Context-aware Visual Storytelling with Visual Prefix Tuning and Contrastive Learning

Yingjin Song, Denis Paperno and Albert Gatt

Utrecht University, Utrecht, The Netherlands
{y.song5, d.paperno, a.gatt}@uu.nl

Abstract

Visual storytelling systems generate multi-sentence stories from image sequences. In this task, capturing contextual information and bridging visual variation bring additional challenges. We propose a simple yet effective framework that leverages the generalization capabilities of pretrained foundation models, only training a lightweight vision-language mapping network to connect modalities, while incorporating context to enhance coherence. We introduce a multimodal contrastive objective that also improves visual relevance and story informativeness. Extensive experimental results, across both automatic metrics and human evaluations, demonstrate that the stories generated by our framework are diverse, coherent, informative, and interesting.

1 Introduction

Visual storytelling (VIST; Huang et al., 2016) aims at crafting a narrative from a sequence of ordered images. This task involves a number of key challenges, some of which are well-studied problems in computational narrative generation, while others arise from the visually grounded nature of the task: VIST image sequences exhibit semantic and temporal gaps, so that (i) a successful VIST system needs to balance textual **coherence** (Redeker, 2000; Callaway and Lester, 2001) with (ii) visual **grounding** (Wang et al., 2022; Surikuchi et al., 2023). At the same time, (iii) generated narratives should capture the reader's attention, necessitating a degree of creativity and **interestingness** (Gervás, 2009), but should also (iv) be **informative** (Li et al., 2019a; Chen et al., 2021), that is, incorporate relevant details of the entities and activities in the visual content.

Existing models usually include a vision encoder and language decoder either trained from scratch or finetuned (Kim et al., 2018; Wang et al., 2018b; Hu et al., 2020; Li et al., 2022; Fan et al., 2022; Yang and Jin, 2023; Wang et al., 2024) on the VIST task. This requires a large amount of computational resources. Instead, we propose to benefit from pre-trained models that have already learned meaningful representations from vast amounts of data, following the ClipCap approach (Mokady et al., 2021) that integrates pretrained CLIP (Radford et al., 2021) and GPT2 (Radford et al., 2019) via a lightweight mapping network. ClipCap trains only the mapping network to construct soft visual prefixes from CLIP embeddings to guide GPT2 to generate text, while both CLIP and GPT2 can be kept frozen. Although visual prefix tuning has been widely used for image captioning, it has not been adapted for visual storytelling, and its potential here is yet to be explored.

Our new framework incorporates a context-aware mappping network, while addressing coherence by incorporating previous story sentences. To enhance visual grounding and informativeness, we employ a multimodal training objective. We further compare four common decoding strategies (beam, top-k, nucleus and contrastive search), showing that they have substantial impact on the generation quality, especially as reflected in human evaluation, in contrast to standard metrics.

The main contributions of this work are:[1]

- a framework to incorporate textual coherence in VIST, while leveraging pretrained models;
- contrastive training to improve informativeness and visual grounding;
- a comprehensive human evaluation targeting the four challenges outlined above;
- extensive evaluation demonstrating competitiveness with state-of-the-art baselines.

[1] Our code and model are available at https://github.com/yjsong22/ContextualVIST

Proceedings of the 17th International Natural Language Generation Conference, pages 384–401
September 23–27, 2024. ©2024 Association for Computational Linguistics

2 Related Work

Visual Storytelling. The Visual Storytelling (VIST) task (Huang et al., 2016) aims to create narrative continuity between images for a fluent, coherent story. Early attempts extended image captioning models by combining global-local visual attention (Kim et al., 2018) and learning contextualized image representations (Gonzalez-Rico and Fuentes-Pineda, 2018). Considerable efforts explored Reinforcement Learning (RL) with custom reward functions for visual storytelling (Wang et al., 2018a,b; Huang et al., 2019; Hu et al., 2020). Given that storytelling involves imagination and reasoning, many works (Yang et al., 2019; Hsu et al., 2020; Wang et al., 2020; Chen et al., 2021; Xu et al., 2021; Zheng et al., 2021; Li et al., 2022; Wang et al., 2024) also integrate external knowledge to introduce commonsense concepts not directly present in visual input.

Recent research leverages Transformer-based architectures to learn multimodal feature embeddings, integrating image regions with semantic relationships (Qi et al., 2021). Several studies have focused on utilizing pre-trained models for visual storytelling, either by fine-tuning pre-trained Transformer encoders (Fan et al., 2022), or jointly tuning pre-trained LMs with pre-trained image encoders (Yu et al., 2021). Other variants consider additional factors such as emotion/sentiment (Li et al., 2019b), personas (Chandu et al., 2019; Liu and Keller, 2023; Hong et al., 2023), and writing style (Wang et al., 2023; Yang and Jin, 2023). Unlike prior work, our approach efficiently adapts frozen VLMs and LLMs, conditioning on both textual context and visual input to ensure story continuity and coherence.

Prompt and Prefix Tuning. Prompting means designing "instructions" for pretrained language models (LM) to generate desired outputs, conditioning them on either human-crafted templates or automatically optimized tokens (Liu et al., 2023b). Much research proposes to automate prompt engineering by learning discrete (Jiang et al., 2020; Haviv et al., 2021; Ben-David et al., 2022) or continuous prompts (Li and Liang, 2021; Lester et al., 2021). The latter can be updated via backpropagation, making them less constrained than (Zhong et al., 2021; Petrov et al., 2024). With large frozen LMs, Prompt Tuning (Lester et al., 2021) simply adds a tunable, real-valued embedding to the input of the decoder, achieving results comparable to full model fine-tuning. On the other hand, Prefix Tuning (Li and Liang, 2021) optimizes the inputs of every attention layer in the pretrained LMs.

Constructing soft visual prompts for a frozen LLM is an effective way to achieve vision-language alignment (Merullo et al., 2023; Koh et al., 2023). Flamingo (Alayrac et al., 2022) adds cross-attention layers to the LLM for incorporating visual features, pretrained on billions of image-text pairs. BLIP-2 (Li et al., 2023) adopts a Q-Former module to link a frozen image encoder to a frozen LLM, learning visual features relevant to text. LLaVA (Liu et al., 2023a), trained on multimodal instruction-following, uses a linear layer to map image features from pre-trained CLIP to the word embedding space of Vicuna (Chiang et al., 2023). Inspired by the widespread application of visual prefix tuning in V&L tasks, we explore its potential in visual storytelling while also considering the context when tuning the prefix.

3 Method

In visual storytelling, the input is a sequence of N images $\mathcal{I} = \{I_1, \ldots, I_N\}$, where $N = 5$ in the VIST dataset (Huang et al., 2016). Our model aims to generate a multi-sentence story $\mathcal{S}$ by predicting the probability $P(\mathcal{S}|\mathcal{I})$. In this section, we introduce a visual storytelling pipeline enhanced with prefix tuning (§3.1), then describe the context-aware components (§3.2), curriculum training (§3.3) and finally the contrastive learning loss involved (§3.4). Figure 1 illustrates an overview of our framework.

3.1 Visual Storytelling with Prefix Tuning

From the perspective of a single image, visual storytelling is very similar to image captioning, where an image-sentence pair $\{I_i, S_i\}$ is given. Motivated by prefix tuning (Li and Liang, 2021), ClipCap (Mokady et al., 2021) only updates the parameters of a lightweight Transformer-based mapping network during training to produce visual prefix vectors that can drive a pretrained frozen language model (LM) to generate text. ClipCap applies frozen CLIP (Radford et al., 2021) as vision encoder to extract visual features from the input image as $\boldsymbol{v}_i = f_{\text{CLIP}}(I_i)$. The visual feature $\boldsymbol{v}_i$ is then processed by a trainable mapping network $\mathcal{MN}_{\text{v}}$ to map the visual features to visual prefix vectors that are in the embedding space of the LM:

$$\mathbf{p}_{I_i} = [p_1, \ldots, p_k] = \mathcal{MN}_{\text{v}}(\boldsymbol{v}_i) = \mathcal{MN}_{\text{v}}(f_{\text{CLIP}}(I_i))$$

where k denotes the prefix size and $\mathcal{MN}_{\text{v}}$ is a Transformer with 8 multi-head self-attention lay-

