we present the VHED (VIST Human Evaluation Data) dataset, which first re-purposes human evaluation results for automatic evaluation; hence we develop Vrank (VIST ranker), a novel reference-free VIST metric for story evaluation. We first show that the results from commonly adopted automatic metrics for text generation have little correlation with those obtained from human evaluation, which motivates us to directly utilize human evaluation results to learn the automatic evaluation model. In the experiments, we evaluate the generated texts to predict story ranks using our model as well as other reference-based and reference-free metrics. Results show that Vrank prediction is significantly more aligned to human evaluation than other metrics with almost 30% higher accuracy when ranking story pairs. Moreover, we demonstrate that only Vrank shows human-like behavior in its strong ability to find better stories when the quality gap between two stories is high. Finally, we show the superiority of Vrank by its generalizability to pure textual stories, and conclude that this reuse of human evaluation results puts Vrank in a strong position for continued future advances.

1 Introduction

In visual storytelling (VIST) (Huang et al., 2016), a generation model tells a short story to describe the given five images. Automatic generation of visual stories is challenging because it has the complexity of cross-modal understanding with the diversity and sophistication of creative writing (Zhu et al., 2020). Extensive efforts in model developments have decreased the distance between machine-generated and human-written stories, but research on VIST evaluation remains stagnant.

Reference: i decided my dog would like a train ride. off to the train station we go. this is the train we will be taking our short trip on. my friend is the conductor. he is getting ready to attach the cars. here is the train all together. as you can see, my dog had a fantastic time.

Model 1 (BLEU-1: 0.605, Human Rankers: 👍)
the city was very busy. there were many different kinds of bikes. some were very unique. they were all very fast. i had a great time.

Model 2 (BLEU-1: 0.354, Human Rankers: 👍👍👍👍)
i went to the park station. it was a train trip to the museum. the train was very long. we had to go on our way out of the trains. this dog is so happy to see us.

Figure 1: Ranking of two stories generated by Model 1 and 2, by human rankers versus BLEU-1 score. BLEU-1 mispredicts due to unreasonable matches, correlating poorly with human ranking judgment.

Automatic metrics and human evaluation are widely used to examine natural language generation. Traditional n-gram-based or reference-based autometrics such as BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), and METEOR (Banerjee and Lavie, 2005) are common for VIST evaluation. However, preliminary findings have shown that these metrics have many drawbacks and hence are incompatible with VIST (Wang et al., 2018b). In particular, they assume that human-written stories are always better than machine-generated stories, limiting the advance of models yet not conforming to our observation on human judgment. Rethinking this postulation in evaluation, we believe the dependence on references should be minimized and human evaluation results should be fully utilized instead, because human judgements contain more meaningful signals. Recent hybrid and reference-free metrics such as BLEURT (Sellam et al., 2020) and UNION (Guan and Huang, 2020) have not yet been implemented or studied in VIST. Nevertheless, BLEURT utilizes few human results in finetuning, and UNION still regards human references as gold labels, which results in poor correlation to human judgment. On the other hand, human evaluations are relatively reliable for performance reports, and recent studies often include them to provide more convincing experimental results (Hsu et al., 2020, 2021a,b). However, human evaluations are expensive, time-consuming, and difficult to reproduce. Therefore, results should be recycled to benefit future evaluations.

Accordingly, we re-collected the human evaluation results from multiple published papers and organized the data into story pairs (Wei and Jia, 2021) as the VHED (VIST Human Evaluation Data) dataset. We then re-purposed VHED to create a better metric for VIST named Vrank (VIST Ranker). Vrank is a reference-free SIMCSE (Gao et al., 2021) based metric trained on VHED to learn to rank visual stories. We believe a storytelling metric should be independent of the references because stories are highly diverse by nature (Zhu et al., 2020), and it is reasonable for them to be dissimilar to the references (Guan and Huang, 2020; Wang et al., 2018b), as shown in Fig. 1. The story generated by Model 1 is assigned a higher BLEU score because larger portions of text overlap with the reference. However, human rankers recognize description in isolation and object detection error in Model 1, and instead rank Model 2 better. We conduct experiments to show that Vrank is superior to existing metrics, many of which lack properties essential to evaluating stories in a human-like fashion.

Therefore, we utilize VHED to understand and analyze human judgment in evaluating visual stories, and to provide additional metric assessments to reveal the shortcomings of existing metrics. The metric assessment experiments are conducted as the story-pair ranking task in which two stories are ranked based on their story quality. We observe three characteristics and design corresponding assessments to demonstrate Vrank's merits. First, larger rank differences in story quality are easier for people to differentiate. We measure the performance of metrics in story pairs with large gaps versus small gaps to determine whether all metrics have this property. Our assessment indicates this property is exclusively hold by Vrank. Second, human-written stories are not always better than machine-written stories. Indeed, 38% of machine-generated stories are better than the references, which suggests that the afore-mentioned assumption may need to be revisited (Clark et al., 2021). We examine the ability of metrics to rank such human-machine pairs, which Vrank performs relatively well. Finally, most generated stories still contain many errors, which serve as signals for human rankers (Modi and Parde, 2019). Hence we evaluate the ability of metrics to detect errors and show that Vrank is a better indicator of errors. Also, we show that Vrank is able to generalize to other datasets without bias to VHED. In conclusion, Vrank excels in the above assessments and able to follow human behaviors in ranking, rank machine and human stories decently and is better at detecting story errors.

The contributions of this paper are threefold:

- We re-collect and organize human evaluation results from recent VIST papers to form a new dataset: VHED.
- We propose a novel valid metric Vrank for visual storytelling which appropriately evaluates VIST model performance.
- We propose three assessments for metrics according to human properties and a generalization test to better illustrate the shortcomings of existing VIST metrics.