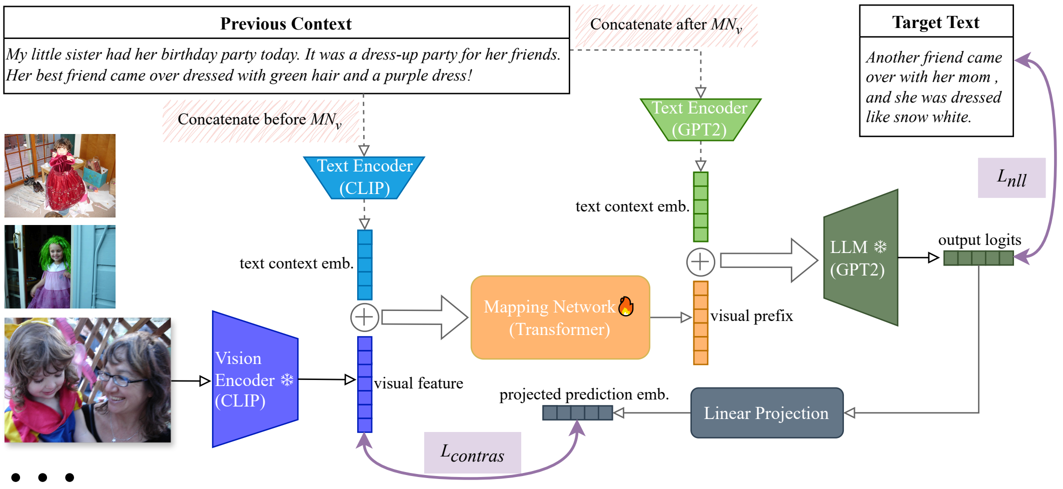

Previous Context
My little sister had her birthday party today. It was a dress-up party for her friends. Her best friend came over dressed with green hair and a purple dress!
Concatenate after MN_v
Target Text
Another friend came over with her mom , and she was dressed like snow white.
Concatenate before MN_v
Text Encoder (GPT2)
Text Encoder (CLIP)
text context emb.
text context emb.
L_{nll}
LLM ❄ (GPT2)
output logits
Mapping Network (Transformer)
visual prefix
Vision Encoder ❄ (CLIP)
visual feature
projected prediction emb.
Linear Projection
$L_{contras}$
• • •

Figure 1: Illustration of the framework. A Transformer-based mapping network ($\mathcal{MN}_\text{v}$) is trained to map visual features from a frozen encoder (CLIP) into a visual prefix for a frozen LLM (GPT2). We incorporate the previous sentences as the context via (1) concatenation after $\mathcal{MN}_\text{v}$: previous context is encoded by the LLM (GPT2), combined with the visual prefix and then fed into the LLM decoder; or (2) concatenation before $\mathcal{MN}_\text{v}$: previous context is encoded by the CLIP text encoder, combined with CLIP visual features and then fed into $\mathcal{MN}_\text{v}$. In addition to the teacher-forcing objective $\mathcal{L}_\text{NLL}$, we further compel the model to produce text that aligns semantically with the image through a contrastive training objective $\mathcal{L}_\text{contras}$.

ers with 8 heads each. We then concatenate the visual prefix vectors $\mathbf{p}_{I_i}$ to the caption tokens $S_i = [s_1, s_2, ..., s_\ell]$, as

$$\mathbf{z}_{I_i} = [p_1, \dots, p_k; s_1, \dots, s_\ell]$$

where ‘ ;’ denotes the concatenation. During training, $\mathbf{z}_i$ is fed into the LM with a teacher-forcing objective in an auto-regressive manner. In other words, the mapping network $\mathcal{MN}_\text{v}$ is trained using Negative Log-Likelihood (NLL) loss:

$$\mathcal{L}_\text{NLL} = -\sum_{j=1}^{\ell} \log p_\theta \left(s_j \mid p_1, \dots, p_k; s_1, \dots, s_{j-1}\right)$$

where θ are the trainable parameters of the model.

3.2 Context-aware Mapping Network

VIST story generation needs to establish informative connections between images in a sequence to bridge the potential visual/semantic gaps between them. We incorporate contextual knowledge into our model in the form of past story sentences. In addition to the image, we use the previous L sentences $[\mathcal{S}_{i-L}, \dots, \mathcal{S}_{i-1}]$ to generate the sentence for the current image I_i. For the first image I_0 in a sequence, we use the title and description of the belonging album[2] as the textual context. We propose two methods to include the previous sentences[3] as additional contextual information: (1) Concatenate $[\mathcal{S}_{i-L}, \dots, \mathcal{S}_{i-1}]$ with visual prefix vectors $\mathbf{p}_{I_i}$; (2) Concatenate $[\mathcal{S}_{i-L}, \dots, \mathcal{S}_{i-1}]$ with visual features $\boldsymbol{v}_i$ and use them together as the input of mapping network.

Concatenate after $\mathcal{MN}_\text{v}$. Following Han et al. (2023), we embed the sentences $[\mathcal{S}_{i-L}, \dots, \mathcal{S}_{i-1}]$ with the language generation model f_LM as

$$\mathbf{C}text_i = [\text{BOS}_\text{text}; f_\text{LM}([\mathcal{S}_{i-L}, \dots, \mathcal{S}_{i-1}]); \text{EOS}_\text{text}]$$

where BOS_text and EOS_text are learnable beginning and end of sequence tokens. The contextual vector $\mathbf{C}text_i$ is concatenated with the prefix vector $\mathbf{p}_{I_i}$ and then fed to the language generation model as a prompt vector (see Figure 1). $\mathcal{MN}_\text{v}$ is trained with NLL loss as:

$$\mathcal{L}_\text{NLL} = -\sum_{j=1}^{\ell} \log p_\theta \left(s_j \mid \mathbf{p}_{I_i}; \mathbf{C}text_i; s_1, \dots, s_{j-1}\right)$$

Concatenate before $\mathcal{MN}_\text{v}$. Since CLIP (Radford et al., 2021) is multimodal, we can use a common embedding space to encode both the image I_i as $f_\text{CLIP}\left(I_i\right)$, and previous sentences

[2] Huang et al. (2016) collected 10,117 Flickr albums that each contains 10 - 50 images. They asked human annotators to select 5 images of each album to form an image sequence, and write a story correspondingly. Album titles, descriptions and other metadata were provided in the original Flickr albums by the album owners.

[3] During training, we use the previous ground-truth sentences as the context, while during inference the past predicted sentences are used instead.

$[\mathcal{S}_{i-L}, \ldots, \mathcal{S}_{i-1}]$ as $f_{\text{CLIP}}([\mathcal{S}_{i-L}, \ldots, \mathcal{S}_{i-1}])$. The two CLIP embeddings are then concatenated and fed into the mapping network to produce visual prefix vectors

$$\mathbf{p}'_{I_i} = \mathcal{MN}_{\text{v}}([f_{\text{CLIP}}\left(I_i\right); f_{\text{CLIP}}([\mathcal{S}_{i-L}, \ldots, \mathcal{S}_{i-1}])]).$$

The $\mathcal{MN}_{\text{v}}$ is trained with NLL loss as:

$$\mathcal{L}_{\text{NLL}} = -\sum_{j=1}^{\ell} \log p_\theta \left(s_j \mid \mathbf{p}'_{I_i}; s_1, \ldots, s_{j-1}\right)$$

3.3 Curriculum Learning

In VIST, reference texts are often too generic and lack concretness to the image content. An example is "There was a lot to see and do" for an image depicting a funfair. The frequency of this phenomenon may compromise the model's ability to ground its linguistic choices in visual data. To address this, we use curriculum learning, which involves training a model with data sorted by difficulty to improve generalization and speed up convergence (Bengio et al., 2009).

We start by training the model on basic image captioning data to enhance grounding abilities before progressing to storytelling from image sequences. The training proceeds as follows: **(1)** Train the mapping network $\mathcal{MN}_{\text{v}}$ with image-caption pairs (Description in Isolation, DII) from VIST (see Section 4.1). **(2)** Switch to visual storytelling data (Stories in Sequence, SIS) once validation loss stops decreasing. **(3)** Return to step **(1)** when validation loss stops decreasing. **(4)** Stop training when no further improvement in validation loss is observed.

3.4 Visually-supervised Contrastive Training

To encourage our model to generate text that is grounded in the image, we leverage a contrastive training objective $\mathcal{L}_{\text{contras}}$ in addition to the teacher forcing objective $\mathcal{L}_{\text{NLL}}$. To maximize the relatedness between a positive pair consisting of a target text sequence and a source image, while minimizing the similarity between the negative pairs, we apply InfoNCE (Noise-Contrastive Estimation) loss (Oord et al., 2018) as:

$$\mathcal{L}_{\text{contras}} = -\log \frac{\exp\left(\text{sim}\left(\boldsymbol{v}_i, \hat{S}_i\right)/\tau\right)}{\sum_{j \neq i}^{|B|} \exp\left(\text{sim}\left(\boldsymbol{v}_i, \hat{S}_j\right)/\tau\right)}$$

where $\hat{S}_i$ is the projected representation of the text decoder's final layer output via a linear projection layer, $\text{sim}(,)$ denotes the cosine similarity of the two vectors, $|B|$ is the batch size, and τ denotes the temperature.

		Original	Ours
Train	No. DII captions	120,465	120,099
	No. SIS stories[5]	40,098	40,071
Val	No. DII captions	14,970	14,940
	No. SIS stories	4,988	4,988
Test	No. DII captions	15,165	15,165
	No. SIS stories	5,050	5,030

Table 1: Data split in original VIST dataset annotations and our experiments. Differences are due to the removal of unavailable images for some samples. DII: Descriptions of Images in Isolation. SIS: Stories of Images in Sequence.

During training, we first train the mapping network with the NLL loss $\mathcal{L}_{\text{NLL}}$ (training DII and SIS data in curriculum training scheme) for the first N_{nll} epochs and then add the contrastive loss $\mathcal{L}_{\text{contras}}$ (using only SIS data). The reason for not using $\mathcal{L}_{\text{contras}}$ from the beginning is that initially the model can only generate random tokens, which cannot be projected to semantically meaningful embeddings for contrasting with the image representation. Overall, our model is trained by minimizing the combined loss $\mathcal{L}$ (Zhu et al., 2023) as:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\text{NLL}}, \textit{epoch} < N_{nll} \\ \mathcal{L}_{\text{NLL}} + \lambda \mathcal{L}_{\text{contras}}, \textit{epoch} \geq N_{nll} \end{cases}$$

where λ is the coefficient of the contrastive loss.

4 Experiments[4]

4.1 Dataset

The visual storytelling (VIST; Huang et al., 2016) dataset includes 210,819 unique photos and 50,200 stories collected from 10,117 Flickr albums. Our experiments follow the data splits in the original VIST, removing the broken or unavailable image files (see Table 1).

4.2 Decoding Strategies

We compare four popular decoding methods for text generation: **Beam search** selects the text continuation with highest probability based on the model's probability distribution; this may result

[4] Experimental details of training, inference and automatic evaluation are listed in the Appendix A.

[5] Each story usually consists of 5 sequences of text corresponding to 5 images.

in low variation (Li et al., 2016) and degeneration (Fan et al., 2018; Holtzman et al., 2020) in the generated text. **Top-k sampling** redistributes the probability mass among only the top k most likely next tokens, avoiding sampling from the unreliable tail of the distribution (Fan et al., 2018). **Nucleus sampling** (Holtzman et al., 2020), also known as top-p sampling, chooses from the smallest set of tokens whose cumulative probability exceeds the probability p. **Contrastive search** (SimCTG, Su et al., 2022) jointly considers the probability predicted by the language model and the similarity with respect to the previous context.

4.3 Baseline Models

For a fair and thorough comparison, we choose four SOTA baselines that don't require additional datasets and have reproducible code/weights. **GLACNet** (Kim et al., 2018) is a seq2seq model using global-local attention and context cascading on visual features. **AREL** (Wang et al., 2018b) is an adversarial framework learning an implicit reward function from human demonstrations and optimizing policy search with a CNN-based reward model. **ReCo-RL** (Hu et al., 2020) is a reinforcement learning model with composite rewards for relevance, coherence, and expressiveness. **TAPM** (Yu et al., 2021) uses an adaptation loss to align a vision encoder with a pretrained LM and a sequential coherence loss to improve temporal coherence by aligning predicted text representations with neighboring visual representations.

4.4 Automatic Evaluation Metrics

In line with prior work on the VIST benchmark, we validate our results over the test set using the standard metrics BLEU (Papineni et al., 2002), ROUGE-L (Lin and Och, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016). We evaluate the generated text in terms of text-text semantic similarity using BLEURT (Sellam et al., 2020), image-text semantic similarity using CLIPScore (Hessel et al., 2021), and language fluency using Perplexity. Following Su et al. (2022), we also assess text degeneration and word diversity using: (1) rep-$n = 1.0 - \frac{|\text{ unique } n\text{-grams }|}{|\text{ total } n\text{-grams }|}$ measures story-level repetition by computing the portion of duplicate n-grams; (2) diversity$= \prod_{n=2}^{4}(1-$ rep-$n)$ measures the diversity of n-grams.

4.5 Human Evaluation

We conduct a human evaluation on a sample of generated texts. We randomly select 100 distinct image sequences and the corresponding generated stories from 8 models (i.e., our model[6] with four decoding strategies, the ground truth texts (GT), GLACNet, AREL and TAPM).

We invite 75 human annotators from Prolific to rate stories on a 5-point Likert scale for the criteria of **Visual Grounding**, **Coherence**, **Interestingness**, and **Informativeness**. As noted in Section 1, we consider these among the most important criteria for visually grounded narrative generation. Each participant answered 32 questions (each question containing ratings for one image sequence and one story across four criteria), resulting in a total of 9600 responses. We evenly distributed 800 pairs of image sequences and stories among all participants, ensuring that each question received ~3 responses. A full explanation of rating criteria, questionnaire instructions and sample questions are in the Appendix B.

5 Results and Analysis

Setting	**B-4**	**M**	**R-L**	**C**	**S**	**BR**	**PPL↓**
GLACNet	13.5	31.6	**30.0**	7.6	8.3	30.7	12.0
AREL	13.5	**31.7**	29.6	8.6	8.9	30.4	13.1
TAPM	11.4	30.7	28.7	9.5	10.0	31.4	18.3
ReCo-RL	13.1	31.5	27.9	11.5	**11.2**	27.7	28.4
no context							
beam	9.8	27.4	27.2	5.0	5.9	26.7	13.9
top-k	4.0	24.1	22.5	2.1	6.6	24.9	39.7
nucleus	3.5	23.6	21.4	1.7	5.7	24.1	42.5
SimCTG	7.3	28.5	25.5	5.7	6.9	25.8	16.6
+context after $\mathcal{MN}_{\text{v}}$							
beam	13.6	31.4	29.0	11.4	9.7	31.5	**10.5**
top-k	4.0	25.1	22.4	5.8	8.9	29.1	32.9
nucleus	3.5	24.2	22.0	5.6	7.9	28.2	41.6
SimCTG	7.9	28.8	26.0	7.5	9.7	30.6	13.3
+context before $\mathcal{MN}_{\text{v}}$							
beam	**14.0**	31.2	29.3	**12.0**	9.9	**32.4**	11.1
top-k	4.9	25.1	23.5	5.8	7.9	28.3	33.2
nucleus	4.2	24.0	22.78	5.5	7.4	27.2	42.2
SimCTG	7.7	29.0	26.1	7.6	8.4	30.9	12.7

Table 2: Automatic evaluation results on VIST test set. All listed models are trained with curriculum learning and contrastive loss using GPT2-xl as language generator. B-4: BLEU-4; M: METEOR; R-L: ROUGE-L; C: CIDEr; S: SPICE; BR: BLEURT; PPL: Perplexity.

[6] We choose GPT2-xl, concatenation before mapping network, with curriculum learning and contrastive training, based on automatic metrics.

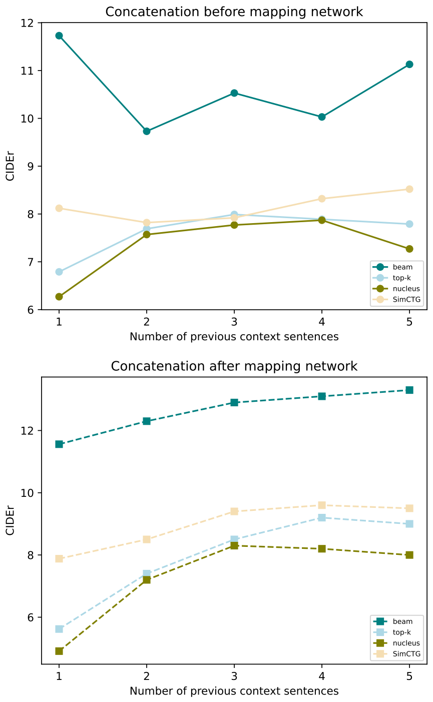


Figure 2: Impact of context length: CIDEr of various number of previous context sentences with concatenation before (top) and after (bottom) $\mathcal{MN}_{\text{v}}$.

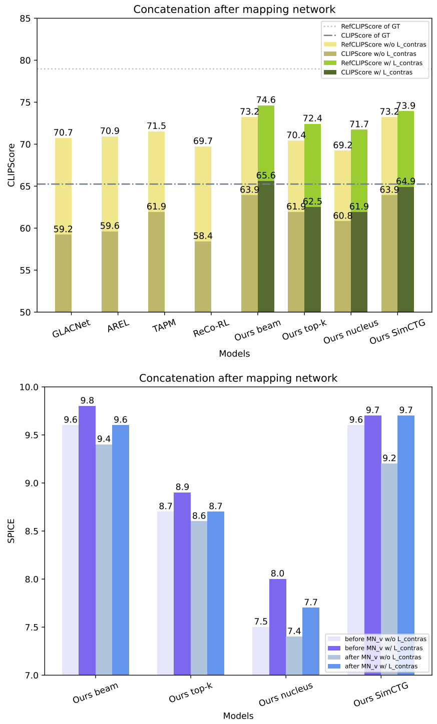


Figure 3: Impact of contrastive training object: CLIPScore (top) and SPICE (bottom) of training our models without or with $\mathcal{L}_{\text{contras}}$.

5.1 Automatic Evaluation

Table 2 outlines the results of automatic metrics among the baselines[7] and our models with curriculum learning, contrastive training and GPT2-xl as the decoder (we consider the impact of different decoder model sizes further below). These results suggest that our model is comparable to or better than the strong baselines on most automatic metrics.