2 Related Work

Visual Storytelling (VIST) Visual storytelling was introduced by Huang et al. (2016) as the task of generating a coherent story given five images. They provided a dataset, Sequential Images Narrative Dataset (SIND), containing images and references in which references are human-written short stories describing images. For every image prompt (one sequence of photos), there are 2 to 5 references. VIST requires deeper understanding of the photo events to prevent descriptions in isolation (i.e., image captions). Researchers have proposed various methods for this task. Knowledge graphs are often integrated in models to encourage diversity of terms and plots in the stories (Hsu et al., 2020, 2021a; Chen et al., 2021). Some studies use reinforcement learning to reward models that generate stories that contain fewer errors and are more topically-focused (Huang et al., 2019; Hu et al., 2020a). However, existing evaluation methods are unable to capture the true quality of the generated stories. Thus we examine automatic metrics to devise a better way for machines to evaluate stories.

VIST-Human Evaluation Several VIST generation models use human evaluation to evaluate model performance. Recent studies apply aspect-based rating evaluation. Hu et al. (2020b) and Wang et al. (2020b) ask workers to rate stories based on pre-defined aspects.[1] However, it is difficult to normalize these aspects as the definition of aspect varies from paper to paper. Also, these aspects are not mutually independent, making it difficult to analyze results based on these ratings. Therefore, we consider the ranking method as it is commonly used (Hsu et al., 2020; Wang et al., 2020b; Hsu et al., 2021a) among authors. Hsu et al. (2020) asks human annotators to rank five stories from different models based on overall quality. Hu et al. (2020b) and Wang et al. (2020b) conduct pairwise human evaluations to rank stories according to different story aspects, where the latter is judged to be closer to human-level. These human evaluation results are valuable resources for observing human judgments in visual storytelling. Hence, in our work we collect this information for analysis and model training.

Automatic Metrics Automatic evaluation metrics are widely used in language generation tasks. Most reference-based metrics (e.g., BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004)) evaluate the n-gram similarity between a generated text and the reference. However, referenced metrics correlate poorly with human judgment (Wang et al., 2018b; Hsu et al., 2019; Modi and Parde, 2019) in dialog generation and story generation tasks: the generated text is given unreasonable scores due to incongruity with the reference. To account for this, several reference-free metrics (Sinha et al., 2020; Guan and Huang, 2020) have been designed to measure generated texts without any reference. BERT-Score (Zhang et al., 2019), for instance, uses contextual embedding to calculate the similarity between candidates and references, and BLEURT (Sellam et al., 2020) uses referenced automatic metrics as supervision signals for pre-training and is fine-tuned on a human judgment evaluation dataset. UNION (Guan and Huang, 2020) uses pre-defined negative samples to train a model in an attempt to provide a metric that specializes in story generation. In our analysis, current metrics remain unable to mimic human judgment to discern quality differences in story pairs.

3 VHED

3.1 Dataset Description

The VHED dataset is a collection of human evaluation results from three VIST studies: KG-Story (Hsu et al., 2020), PR-VIST (Hsu et al., 2021a), and Stretch-VST (Hsu et al., 2021b). All papers followed Hsu et al. (2020)'s human evaluation method using Amazon Mechanical Turk. For each task, the workers were to rank the story by overall quality, from the best story to the worst story. Specifically, each task displayed N stories, and each worker ranked each story from 1 to N. Details about each paper are listed in Table 1.

The construction of VHED is shown in Fig. 2. Collected from the aforementioned papers, we obtained 4,500 task results. Further, we grouped N stories into story pairs, where the number of story pairs per task is C_2^N. The resulting story pairs (x_1, x_2) are either two machine-generated stories from two different models or one reference and one machine-generated story. For each story pair, there are five attributes:

- **Stories:** A story pair consists of a better-ranked story and worse-ranked story. The story pair is either a reference with a machine-generated story, or two machine-generated stories.
- **Image Sequence IDs:** A list of IDs for each of the five images from the SIND dataset (Huang et al., 2016).
- **Average Rank:** The average of the five workers' story rankings and is divided by N for normalization. N varies from paper to paper (Table 1)
- **Ranking Gap:** The ranking gap is calculated as the average ranking of x_1 minus the average ranking of x_2. The ranking gap distribution is shown in the appendix (Table 6).
- **Human Agreement:** Human agreement is when k workers agree that the better-ranked stories are better than the worse-ranked stories. Note that human agreement = 2 is equivalent to human agreement = 3, because 1 person agreeing that story A is better than B is equivalent to 4 people agreeing that story B is better than A. Therefore, we kept human agreements = 3,4,5 for simple notation.

For quality control, we remove story pairs with zero ranking gap. This yields 13,875 story pairs in

[1] Hu et al. define relevance, coherence, and expressiveness, and Wang et al. define focus, coherence, detail, share, grounded, and human.

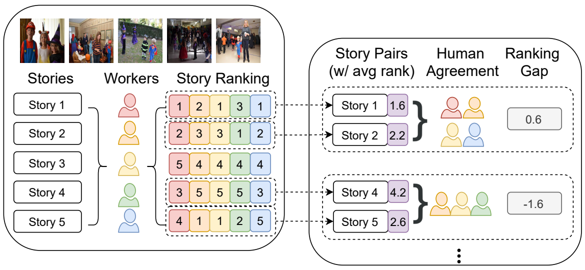

Figure 2: Workflow for creation of VHED dataset

Paper	**Human Evaluation**	**Sampling**	**Tasks**	**N**
KGStory	2	500	1,000	5
PRVIST	6	250–500	1,000	3–4
Stretch-VST	7	250–500	2,500	2–4

Table 1: Statistics of human evaluation results of KGStory (Hsu et al., 2020), PRVIST (Hsu et al., 2021a), and Stretch-VST (Hsu et al., 2021b)

total. [2] The train-test-validation sets were split at a ratio of 8:1:1 to 11,208, 1,351, and 1,316 story pairs. The descriptions of VIST models' generated stories are included in the appendix.