In our experiments, we found that using or not using curriculum learning has no significant impact on automatic metrics (see the full report in the Appendix C). In what follows, we will specifically analyze the impact of the textual context, contrastive training, language model size, and decoding strategies on our method, plus the evaluation of linguistic diversity.

Textual context. Table 2 demonstrates that the combination of textual context (num of previous sentences = 1) brings a consistent improvement, both when concatenation is before and after $\mathcal{MN}_{\text{v}}$. The third and the fourth blocks of Table 2 show that the choice of concatenation strategy does not have much impact on the performance.

Figure 2 shows the impact of concatenating different numbers of previous sentences as context, in both settings. For concatenation before $\mathcal{MN}_{\text{v}}$ (top in Figure 2), we observe that performance tends to decline as context gets longer when decoding with beam search and contrastive search. Whereas, the performance slightly improves for top-k and nucleus sampling when the number of context sentences is less than 3 and 4, respectively. This may be due to the restriction of the maximum length of the input to CLIP to 77 tokens [8]. For the context concatenation after $\mathcal{MN}_{\text{v}}$ (bottom in Figure 2), extending the context length marginally enhances performance, yet it also incurs additional computational costs because of the quadratic complexity of the attention mechanism in GPT2.

Contrastive training. We explore the impact of the contrastive training objective with CLIPScore and RefCLIPScore (Hessel et al., 2021) shown on the top of Figure 3. Contrastive training brings about a clear gain for both CLIPScore and RefCLIPScore, as the contrastive loss serves to minimize the difference between the generated text and

[7] Following the original papers, all the baselines use beam search as decoding strategy.

[8] When the previous context length exceeds 77 tokens, we discard the excess.

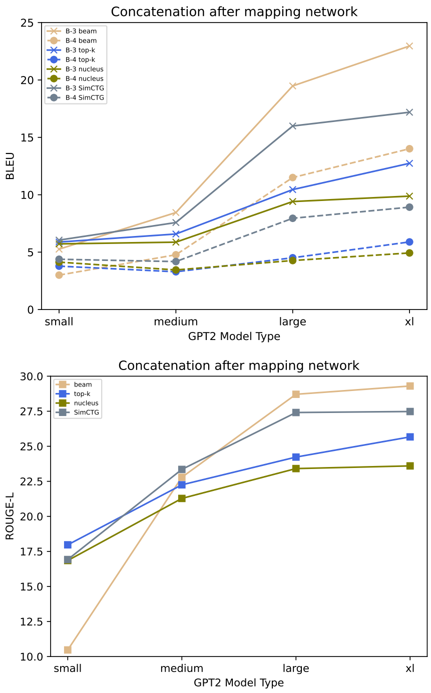


Figure 4: Impact of language model size: BLEU-3, 4 (top) and ROUGE-L (bottom) of our models using GPT2-small, medium, large and xl as text generator with textual context concatenation after $\mathcal{MN}_{\mathrm{v}}$.

the image content in the semantic space of CLIP. In addition to the improvement of text-image similarity, incorporating $\mathcal{L}_{\text{contras}}$ also produces higher SPICE scores, as shown on the bottom of Figure 3. This implies that stories generated with contrastive training are more semantically accurate and detailed, effectively describing important elements and their interrelations in the images.

Language model size. Figure 4 illustrates the performance of various decoding methods applied to different sizes of the GPT2 model. As the model size increases, all decoding methods tend to yield higher BLEU and ROUGE-L scores, especially when comparing GPT2-small to GPT2-large, with limited additional benefits accrued from the larger GPT2-xl. Full results of different language models are in Appendix C.

Decoding strategies. Under identical training, different decoding methods exhibit varying performance across various automatic metrics (as shown in Table 2, Figures 2, 3, 4). Beam search performs the best among all automatic metrics followed by SimCTG, while top-k and nucleus sampling score worse. Though beam search suffers from high repetition and yields very generic text, it seems to align better with the ground truth based on standard automatic metrics in image captioning. On the other hand, decoding methods that aim at alleviating text degeneration, like top-k and nucleus sampling, tend to generate stories that differ from the ground truth, perhaps due to hallucination. SimCTG seems to strike a better balance between grounding and degeneration for VIST. These somewhat counter-intuitive results provide the strongest motivation for our human evaluation, which does not rely on a metric-based comparison of generated text to ground- truth narratives.

	rep-1↓	rep-2↓	rep-3↓	rep-4↓	diversity↑
GT	26.94	4.22	1.03	0.39	94.43
GLACNet	48.43	27.77	20.86	15.97	48.03
AREL	45.20	22.04	15.16	10.98	58.88
TAPM	36.16	10.02	5.16	2.89	82.87
ReCo-RL	33.58	3.14	**0.11**	**0.02**	97.27
Concatenate **before** $\mathcal{MN}_{\mathrm{v}}$, **without** contrastive training, GPT2-xl					
beam	55.33	37.22	29.49	23.91	33.68
top-k	26.80	2.80	0.39	0.08	96.74
nucleus	24.72	2.07	0.23	0.05	97.64
SimCTG	35.02	8.53	2.53	0.89	88.36
Concatenate **before** $\mathcal{MN}_{\mathrm{v}}$, **with** contrastive training, GPT2-xl					
beam	48.31	26.18	18.32	13.38	52.23
top-k	26.55	2.67	0.36	0.08	96.91
nucleus	**24.40**	**2.04**	0.27	0.06	**97.69**
SimCTG	33.16	7.18	1.87	0.61	90.53

Table 3: Text degeneration analysis with rep-1,2,3,4 and diversity score.

Linguistic diversity assessment. The diversity metrics in Table 3 show that beam search suffers from severe text degeneration and 'stammering', that is, generating repeated sequences. In contrast, our models with nucleus sampling provide the most diverse expressions. As shown in the second and third blocks in Table 3, training our model with contrastive loss can also alleviate the degeneration problem with beam search decoding. This further supports the effectiveness of contrastive training in reducing repetitive text.

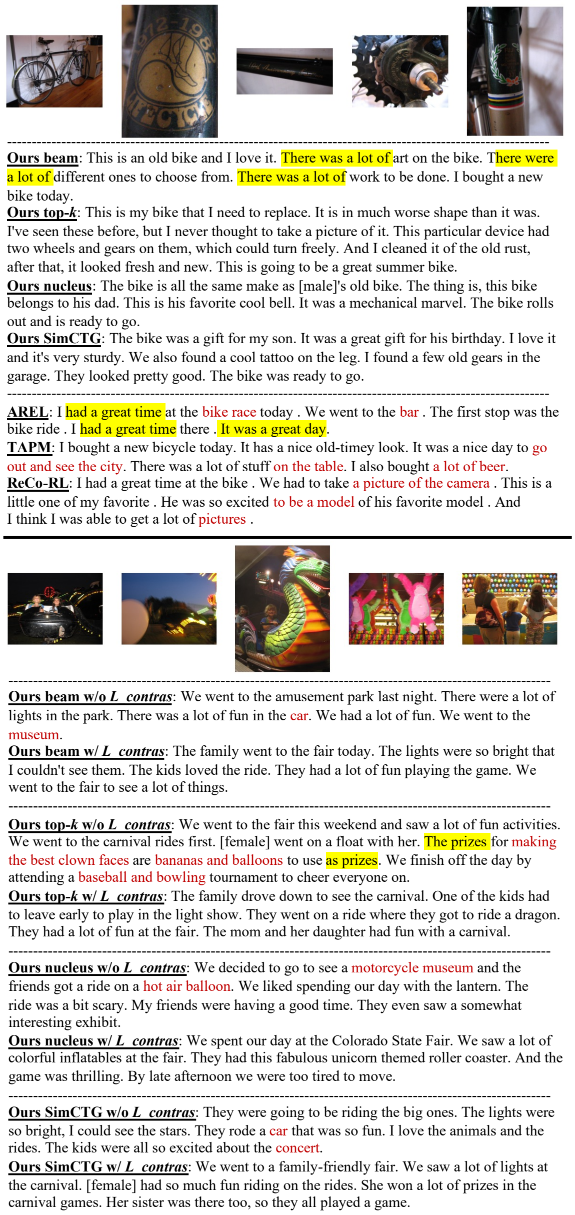
5.2 Human Evaluation

Table 4 displays the means of human rating scores for ground truth (GT), GLACNet, AREL, TAPM and our model with four decoding methods.

Our model with SimCTG decoding outperforms other approaches in terms of Visual Grounding, Coherence and Informativeness. Our model with top-k performs the best in Interestingness. Thus, stories generated by our model compare favorably to baselines in human evaluation. Crucially, we

	Visual Grounding	Coherence	Interestingness	Informativeness
GT	4.10	3.71	3.10	3.61
GLACNet	2.75	2.19	1.78	2.06
AREL	2.85	2.26	1.83	2.20
TAPM	3.16	2.82	2.34	2.61
Ours beam	2.95	2.11	1.80	2.17
Ours top-k	3.01	2.57	**2.40**	2.67
Ours nucleus	2.72	2.42	2.27	2.41
Ours SimCTG	**3.20**	**2.85**	2.27	**2.68**
F(6,293)	6.38	18.46	19.05	15.30
p-value	1.16e-6	6.17e-21	1.22e-21	3.45e-17

Table 4: Human evaluation results: mean rating scores for ground truth (GT), baselines and our models, plus F-statistic and p-value of a one-way ANOVA comparing models on each evaluation dimension.



Ours beam: This is an old bike and I love it. There was a lot of art on the bike. There were a lot of different ones to choose from. There was a lot of work to be done. I bought a new bike today.

Ours top-*k*: This is my bike that I need to replace. It is in much worse shape than it was. I've seen these before, but I never thought to take a picture of it. This particular device had two wheels and gears on them, which could turn freely. And I cleaned it of the old rust, after that, it looked fresh and new. This is going to be a great summer bike.

Ours nucleus: The bike is all the same make as [male]'s old bike. The thing is, this bike belongs to his dad. This is his favorite cool bell. It was a mechanical marvel. The bike rolls out and is ready to go.

Ours SimCTG: The bike was a gift for my son. It was a great gift for his birthday. I love it and it's very sturdy. We also found a cool tattoo on the leg. I found a few old gears in the garage. They looked pretty good. The bike was ready to go.

AREL: I had a great time at the bike race today . We went to the bar . The first stop was the bike ride . I had a great time there . It was a great day.

TAPM: I bought a new bicycle today. It has a nice old-timey look. It was a nice day to go out and see the city. There was a lot of stuff on the table. I also bought a lot of beer.

ReCo-RL: I had a great time at the bike . We had to take a picture of the camera . This is a little one of my favorite . He was so excited to be a model of his favorite model . And I think I was able to get a lot of pictures .

Ours beam w/o *L_contras*: We went to the amusement park last night. There were a lot of lights in the park. There was a lot of fun in the car. We had a lot of fun. We went to the museum.

Ours beam w/ *L_contras*: The family went to the fair today. The lights were so bright that I couldn't see them. The kids loved the ride. They had a lot of fun playing the game. We went to the fair to see a lot of things.

Ours top-*k* w/o *L_contras*: We went to the fair this weekend and saw a lot of fun activities. We went to the carnival rides first. [female] went on a float with her. The prizes for making the best clown faces are bananas and balloons to use as prizes. We finish off the day by attending a baseball and bowling tournament to cheer everyone on.

Ours top-*k* w/ *L_contras*: The family drove down to see the carnival. One of the kids had to leave early to play in the light show. They went on a ride where they got to ride a dragon. They had a lot of fun at the fair. The mom and her daughter had fun with a carnival.

Ours nucleus w/o *L_contras*: We decided to go to see a motorcycle museum and the friends got a ride on a hot air balloon. We liked spending our day with the lantern. The ride was a bit scary. My friends were having a good time. They even saw a somewhat interesting exhibit.

Ours nucleus w/ *L_contras*: We spent our day at the Colorado State Fair. We saw a lot of colorful inflatables at the fair. They had this fabulous unicorn themed roller coaster. And the game was thrilling. By late afternoon we were too tired to move.

Ours SimCTG w/o *L_contras*: They were going to be riding the big ones. The lights were so bright, I could see the stars. They rode a car that was so fun. I love the animals and the rides. The kids were all so excited about the concert.

Ours SimCTG w/ *L_contras*: We went to a family-friendly fair. We saw a lot of lights at the carnival. [female] had so much fun riding on the rides. She won a lot of prizes in the carnival games. Her sister was there too, so they all played a game.

Figure 5: Qualitative examples of our model and baselines. Words highlighted in yellow are repetitive expressions, and words in red represent content that is not relevant to the image sequence.

observe a strong discrepancy between the human evaluation results and automatic metrics. In particular, our model with beam search decoding is ranked low on human judgments, whereas it tends to be ranked highly on automatic metrics, especially those relying on a token-matching comparison to the reference texts. A one-way ANOVA (see the last row of Table 4) shows that overall, differences between models on each of the human evaluation criteria are significant. We follow this up with pairwise comparisons using Tukey's HSD to identify the significant pairwise differences; see the Appendix D for complete results. We find that our model significantly outperforms GLACNet and AREL on human judgments, but is not statistically distinguishable from the other models, though our model leads in terms of mean values.

5.3 Qualitative Case Analysis

The first image sequence in Figure 5 shows stories generated by our models and the baselines in which our model's productions better ground to the input images involving an old bicycle and its various parts. For example, TAPM includes unrelated expressions such as "*on the table*" and "*a lot of beer*", and ReCo-RL erroneously mentions "*a picture of the camera*" and "*to be a model*". In contrast, our model consistently generates text closely relevant to the theme of "*bike*" and provides more specific details, avoiding generic expressions like "*had a great time*". Our model demonstrates a promising ability to write coherent narratives with clear storylines, even for the challenging image sequence in the upper part of Figure 5 (where all images are of objects). For instance, the story from our model with top-k decoding features a fairly clear narrative arc, wherein the narrator discovers a bike in poor condition that is restored after repair and cleaning. This further confirms our model's ability to generate more relevant and engaging stories.

The second image sequence of Figure 5 compares the stories generated by our models without and with contrastive training. The contrastive training forces the model to generate more visually grounded stories with fewer irrelevant elements, that is, hallucinations. However, defining hallucinations in open-ended generation tasks like

VIST remains challenging. While hallucinations can disrupt the story-image correspondence, they can also create intriguing narratives. The storytelling based on images is expected to incorporate elements which are not strictly descriptive of visual contents. For example, the last sentence in the story by our model with top-k decoding and contrastive training, "*We finish off the day by attending a baseball and bowling tournament to cheer everyone on*" is not directly reflected in the images but adds relevant context and imaginative extension. Balancing hallucination and creativity is left for future work.

6 Conclusion

We present a simple yet effective framework for visual storytelling that utilizes pretrained multimodal models with a lightweight vision-language mapping network to construct prefixes for LLMs. Our model enhances the coherence of multi-sentence stories by integrating contextual information. In addition to teacher-forcing loss, we use a curriculum training scheme and image-text contrastive loss to enhance the concretness and visual grounding of generated stories. Extensive evaluation on the VIST benchmark using both automatic metrics and human assessment shows that our model obtains strong results compared to SOTA methods. We empirically confirm that our model demonstrates the ability to generate coherent stories that are closely tied to visual content, and possess more creative and engaging details with minimal degeneration. Our study contributes to improved evaluation practices in text generation, recommending a specific human evaluation setup for visual storytelling that assesses four key output qualities. Such evaluation enables informative model comparisons and better insight into the relative strengths of different systems. Results show that automatic metrics, particularly token overlap measures like BLEU, often poorly correspond to human judgments and should not be fully trusted for open-ended tasks like visual storytelling. This echoes similar observations made in other NLG domains (Belz and Reiter, 2006; Reiter and Belz, 2009; Reiter, 2018; Moramarco et al., 2022).

Limitations. Despite having employed diverse automatic metrics and comprehensive human evaluations to assess our models' generated stories, we recognize substantial opportunity for enhancing the evaluation methodology of visual storytelling. As discussed above, correlating with ground-truth text or grounding to the visual content represents just a one-sided view, which downplays the role of diversity and creativity in storytelling. While our proposed human evaluation aims for thorough assessment, human annotation is costly and cannot be continuously applied during model development. Future research could explore the balance in visual storytelling between factuality and groundedness on the one hand, and *justified* deviation from the images in the interest of creativity on the other.

Additionally, our model exhibits certain biases, such as producing wedding-related stories from images of churches, even though there are no wedding-related elements in the images. This may stem from the biases in VIST dataset or the pretraining data of CLIP and GPT2.

Lastly, this study primarily investigates the utility and performance of two specific pre-trained models, CLIP and GPT-2. While these models have demonstrated broad applicability and strong performance across various tasks, they represent only a subset of the rapidly evolving landscape of pre-trained vision an language models. Future work could benefit from incorporating a wider array of models, such as BLIP-2 (Li et al., 2023), LLaVA (Liu et al., 2023a), Llama 3 (Meta AI, 2024) and Mistral (Mistral AI, 2024), to provide a more comprehensive understanding of the strengths and limitations inherent to different foundation models.

Ethics Statement. In this research, we employ pretrained multimodal models LLMs to transform images into narratives. There's a possibility that any biases inherent in the pre-training data may unintentionally be reflected in the text generated, potentially leading to uncontrolled biases. While our examination did not observe such problems, we recognize it as a potential concern that might affect the integrity of the generated content. Regarding the VIST dataset and the models used in this study, we are not aware of any major ethical concerns they may pose on their own. However, we acknowledge the potential for biases present in the original VIST data to influence both our models and their evaluations. Our research has received approval from the Ethics Board of our institution, ensuring compliance with ethical standards in human evaluation processes. All the human evaluation data collected has been de-identified to protect the privacy and security of all participants involved.