3.2 Data Analysis and Findings

As we acquired data about human preferences in story pairs, we conducted analyses to understand the potential patterns for workers when assigning story ranks, the quality gap between machine-generated and human-written stories, and the errors in the stories. The results of this observation are crucial for assessing the performance of a metric.

Worker Ranking Analysis Story pairs are grouped by the same human agreement. Ω^k denotes a sub-dataset containing story pairs with human agreement = k. In Table 2, we calculate the number of story pairs as well as the averaged ranking gap of each sub-dataset. For story pairs, we note that story pairs with $k = 3$ account for 53% of the dataset, meaning that half of the tasks have inconsistent annotations. Regardless, this paper evaluates the story pairs with $k \geq 4$ to filter out inconsistent human annotations. We also note that the ranking gap increases as human agreement increases. The ranking gap indicates the quality difference between a better-ranked and a worse-ranked story. That is, the difference between a ranked 1 story and a ranked 5 story should be larger than that between a ranked 2 story and a ranked 3 story. From Table 2, we find that story-pairs with lower agreement are closer in ranking. In other words, a story pair with a marginal quality difference easily leads to inconsistent worker annotations, because it is harder to rank two similar stories. Essentially, we expect the metrics to exhibit similar behavior: *the larger the ranking gap, the easier it is to rank.*

	Story pairs	Ranking gap	Machine better
Ω^3	6,494 (53%)	0.123	918(45%)
Ω^4	3,677 (30%)	0.247	523(35%)
Ω^5	2,110 (17%)	0.416	110(22%)

Table 2: The number and percentage of story pairs, average ranking gap of each sub-dataset. Machine better is the number and percentage of machine stories better than references in story pairs containing only a reference and a machine-generated story.

Who Wins? Machine vs. Human Stories Next we revisit the assertion that references are always superior. We select story pairs with a reference and a machine-generated story. We analyze the number and percentage of references that are ranked better than the generated stories on three human agreements. From Table 2, we observe that when more humans agree on the ranking results, the percentage of the reference being better also increases. In addition, further analysis shows that, on average, 38% of the machine-generated stories are in fact better than the references, showing that references are not always better than machine-generated stories.

Error Analysis To understand the difference between better- and worse-ranked stories, deeper analysis into the story content is necessary. We randomly sampled 200 stories from VHED (67 human and 134 machine generated) and manually labeled the stories according to the following error aspects:

- **Grammatical error (Gram):** Erroneous usage of past/current tense and mistakes in misplaced modifiers.
- **Repetition (Rep):** Repetitive sentences or phrases at sentence- and story-level.
- **Description in isolation (Desc):** Sentences that lack consistency, resulting in isolated captions instead of a fluent story.
- **Absurdity (Abs):** Ambiguous sentences or nonsensical phrases that are incomprehensible to humans.
- **Event mismatch (Event):** Stories that are off-topic, which present events that are not relevant to the image stream.
- **Object mismatch (Obj):** Irrelevant nouns that do not appear in the images and are not semantically related.

[2] VHED will be released upon acceptance

Type	Gram	Rep	Desc	Abs	Event	Obj
Percentage	.186	.141	.306	.351	.313	.186

Table 3: Error percentage of the sampled stories.

We first labeled stories based on all 11 error aspects defined in (Modi and Parde, 2019) and we select the most occurring errors, which are grammar, repetition, description in isolation, and absurdity. These four error aspects focus primarily on story coherence and within-story consistency. However, visual storytelling requires generated stories to fit the given story images. Rohrbach et al. (2019) show that humans are aware of the correctness of image descriptions. Also, Wang et al. (2020a) show that mismatched events in stories can lead to poor story quality. Therefore, we added event and object mismatch into our analysis. The error examples and correlation between the error are illustrated in the appendix (Table 9 and Figure 5).

From our observation, 79.8% of the sampled machine generated stories contained at least one of the errors in the categories, meaning most VIST models are unable to generate perfect stories. In Table 3, the high percentage of object and event mismatch errors also show that current VIST models do not capture visual groundings accurately. This can lead to humans assigning higher scores to human-written stories since they are most likely to be relevant to the given images. Grammatical errors and absurdities are also common in generated text, which can lead to ambiguous stories that humans are unable to comprehend. The prevalence of errors makes it essential for evaluation metrics to automatically detect these errors.

4 Vrank

We propose Vrank, a reference-free automatic metric that inputs story pairs to predict human preferences between the two stories. We utilize SIMCSE(Gao et al., 2021) to leverage better sentence representations. SIMCSE uses contrastive learning with dropout as augmentation, then trained on natural langauge inference datasets to obtain better sentence embeddings from BERT (Devlin et al., 2018). First, we pre-trained the SIMCSE model using SIND reference stories with the Masked Language Model objective. Then, we input two stories with a [SEP] token in between through the pre-trained model. We use the acquired sentence embeddings and feed it through a regression layer to predict a ranking gap. We used mean squared error to calculate the loss between the predicted ranking gap and true ranking gap. After obtaining the ranking gap, we predict which story is better according to the sign of the predicted ranking gap. Although Vrank is a simple model fine-tuned solely on human judgment, it still outperforms current existing metrics in our assessments. This suggests further potential for use with VHED; more studies can be conducted to replace Vrank with stronger neural network models.