Acknowledgements

We thank the three anonymous reviewers for helpful comments, and the participants in our human evaluation for their role in evaluating our systems.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

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

Anja Belz and E Reiter. 2006. Comparing Automatic and Human Evaluation of NLG Systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 313–320.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. PADA: Example-based Prompt Learning for on-the-fly Adaptation to Unseen Domains. *Transactions of the Association for Computational Linguistics*, 10:414–433.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Charles B. Callaway and James C. Lester. 2001. Narrative prose generation. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001*, pages 1241–1250. Morgan Kaufmann.

Khyathi Chandu, Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. 2019. "my way of telling a story": Persona based grounded story generation. In *Proceedings of the Second Workshop on Storytelling*, pages 11–21, Florence, Italy. Association for Computational Linguistics.

Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. 2021. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 999–1008.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Ruichao Fan, Hanli Wang, Jinjing Gu, and Xianhui Liu. 2022. Visual storytelling with hierarchical BERT semantic guidance. In *ACM Multimedia Asia*, MMAsia '21, pages 1–7. Association for Computing Machinery.

Pablo Gervás. 2009. Computational approaches to storytelling and creativity. *AI Magazine*, 30(3):49–62.

Diana Gonzalez-Rico and Gibran Fuentes-Pineda. 2018. Contextualize, show and tell: A neural visual storyteller.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023. Autoad: Movie description in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18930–18940.

Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. BERTese: Learning to speak to BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3618–3623, Online. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual Writing Prompts: Character-Grounded Story Generation with Curated Image Sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.

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 *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7952–7960. AAAI Press.

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

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

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

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. abs/1805.10973.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17283–17300. PMLR.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. Association for Computational Linguistics.

Jiacheng Li, Haizhou Shi, Siliang Tang, Fei Wu, and Yueting Zhuang. 2019a. Informative visual storytelling with cross-modal rules. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 2314–2322. Association for Computing Machinery.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Nanxing Li, Bei Liu, Zhizhong Han, Yu-Shen Liu, and Jianlong Fu. 2019b. Emotion reinforced visual storytelling. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 297–305.

Tengpeng Li, Hanli Wang, Bin He, and Chang Wen Chen. 2022. Knowledge-enriched attention network with group-wise semantic for visual storytelling. pages 1–12. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 605–612. ACL.

Danyang Liu and Frank Keller. 2023. Detecting and grounding important characters in visual stories. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13210–13218. AAAI Press.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations*.

Meta AI. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. Accessed: [Jul 29 2024].

Mistral AI. 2024. Mistral llms. `https://docs.mistral.ai/getting-started/models/`. Accessed: [Jul 29 2024].

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation. ArXiv: 2204.00447.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Aleksandar Petrov, Philip Torr, and Adel Bibi. 2024. When do prompting and prefix-tuning work? a theory of capabilities and limitations. In *The Twelfth International Conference on Learning Representations*.

Mengshi Qi, Jie Qin, Di Huang, Zhiqiang Shen, Yi Yang, and Jiebo Luo. 2021. Latent memory-augmented graph transformer for visual storytelling. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4892–4901. ACM.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Gisela Redeker. 2000. Coherence and structure in text and discourse. *Abduction, belief and context in dialogue*, 233(263).

Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401. Place: Cambridge, MA Publisher: MIT Press.

Ehud Reiter and Anja Belz. 2009. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistcs*, 35(4):529–558.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 21548–21561. Curran Associates, Inc.

Aditya Surikuchi, Sandro Pezzelle, and Raquel Fernández. 2023. GROOViST: A metric for grounding objects in visual storytelling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3331–3339, Singapore. Association for Computational Linguistics.

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

Eileen Wang, Caren Han, and Josiah Poon. 2022. RoViST: Learning robust metrics for visual storytelling. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2691–2702, Seattle, United States. Association for Computational Linguistics.

Eileen Wang, Caren Han, and Josiah Poon. 2024. SCO-VIST: Social interaction commonsense knowledge-based visual storytelling. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1602–1616, St. Julian's, Malta. Association for Computational Linguistics.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. Issue: 1.

Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. 2020. Storytelling from an image stream using scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page 8.

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018b. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 899–909. Association for Computational Linguistics.

Yuechen Wang, Wengang Zhou, Zhenbo Lu, and Houqiang Li. 2023. Text-only training for visual storytelling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 3686–3695. ACM.

Chunpu Xu, Min Yang, Chengming Li, Ying Shen, Xiang Ao, and Ruifeng Xu. 2021. Imagine, reason and write: Visual storytelling with graph knowledge and relational reasoning. 35(4):3022–3029. Number: 4.

Dingyi Yang and Qin Jin. 2023. Attractive storyteller: Stylized visual storytelling with unpaired text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11053–11066. Association for Computational Linguistics.

Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu Sun. 2019. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, page 7.

Youngjae Yu, Jiwan Chung, Heeseung Yun, Jongseok Kim, and Gunhee Kim. 2021. Transitional adaptation of pretrained models for visual storytelling. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12653–12663. IEEE.

Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. 2021. Two heads are better than one: Hypergraph-enhanced graph reasoning for visual event ratiocination. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12747–12760. PMLR. ISSN: 2640-3498.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Wang, Miguel Eckstein, and William Yang Wang. 2023. Visualize before you write: Imagination-guided open-ended text generation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 78–92, Dubrovnik, Croatia. Association for Computational Linguistics.

A Experimental Details of Training, Inference and Automatic Evaluation

We use CLIP RN50x4 as the image encoder backbone to extract visual features offline[9] and GPT2-small, medium, large and xl as the language decoder. The mapping network is a Transformer-based model with 8 multi-head self-attention layers with 8 heads each. We set the CLIP embedding length as 20 and visual prefix length as 20. We stop the text generation when an end of sequence token is predicted, otherwise we limit the maximum length to 30 tokens. For each experiment, we use a single NVIDIA A100 for training and inference. Other empirically tuned hyperparameters are listed in the Table 5.

Hyperparameters	Value
Batch size	50
Training epochs	10
N_{nll}	6
λ	0.3
Optimizer	Adam
Learning rate	2e-5
Weight decay	1e-4
Warmup steps	1300
Max length	30
Num of beams	5
k in top-k	50
p in nucleus sampling	0.9
Top-k in SimCTG	5
Degeneration penalty in SimCTG	0.8
Temperature	1.0

Table 5: Hyperparameter settings.

As for the automatic evaluation, we use pycocoevalcap[10] library to compute BLEU, ROUGE-L, CIDEr and SPICE, and use the official VIST challenge evaluation code[11] to compute METEOR. We report BLEURT[12] score with BLEURT-20 as the checkpoint, CLIPScore and RefCLIPScore[13] with ViT-B/32 as the base model, and the mean perplexity[14] score calculated by GPT2.

B Human Evaluation Survey

For the human evaluation survey, participants were asked to rate each pair, consisting of a story and an image sequence, on the following criteria: (1) **Visual Grounding** assesses how accurately and reasonably the story corresponds to the content in the image sequence; (2) **Coherence** evaluates how logical and consistent the story is; (3) **Interestingness** measures how the story captures the reader's interest through unique ideas or expressions; (4) **Informativeness** evaluates how specific and detailed the story is in narrating the scene, objects, and events depicted in the images, rather than relying on highly generic descriptions.

Figure 6 presents the instruction, sample image sequence stories provided in the human evaluation questionnaire. The introduction aims to make participants fully understand the specific meaning of the four evaluation criterion and the corresponding score scale. The samples are intended to help participants build a mental expectation of the image sequences and stories they will see, in order to avoid the order in which the images and stories appear influencing their judgment. In Figure 7, we show an example question that consists of a story generated by 1 out of 8 models, a sequence of 5 images, and 4 direct rating questions. We randomly shuffled all 100 image sequences and their corresponding 8 stories generated by different models in an even manner. In each participant's survey, which includes 32 questions, the same image sequence will not appear twice, and stories from all 8 models are included. We only asked each participant to complete 32 questions (median completion time is 20mins 8secs), avoiding their judgment being affected due to excessively long periods of focus at a single survey task. We hired 75 annotators (38 females, 37 males) on Prolific at a hourly rate of £13.41, all of whom are proficient in English with at least the college education level.

[9] We tried both CLIP RN50x4 and CLIP ViT/B-32 in the preliminary experiments, and RN50x4 performs a little bit better than ViT/B-32.