During model training, since the number of positives and negatives were not balanced in the original dataset, we augmented the data to create a symmetric dataset of VHED to minimize dataset bias.[3] The ranking gap in the resulting dataset was close to normally distributed. We hypothesize that thus doing makes it possible to extract more information, making it easier for the model to learn human judgment for story pairs. However, due to the small amount of data available, high variance is likely (Mosbach et al., 2021) to occur during inference. Hence, we used all data from VHED, including human agreement=3 to increase the stability of our model following Mosbach et al. (2021).

5 Metric Assessment

In this section, we describe a series of assessments conducted on existing metrics on VHED, in which the assessment methods are based on the analyses in VHED. The objective is to examine whether Vrank is superior to other metrics based on our analysis of VHED.

5.1 Experimental Settings

Story-Pair Ranking A recent study (Wei and Jia, 2021) illustrates that pairwise accuracy reflects metric performance better than using correlation with human evaluation. Hence, we propose simple story-pair ranking to evaluate automatic evaluation metrics for visual storytelling. The task is to determine the correct ranking order of the stories in a story pair based on the story quality scores predicted by the automatic evaluation metrics being assessed. Given the story pair (x_1, x_2), the auto-metric being assessed predicts the corresponding story quality scores (s_1, s_2) which we compare to the averaged ranks y_1 and y_2 of x_1 and x_1 from human evaluation. The performance of the evaluation

[3] Other configurations, including utilizing visual features and changing the task objective to classifying better- and worse-ranked stories did not perform better.

Dataset	VHED			VIST-Edit		VHED	
Subset	Ω^4	Ω^5	$\Omega^{\{4,5\}}$	AREL-edit	GLAC-edit	R&M	M&M
Metrics	Reference-based metric					Reference-based metric	
Random	.516	.495	.511	.503	.481	.481	.528
BLEU-1	.470	.413	.459	.482	.405	.346	.529
BLEU-4	.205	.134	.192	.146	.103	.346	.097
SacreBLEU	.531	.557	.536	.424	.456	.528	.541
METEOR	.493	.432	.481	.437	.501	.461	.494
ROUGE-L	.506	.480	.501	.375	.389	.519	.491
BERT-Score	.527	.548	.531	.567	.450	.533	.529
	Reference-free/hybrid metric					Reference-free/hybrid metric	
BLEURT	.497	.451	.489	.546	.532	.509	.476
UNION-ROC	.488	.521	.496	.727	.475	.445	.525
UNION-WP	.449	.504	.461	**.740**	.612	.507	.435
Vrank	**.786**	**.826**	**.796**	.696	**.626**	**.816**	**.789**

Table 4: (Left) Average ranking accuracy for each metric on VIST-Eval and VIST-Edit. (Right) Evaluation results for reference-and-machine (R&M) story pairs and machine-and-machine (M&M) story pairs from $\Omega^{\{4,5\}}$. The Random baseline indicates that metrics that perform around 50% correspond to random guesses. Vrank's standard deviation for accuracy is calculated by training over 10 different seeds and taking the average.

metric on the i-th story pair is formulated as

$$\texttt{ranking\_acc}_i = \begin{cases} 1, & \text{if } s_1 > s_2 \text{ and } y_1 < y_2 \\ 1, & \text{if } s_1 < s_2 \text{ and } y_1 > y_2 \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $\texttt{ranking\_acc}_i = 1$ indicates correct (incorrect) prediction. Note that low scores indicate high rank. The overall metric performance is defined as

$$\texttt{avg\_ranking\_acc} = \frac{1}{M} \sum_{i=1}^{M} \texttt{ranking\_acc}_i, \tag{2}$$

where M denotes the number of story pairs for evaluation.

Datasets In addition to VHED, we also collected VIST-Edit[4] (Hsu et al., 2019) for story-pair ranking. VIST-Edit includes 2,981 visual stories generated by AREL (Wang et al., 2018a) and GLAC (Kim et al., 2018), and 14,905 human-edited visual stories, that is, AREL and GLAC-generated stories edited by workers. Their paper shows that the crowd workers' edits systematically increased the lexical diversity of the stories. Since the purpose of the editing was to improve the machine-generated stories, we paired up human-edited stories and machine-generated stories as better-ranked and worse-ranked samples (labeled as 1 and 2), resulting in 14,905 story pairs. Comparing VHED to VIST-Edit, VHED contains reference and multiple models' generated stories, but VIST-Edit has only human-machine story pairs. Additionally, VIST-Edit is not in Vrank's training data. VIST-Edit is utilized only for metric performance reports, serving as an unseen dataset for Vrank.

Baseline Automatic Metrics We implemented BLEU, ROUGE-L, METEOR, and SacreBLEU (Keenan, 2017), all traditional n-gram-based reference-based metrics. We also considered the more recent BERT-Score, BLEURT, and UNION as baseline metrics. In addition to the above automatic metrics, we also included a random baseline to provide a random score for each story, shown as Random in Table 4, as the lower bound.

A common practice for reference-based metrics: a candidate story is scored against each reference r_j in a gold reference set $R = \{r_i\}_{i=1}^{n}$; the highest score was used. However, applying this method on a reference-machine story pair would always result in reference having a full score, because of the exact match between reference and the gold reference set. To ensure a fair evaluation and avoid meaningless matching, we first check that the gold references do not include the reference. To this end, we propose the Reference Absent Algorithm for evaluating story pairs containing the reference story (or stories) as in Eq. 3, which removes the r_j from R when any of the candidate stories in a story pair ($x = \{x_1, x_2\}$) is identical to r_j.

$$s_j = \max(\text{metric}(x_j, R - x)), j = \{1, 2\}, \tag{3}$$

where $\text{metric}(\cdot)$ can be any reference-based metric and s_j is the story quality scores for the j-th story in a story pair. The Reference Absent Algorithm only applies when evaluating story pairs containing references, i.e., reference-machine pairs in this paper.