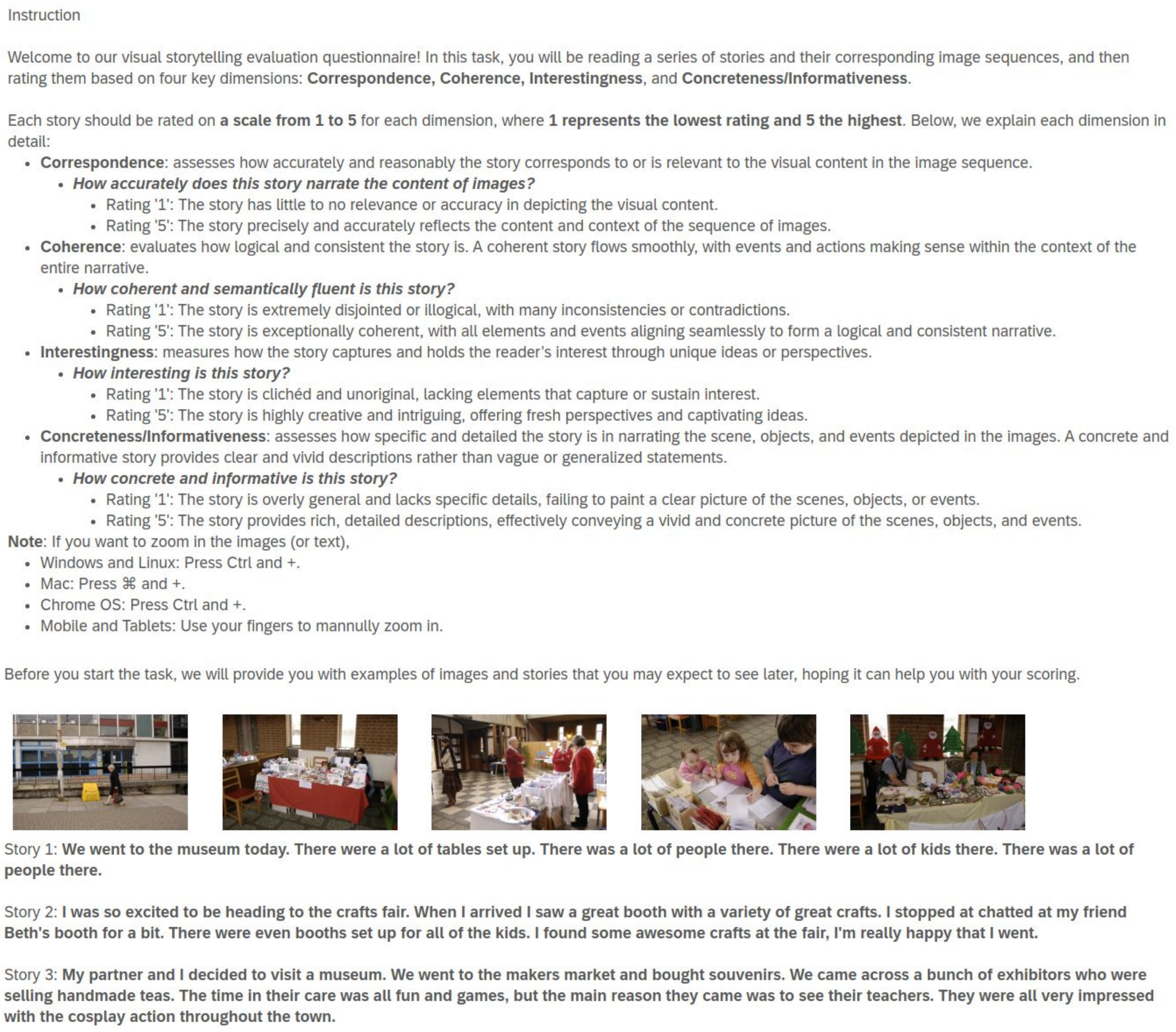
[10] https://github.com/tylin/coco-caption

[11] https://github.com/windx0303/VIST-Challenge-NAACL-2018

[12] https://github.com/google-research/bleurt

[13] https://github.com/jmhessel/clipscore

[14] https://huggingface.co/spaces/evaluate-metric/perplexity



Instruction

Welcome to our visual storytelling evaluation questionnaire! In this task, you will be reading a series of stories and their corresponding image sequences, and then rating them based on four key dimensions: **Correspondence, Coherence, Interestingness**, and **Concreteness/Informativeness**.

Each story should be rated on **a scale from 1 to 5** for each dimension, where **1 represents the lowest rating and 5 the highest**. Below, we explain each dimension in detail:

- **Correspondence**: assesses how accurately and reasonably the story corresponds to or is relevant to the visual content in the image sequence.
  - ***How accurately does this story narrate the content of images?***
    - Rating '1': The story has little to no relevance or accuracy in depicting the visual content.
    - Rating '5': The story precisely and accurately reflects the content and context of the sequence of images.
- **Coherence**: evaluates how logical and consistent the story is. A coherent story flows smoothly, with events and actions making sense within the context of the entire narrative.
  - ***How coherent and semantically fluent is this story?***
    - Rating '1': The story is extremely disjointed or illogical, with many inconsistencies or contradictions.
    - Rating '5': The story is exceptionally coherent, with all elements and events aligning seamlessly to form a logical and consistent narrative.
- **Interestingness**: measures how the story captures and holds the reader's interest through unique ideas or perspectives.
  - ***How interesting is this story?***
    - Rating '1': The story is clichéd and unoriginal, lacking elements that capture or sustain interest.
    - Rating '5': The story is highly creative and intriguing, offering fresh perspectives and captivating ideas.
- **Concreteness/Informativeness**: assesses how specific and detailed the story is in narrating the scene, objects, and events depicted in the images. A concrete and informative story provides clear and vivid descriptions rather than vague or generalized statements.
  - ***How concrete and informative is this story?***
    - Rating '1': The story is overly general and lacks specific details, failing to paint a clear picture of the scenes, objects, or events.
    - Rating '5': The story provides rich, detailed descriptions, effectively conveying a vivid and concrete picture of the scenes, objects, and events.

Note: If you want to zoom in the images (or text),

- Windows and Linux: Press Ctrl and +.
- Mac: Press ⌘ and +.
- Chrome OS: Press Ctrl and +.
- Mobile and Tablets: Use your fingers to mannully zoom in.

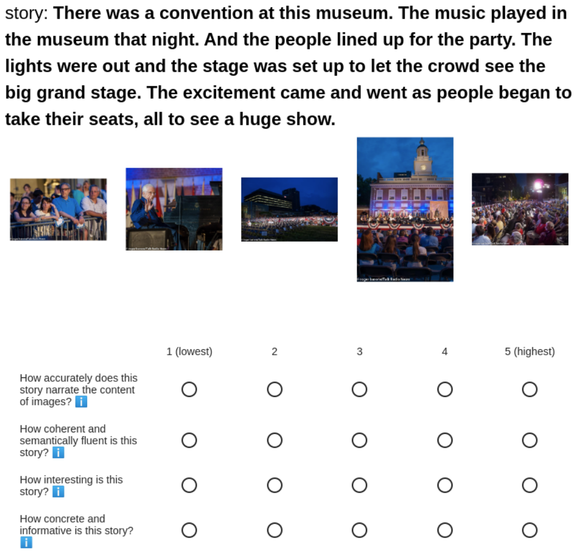
Before you start the task, we will provide you with examples of images and stories that you may expect to see later, hoping it can help you with your scoring.

Story 1: **We went to the museum today. There were a lot of tables set up. There was a lot of people there. There were a lot of kids there. There was a lot of people there.**

Story 2: **I was so excited to be heading to the crafts fair. When I arrived I saw a great booth with a variety of great crafts. I stopped at chatted at my friend Beth's booth for a bit. There were even booths set up for all of the kids. I found some awesome crafts at the fair, I'm really happy that I went.**

Story 3: **My partner and I decided to visit a museum. We went to the makers market and bought souvenirs. We came across a bunch of exhibitors who were selling handmade teas. The time in their care was all fun and games, but the main reason they came was to see their teachers. They were all very impressed with the cosplay action throughout the town.**

Figure 6: Instructions, sample image sequence and corresponding stories we displayed at the beginning of the human evaluation questionnaire.



story: **There was a convention at this museum. The music played in the museum that night. And the people lined up for the party. The lights were out and the stage was set up to let the crowd see the big grand stage. The excitement came and went as people began to take their seats, all to see a huge show.**

	1 (lowest)	2	3	4	5 (highest)
How accurately does this story narrate the content of images? ℹ	○	○	○	○	○
How coherent and semantically fluent is this story? ℹ	○	○	○	○	○
How interesting is this story? ℹ	○	○	○	○	○
How concrete and informative is this story? ℹ	○	○	○	○	○

Figure 7: One example question in the human evaluation questionnaire.

C Additional Results

Table 6: Results of our model with GPT2-xl, textual context concatenation before and after mapping network, +/- contrastive learning and +/-curriculum training.