[4] VIST-Edit: https://github.com/tingyaohsu/VIST-Edit

5.2 Results and Discussion

Pairwise Story Evaluation Accuracy: Metric's ability to determine the correct ranking order in story pairs. The average ranking accuracy of each automatic metric on VHED and VIST-Edit are presented in Table 4 (left). Around 50% corresponds to random guessing, as shown as Random in the table. Vrank shows superior performance in VHED and VIST-Edit, which VIST-Edit is the unknown dataset to Vrank. High performance on VIST-Edit and VHED indicates Vrank has the ability to distinguish diverse story pairs. In contrast, we observe unexpectedly low performance for most baseline metrics, as they perform no better than the Random baseline. BLEU-4 especially struggles to rank the stories in both datasets. Further analysis suggests that BLEU-4 marked ∼80% of the stories as 0, and Equation 1 coincidentally treated them as incorrect prediction because it discourages ties. BLEURT, in turn, also performed poorly because it relies on reference-based metrics as signals for training. Reference-free metrics, especially UNION, perform well on VIST-Edit. However, its design is not generalizeable to VHED.

Worker Ranking Behavior on Metrics: The larger the ranking gap, the easier is it to rank. The ranking gap is the difference between a better-ranked and worse-ranked sample's average ranks. VHED is categorized into four sub-datasets with different ranking gaps. This assessment tests each metrics' ability to mimic worker ranking behavior observed in the analysis. Story pairs with larger gaps suggest stronger linguistic differences and are likely easier to rank, whereas those with smaller gaps are likely more difficult. In Fig. 3, all baseline automatic metrics, including metrics not reported in the figure, show randomly distributed scores, most of which remain around 50%, thus failing to exhibit such behavior. On the contrary, Vrank yields an ideal decrease. Starting with ranking gaps over 0.3, the accuracy reaches ∼0.85 and a gradual decrease afterward. We believe such behavior reveals Vrank to be a more preferable metric for visual story evaluation.

Machine and Human on Metrics: Machines are sometimes better than humans. Two aspects are studied in this section. First, we evaluate the ability of Vrank and reference-based metrics to rank reference-machine (R&M) pairs. Although some machine texts have progressed to human-level, to our knowledge, there has been little investigation of metrics' ability to evaluate references and machines. We apply reference-based metrics with Eq. 3. This results in poor performance for reference-based metrics as shown in R&M in Table 4[5]. An explanation is that since the reference is removed from the reference set by Eq. 3, the reference needs to match with the remaining references in the reference set. Although most references are on topic, the stories are highly diverse (Zhu et al., 2020). These metrics are unable to calculate the similarity to semantic levels; thus, they result in poor performance. On the contrary, Vrank is a deep learning model, trained on VHED and thus learned to rate based on story quality rather than similarity. Another analysis to study ability of Vrank to rank correctly when machine is better than reference shows that Vrank yields 26.5% recall when the other metrics have 0 recall without Eq. 3 and ∼18% with Eq. 3.

Second, we observe the performance of metrics on M&M (machine-machine pairs). M&M ranking gaps are smaller than those of R&M pairs (0.18 v.s. 0.21), making them harder to rank because their story qualities are closer. However, Vrank still shows promising performance when ranking such story pairs, outperforming existing metrics.

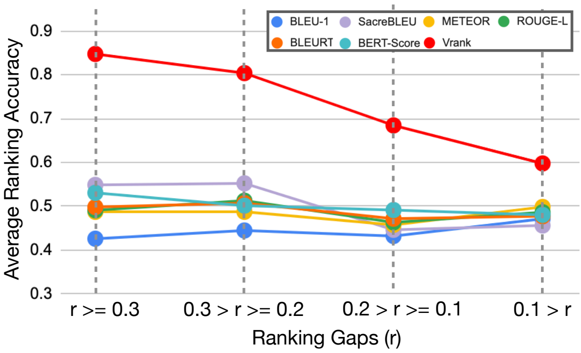

Figure 3: Average ranking accuracy for each metric on four sub-datasets with different ranking gaps r. Based on our analyses, metrics should perform better when the ranking gap is larger, and gradually decrease when the gap is smaller.

Errors in Metrics: Metric's ability to detect errors. Current generated stories often contain errors which prompt human evaluators to assign lower scores. It is crucial for automatic metrics to also recognize such errors to judge generated text. To do this, we adapted the point-biserial correla-

[5] A complete table without Eq. 3 can be found in the appendix (Table 7)

Error Types	Human	Vrank	UNION-ROC	UNION-WP	BLEURT	BERT-Score	ROUGE-L	METEOR
Gram	-0.107	-0.021	-0.099	-0.087	**-0.228**	-0.124	0.024	-0.167
Desc	-0.212	**-0.154**	-0.149	0.154	-0.081	0.080	0.114	-0.018
Rep	-0.130	-0.042	-0.120	**-0.411**	0.168	0.134	0.079	-0.034
Abs	-0.309	**-0.308**	0.003	0.120	-0.113	0.105	0.092	-0.025
Obj	-0.067	-0.157	-0.089	0.158	**-0.302**	-0.111	-0.048	-0.098
Event	-0.191	-0.093	0.008	-0.001	**-0.131**	0.043	0.138	-0.099

Table 5: This table shows the correlation of human rankings, automatic metric scores with the corresponding error categories. An ideal correlation should be closer to Human. Negative correlation illustrates that higher rankings (average ranking closer to 1) co-occur with few errors in the story. Hence, an high error detection rate is a correlation coefficient closer to -1.

tion coefficient to analyze the correlation between binary annotated errors and metric scores.

The correlation between metrics and errors is presented in Table 5: existing metrics are not able to detect errors as the correlation coefficients are low. From the correlation coefficients between the human ranking score and each error aspect, we observe that human evaluation for stories may be influenced by error aspects, especially absurdity and description in isolation. In general, Vrank performs best in detecting absurdity and description in isolation. UNION-WP performs best in correlation with repetition, which is reasonable since UNION is trained to discriminate erroneous stories that are repetitive in structure. In summary, current metrics remain unable to detect errors to evaluate coherency efficiently. Metrics ability to detect errors may give clearer indications of the quality of generated texts.

6 Dataset Generalization

In addition to VIST, we expect Vrank to reasonably evaluate the quality of text as well. To determine whether Vrank generalizes to textual stories, we selected as the benchmark the MANS dataset (Guan et al., 2021), an image-free storytelling dataset in which the stories are derived from the ROCStories corpus. This dataset includes 200 story prompts, where each prompt includes five model-generated stories and a reference. However, it does not contain human story rankings. Thus, for each story prompt, we asked five workers from Amazon Mechanical Turk to rank the five stories to obtain ranking scores.

Following the VHED construction procedure, the ranked stories were converted into story pairs, making for 1,112 story pairs for which 3 workers agreed on the ranking, 605 story pairs for which 4 workers agreed, and 132 story pairs for which 5 workers agreed. Likewise, we evaluate story pairs with $k \geq 4$.

Reference-based metric			
Subset	Ω^4	Ω^5	$\Omega^{\{4,5\}}$
BLEU-1	.486	.530	.494
BLEU-4	.007	.030	.001
SacreBLEU	.537	.545	.539
METEOR	.489	.576	.505
ROUGE-L	.506	.508	.506
BERT-Score	.509	.530	.513
Reference-free/Hybrid metric			
BLEURT	.531	.538	.532
UNION-ROC	.493	.553	.503
UNION-WP	.444	.500	.455
Vrank	**.575**	**.644**	**.588**

Table 6: Average ranking accuracy for generalizing to MANS.

The results of Vrank and the baseline automatic metrics when ranking MANS are shown in Table 6. We find that Vrank outperforms baseline metrics in story pairs with $k \geq 4$, whereas the latter still show limited abilities to rank the MANS dataset. In general, the accuracy of automatic evaluation on MANS is lower than that on VHED. This may be due to the comparably unconstrained writing styles of pure textual stories. An example of the evaluation on stories is given in the appendix(Table 8).

7 Conclusion and Discussion

We present VHED and Vrank, the first dataset of human evaluation results and evaluation metric for VIST. We show that Vrank performs significantly better in three assessment tasks and generalizes to other datasets. Also, recent automatic metrics are ill-suited to evaluating visual stories, especially human-level written stories. We welcome researchers to share their human evaluation results to the community to broaden the data domain to obtain more knowledge about human judgment and improve the performance of Vrank. As the gap between machines and humans continues to decrease, stronger metrics will be needed to evaluate machine and human stories. Improving Vrank performance to replace reference-based metrics is our future goal.

References

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

Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. 2021. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. *CoRR*, abs/2102.02963.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional Transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jian Guan and Minlie Huang. 2020. UNION: An unreferenced metric for evaluating open-ended story generation. *arXiv preprint arXiv:2009.07602*.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. Openmeva: A benchmark for evaluating open-ended story generation metrics.

Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao (Kenneth) Huang, and Lun-Wei Ku. 2020. Knowledge-enriched visual storytelling. In *Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Chi-Yang Hsu, Yun-Wei Chu, T. Huang, and Lun-Wei Ku. 2021a. Plot and rework: Modeling storylines for visual storytelling. In *FINDINGS*.

Chi-Yang Hsu, Yun-Wei Chu, Tsai-Lun Yang, Ting-Hao Huang, and Lun-Wei Ku. 2021b. Stretch-vst: Getting flexible with visual stories.

Ting-Yao Hsu, Chieh-Yang Huang, Yen-Chia Hsu, and Ting-Hao'Kenneth' Huang. 2019. Visual story post-editing. *arXiv preprint arXiv:1906.01764*.

Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. 2020a. What makes a good story? designing composite rewards for visual storytelling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7969–7976.

Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. 2020b. What makes a good story? Designing composite rewards for visual storytelling. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, New York, USA.

Qiuyuan Huang, Zhe Gan, Asli Celikyilmaz, Dapeng Wu, Jianfeng Wang, and Xiaodong He. 2019. Hierarchically structured reinforcement learning for topically coherent visual story generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8465–8472.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

William JF Keenan. 2017. Sacre Bleu: Faith, fashion and freedom: Marist foundation garments 1817–1862. In *Materializing Religion*, pages 132–153. Routledge.

Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. 2018. GLAC Net: GLocal Attention Cascading Networks for multi-image cued story generation. *arXiv preprint arXiv:1805.10973*.

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

Yatri Modi and Natalie Parde. 2019. 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*, pages 47–57, Minneapolis, Minnesota. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning robust metrics for text generation.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jing Wang, Jianlong Fu, Jinhui Tang, Zechao Li, and Tao Mei. 2018a. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ruize Wang, Zhongyu Wei, Ying Cheng, Piji Li, Haijun Shan, Ji Zhang, Qi Zhang, and Xuanjing Huang. 2020a. Keep it consistent: Topic-aware storytelling from an image stream via iterative multi-agent communication.

Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. 2020b. Storytelling from an image stream using scene graphs. In *AAAI 2020*.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018b. No metrics are perfect: Adversarial reward learning for visual storytelling. *arXiv preprint arXiv:1804.09160*.

Johnny Tian-Zheng Wei and Robin Jia. 2021. The statistical advantage of automatic nlg metrics at the system level.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Wanrong Zhu, Xin Eric Wang, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. 2020. Towards understanding sample variance in visually grounded language generation: Evaluations and observations. *arXiv preprint arXiv:2010.03644*.

8 Appendix

Application In this section, we introduce an application for Vrank and other reference-free metrics. Our assessment indicates that Vrank's predictions strongly agree with human judgment. We quantify the distance between humans and machines by pairing up reference and generated stories and calculating the ratio of generated stories that outmatch the references. Unlike human evaluation, which can be conducted only on a portion of the testing data, this method allows researchers to evaluate the proposed model over the entire testing dataset.

After applying Vrank to assess five recent VIST models, we present the results in Fig. 4: the models are gradually approaching human-level writing, outlining an exciting development of NLG in VIST.

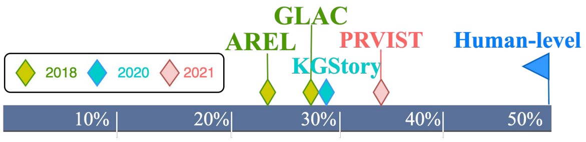

Figure 4: Ratio of generated stories that outmatch the references. Colors denote the publication years. A result of 50% indicates half of them outmatch the references.

Error Type Examples and Correlation In Table 9, we show examples of error types mentioned in our error analysis. We also show the correlation between different error types in Fig. 5. As the error types are mutually independent, there is the potential to construct tools to automatically detect each error, since they do not overlap with each other.

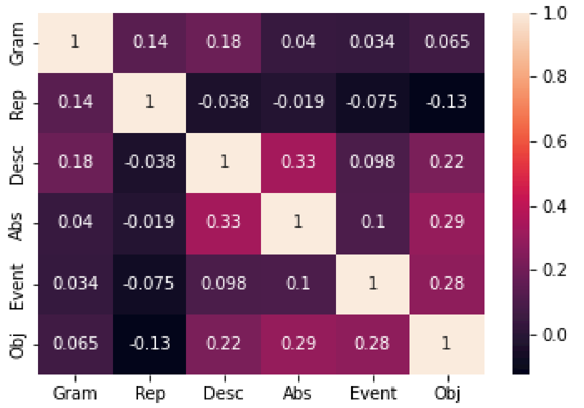

Figure 5: Correlation matrix between different kinds of errors, including **Gram**matical errors, **Rep**etitions, **Desc**riptions in isolation, **Abs**urdity, **Event** mismatches, and **Obj**ect detection errors

Reference-based metric					
Subset	Ω^4	Ω^5	$\Omega^{\{4,5\}}$	R&M	M&M
BLEU-1	.597	.647	.607	.657	.571
BLEU-4	.569	.689	.593	.657	.547
SacreBLEU	.533	.647	.556	.657	.482
METEOR	.546	.638	.564	.657	.497
ROUGE-L	.541	.647	.563	.657	.494
BERT-Score	.516	.663	.546	.657	.464
Hybrid metric					
BLEURT	.552	.664	.575	.657	.514

Table 7: Ranking accuracy for each metric on VIST-Eval. Reference-based metrics without Reference Absent Algorithm accuracy results. Reference-free metrics are not affected by this algorithm.

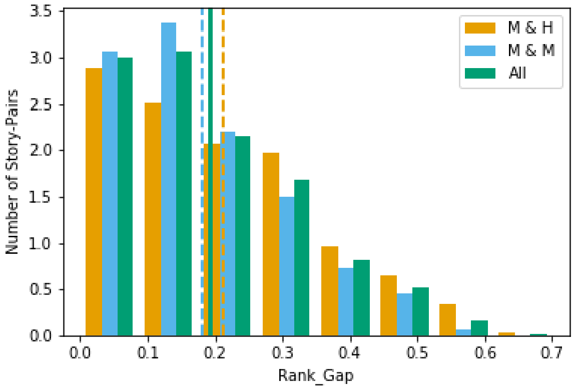

Figure 6: Normalized ranking gap distribution of machine-machine pairs, human-machine pairs, and all story pairs. The dashed lines are the ranking gap means for H&M and M&M pairs, and the full line is the mean for all story-pair ranking gaps.

Ranking Gap Distribution The ranking gap distribution is shown in Fig. 6, in which both the ranking gaps and the number of stories are normalized. Also, since the ranking gaps contain both negative and positive values, we took the absolute value of the gap for the histogram. We observe that the machine-machine pairs are centered closer to zero. However, the human-machine pairs are distributed more evenly than the M&M pairs, which indicates that human-machine pairs are easier to distinguish than machine-machine pairs.

Without Reference Absent Algorithm Here, we show the results of automatic metric accuracy in story-pair ranking without the proposed Reference Absent Algorithm. As expected, the accuracies for H&M pairs are the same. Since all references are regarded as ground truth for reference-based automatic metrics, the accuracy is shown as the percentage of the human-written stories that are better than machine-generated stories. Hence, these metrics are unable to identify any machine-generated stories that are better than human-written stories. This demonstrates the importance of our proposed algorithm in the experiment results.

Data Collection Details We sampled 250 to 500 image prompts from SIND's testing dataset and hired crowd workers from Amazon Mechanical Turk to evaluate the visual stories that were generated based of these image prompts. The workers were adult workers in the US with 98% assignments approved and who had completed at least 3,000 HITs. A user interface for workers to complete was called a task. A task displayed one image prompt on the top with several stories at the bottom, and five workers were recruited to rank the stories. The stories usually included a reference, stories generated using the proposed model, and several baseline stories. The compensation was USD 0.10 per task.

Training Details We use the pre-trained base model from Huggingface (Wolf et al., 2020) and fine-tune it to our regression objective. We utilized Adam as optimizer with learning rate 2e-5 and trained for 30 epochs. The batch size is set as 32 and the random seed for training can be set as 7777 for reproduction. Checkpoints are stored for every 500 steps and we also utilized mixed precision training for more efficient training. The environment of our operating system is Ubuntu 20.0.4. Training was completed on two NVidia RTX 3090 GPUs, each of which contains 24 GB of memory.

Model Design Before we came up with the final model using SIMCSE, we tried several settings. Formulating the task as a binary classification task didn't achieve good accuracy, we speculate that this is because the boundaries for a good and bad story is hard to find. Also, we tried to augment the story-pairs with agreement=5. We found out that it didn't improve the performance. Moreover, we tested using CLIP(Radford et al., 2021) to extract image features for additional features and vision-language models also did not improve performance. Hence, we picked a simple model architecture to demonstrate our performance.

Details of Story Generation Models in VHED

- GLAC (Kim et al., 2018): combines global and local attention to construct image-dependent sentences. A context cascading mechanism is incorporated to improve story coherency.
- AREL (Wang et al., 2018a): uses a policy model and reward model to associate reward learning. The policy model is used to generate stories, and the reward model learns from human demonstrations.
- KGStory (Hsu et al., 2020): a three-stage framework which distills a set of representive words from the input text and utilizes knowledge graphs to enrich the content. It generates stories from the enriched word set.
- PRVIST (Hsu et al., 2021a): a two-stage framework that finds an optimal path through the constructed story graph which forms the best storyline. This path is then used to generate the story.
- Stretch-VST (Hsu et al., 2021b): a modification of KGStory that produces more sentences in the story while maintaining quality. Appropriate knowledge added to the story results in a more detailed story.

	Story	Vrank	UNION	BLEURT	Human
Story1	i learned of my baby 's birthday . i was very sad because my parents made her cake . i went to get my cake . [FEMALE] family surprised me and made me a very happy face .	Rank 2	Rank 1	Rank 1	Rank 2
Story2	one night , my parents and i decided to go to the movies . afterwards , we decided to sleep together . i fell asleep while my dad was watching movies . i was never able to sleep with my parents since my parents were away .	Rank 1	Rank 2	Rank 2	Rank 1
Reference	i told my mother bye as i went to school . after school later that day my brother picked me up . he told me and my twin brother our mother had died . i went home and cried my eyes out .	NaN	NaN	NaN	NaN

Table 8: Example of stories in MANS datasets, and the each metrics' rankings for stories.

Error types	Examples
Grammatical error	there was a lot of students . the space was very small . this is our hotel area . we got to town on our trip . everyone had a great game .
Repetitions	i went to the city yesterday . the streets were empty and the streets were empty . the city was very tall . the city was very tall . it was a beautiful day .
Description in isolation	i went on a hike . i met some people there . they were playing around the house . we got very scared . it was a sheep .
Absurdity	the city was beautiful . there was a lot of traffic . it was a nice day . and the streets were empty . but i had a great time .
Object mismatch	our trip to the town were amazing . it was a long trip to many different formations . my dad took pictures of the view . it was a great view of the snow . i also saw water in the stone .(See Figure 7)
Event mismatch	the parade was a lot of fun . there were many people there . they were all very excited . it was a great time . everyone was dressed up .(See Figure 8)

Table 9: Error types with examples

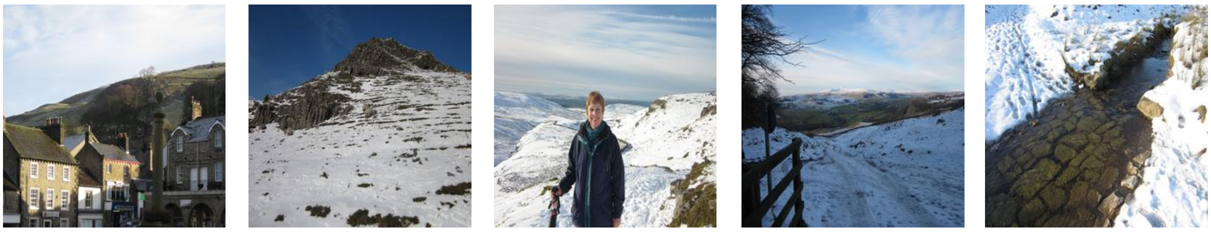

Figure 7: Illustration for object detection error. The last sentence:"i also saw water in the stone" is incorrect. Since there isn't water seen in the photo, it should be snow instead. StoryID:47608.

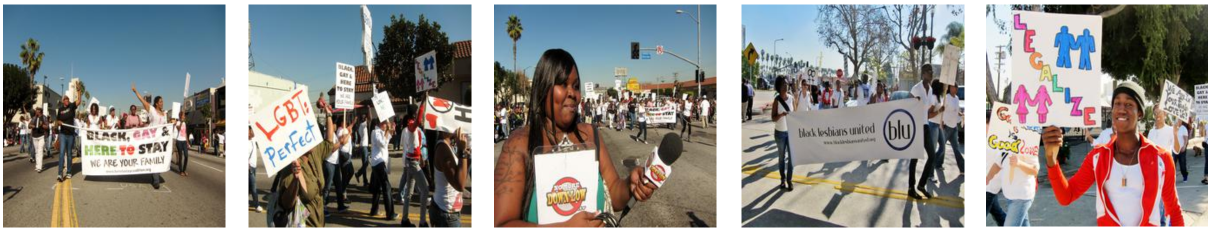

Figure 8: Illustration for event mismatch error. The event should be a peaceful protest for civil rights, while the example story regard the event as a festival parade. StoryID:47670.