	B-1	B-2	B-3	B-4	M	R-L	CIDEr	SPICE	BLEURT	PPL	CLIPS.	RefCLIPS.
+curriculum learning, +context before mapping network,-contrastive loss												
Beam	62.76	37.95	22.8	13.91	32.70	30.53	12.02	8.49	31.63	12.23	63.81	72.24
Top-k	46.34	20.65	8.22	3.45	28.17	21.51	5.83	7.82	29.09	32.27	60.73	69.54
Nucleus	43.12	18.36	6.92	2.83	26.53	20.72	5.59	7.38	28.05	44.12	59.26	68.37
SimCTG	56.27	29.84	14.45	7.07	27.96	25.98	8.70	8.87	30.71	13.39	62.65	72.35
+curriculum learning, +context after mapping network,-contrastive loss												
Beam	60.19	35.67	20.45	13.90	32.52	27.84	10.95	8.46	32.37	11.62	62.63	72.66
Top-k	52.73	24.91	10.48	4.67	26.37	23.05	4.66	7.51	29.23	30.22	61.60	70.13
Nucleus	50.65	23.02	9.25	4.04	25.55	22.36	3.83	7.02	28.14	41.07	60.94	70.01
SimCTG	59.76	32.13	15.43	7.58	27.13	25.47	6.94	8.28	31.19	12.82	62.88	72.29
-curriculum learning, +context before mapping network,+contrastive loss												
Beam	63.12	38.41	23.10	14.24	31.68	29.29	11.73	9.79	32.21	11.12	65.61	74.58
Top-k	46.58	22.10	9.16	5.93	25.28	25.71	6.79	8.86	28.20	33.67	62.50	72.37
Nucleus	44.91	20.43	8.19	4.91	24.26	23.59	6.27	8.03	27.13	40.91	61.89	71.68
SimCTG	56.79	31.65	15.93	8.89	29.02	27.54	8.12	9.71	30.56	13.03	64.87	73.92
-curriculum learning, +context after mapping network,+contrastive loss												
Beam	62.83	38.04	22.87	14.12	31.84	29.20	11.56	9.63	32.43	10.41	64.82	74.17
Top-k	47.25	22.12	9.14	4.29	25.12	22.67	5.62	8.74	29.81	33.28	63.32	72.11
Nucleus	44.40	19.76	7.71	3.73	24.03	21.75	4.91	7.72	28.18	43.92	62.75	71.04
SimCTG	56.90	31.11	15.27	8.37	29.21	26.32	7.88	9.65	31.08	12.46	64.59	73.72

Table 7: Results of our model with different GPT2 language models, textual context concatenation after mapping network, and without contrastive learning and curriculum training.

	B-1	B-2	B-3	B-4	M	R-L	CIDEr	SPICE	BLEURT	PPL	CLIPS.	RefCLIPS.
GPT2-**small**												
Beam	23.63	10.53	5.26	3.00	7.16	10.47	11.41	4.66	26.46	13.90	53.52	60.73
Top-k	24.75	13.73	5.88	3.78	9.89	17.97	6.62	4.96	24.07	43.87	50.87	59.26
Nucleus	26.44	13.90	5.72	4.13	10.05	16.85	5.98	5.14	28.20	53.99	50.74	58.96
SimCTG	26.92	14.19	6.05	4.38	10.76	16.92	5.48	4.91	25.59	22.53	51.18	59.44
GPT2-**medium**												
Beam	33.16	15.80	8.45	4.77	9.88	22.79	18.37	7.22	28.63	13.25	57.30	63.48
Top-k	31.83	13.86	6.58	3.29	9.24	22.25	6.91	6.74	26.09	40.23	56.47	64.12
Nucleus	30.49	13.58	5.87	3.45	8.93	21.18	6.33	6.05	25.05	56.75	55.85	63.91
SimCTG	34.81	16.76	7.58	4.18	9.42	23.35	12.90	7.63	28.71	21.59	57.19	63.93
GPT2-**large**												
Beam	56.67	33.23	19.48	11.50	13.36	28.71	18.40	7.66	31.19	11.36	61.22	71.15
Top-k	51.64	24.50	10.45	4.51	13.68	24.23	8.41	7.72	28.17	35.11	59.54	69.35
Nucleus	49.71	22.76	9.41	4.27	13.14	23.41	6.37	7.12	27.07	50.06	58.37	68.19
SimCTG	59.08	32.34	15.99	7.95	13.82	27.41	12.59	7.98	30.64	19.62	61.34	71.14
GPT2-**xl**												
Beam	62.88	38.04	22.96	14.01	14.95	29.30	17.64	9.37	32.37	10.73	62.08	71.77
Top-k	55.76	28.01	12.74	5.89	13.13	25.67	5.61	8.61	29.23	35.68	60.06	69.75
Nucleus	49.29	22.55	9.88	4.93	12.86	23.60	3.86	7.36	28.14	46.17	59.16	68.81
SimCTG	60.52	33.76	17.19	8.92	13.65	27.48	8.01	9.18	31.09	13.92	62.02	71.66

D Human Evaluation Significance Test

We conduct Tukey's HSD pairwise group comparisons of human evaluation scores we collected as shown in Figure 12.

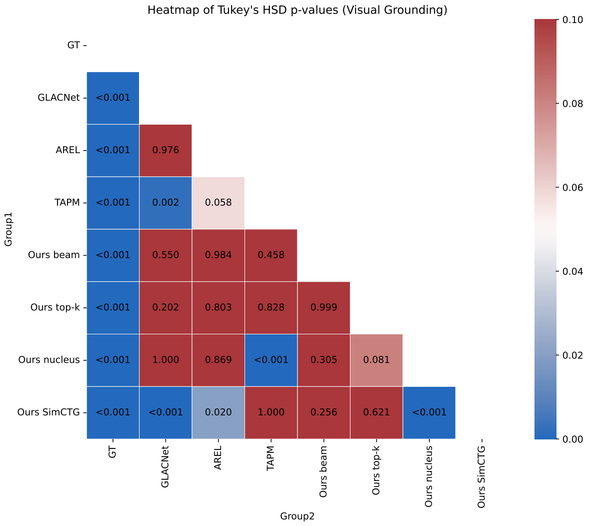


Figure 8: Visual Grounding

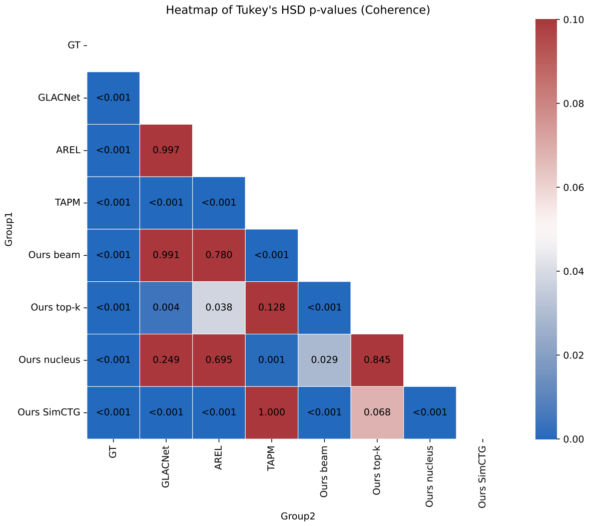


Figure 9: Coherence

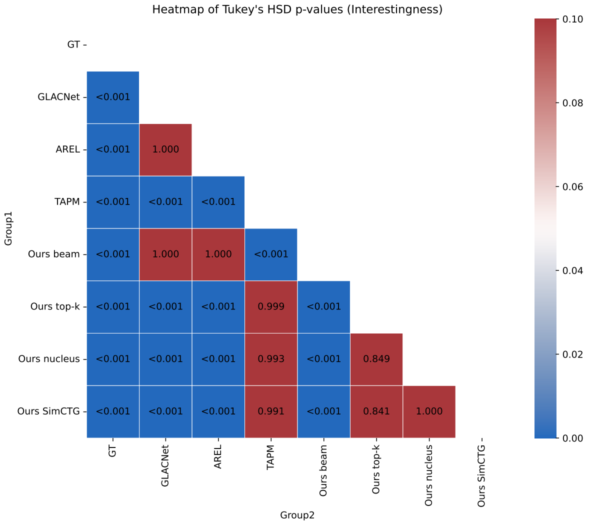


Figure 10: Interestingness

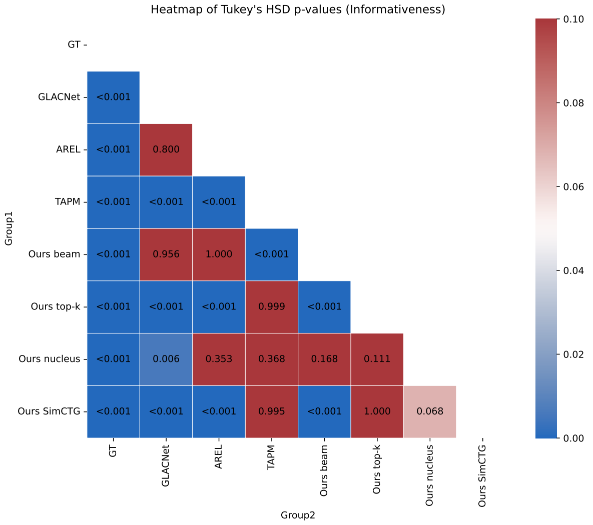


Figure 11: Informativeness

Figure 12: p-values of Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval)