Utrecht University

11-11-2024

From Image Captioning to Visual Storytelling

MSc Thesis Presentation

Admitos Rafail Passadakis

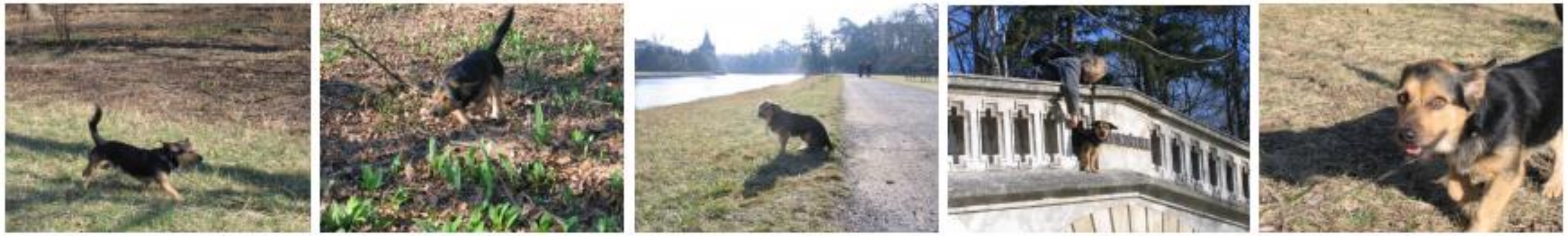
CONTENTS

1. Introduction
2. Related Work
3. Methods
4. Experimental Set-up
5. Results
6. Conclusion
7. Comparisons & My Story

Introduction

OUR TASKS:

- **Image Captioning (IC):** Creating mere descriptions of the depicted elements in the images which can be in isolation (they can be visually uncorrelated).
- **Visual Storytelling (VS):** Conveying a narrative through a sequence of images (can be visually correlated) with the goal of telling a story, capturing the contextual relationship among the events or characters in the images.

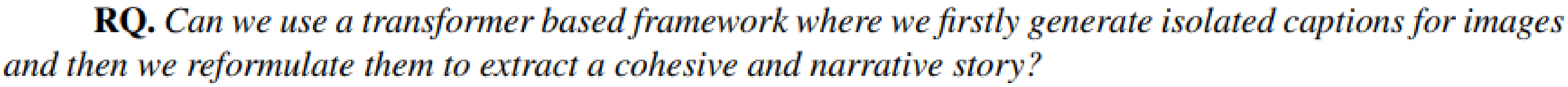


Captions:	*A small dog is running through the grass.*	*A small dog is walking through the leaves.*	*A dog is standing on the side of a road.*	*A woman and her dog on the edge of a bridge*	*A dog standing in the grass with it's tongue out*
Story:	*The black dog was ready to leave.*	*He had a great time on the hike.*	*And was very happy to be in the field.*	*His mom was so proud of him.*	*It was a beautiful day for him.*

Introduction

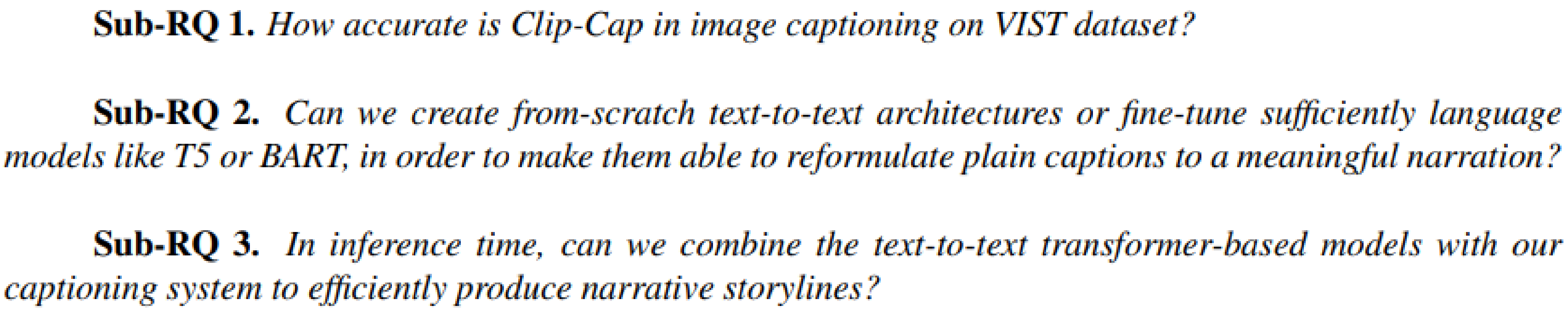
RESEARCH QUESTIONS:

- Previous Approaches: VS and IC are two distinct tasks.
- Our Approach/Motivation: Can we work altogether on those tasks by considering IC as subset of VS? (i.e. using the central information/meaning of captions so to create the story)



RQ. *Can we use a transformer based framework where we firstly generate isolated captions for images and then we reformulate them to extract a cohesive and narrative story?*

But by breaking it down …



Sub-RQ 1. *How accurate is Clip-Cap in image captioning on VIST dataset?*

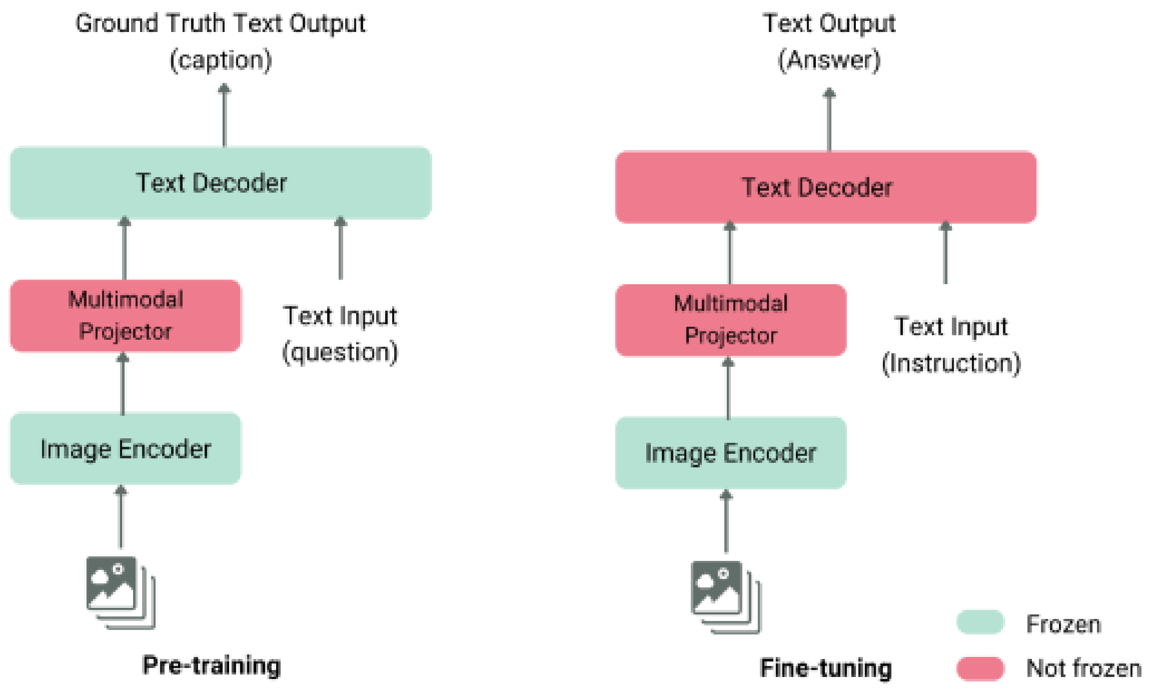
Sub-RQ 2. *Can we create from-scratch text-to-text architectures or fine-tune sufficiently language models like T5 or BART, in order to make them able to reformulate plain captions to a meaningful narration?*

Sub-RQ 3. *In inference time, can we combine the text-to-text transformer-based models with our captioning system to efficiently produce narrative storylines?*

Related Work

VISION & LANGUAGE

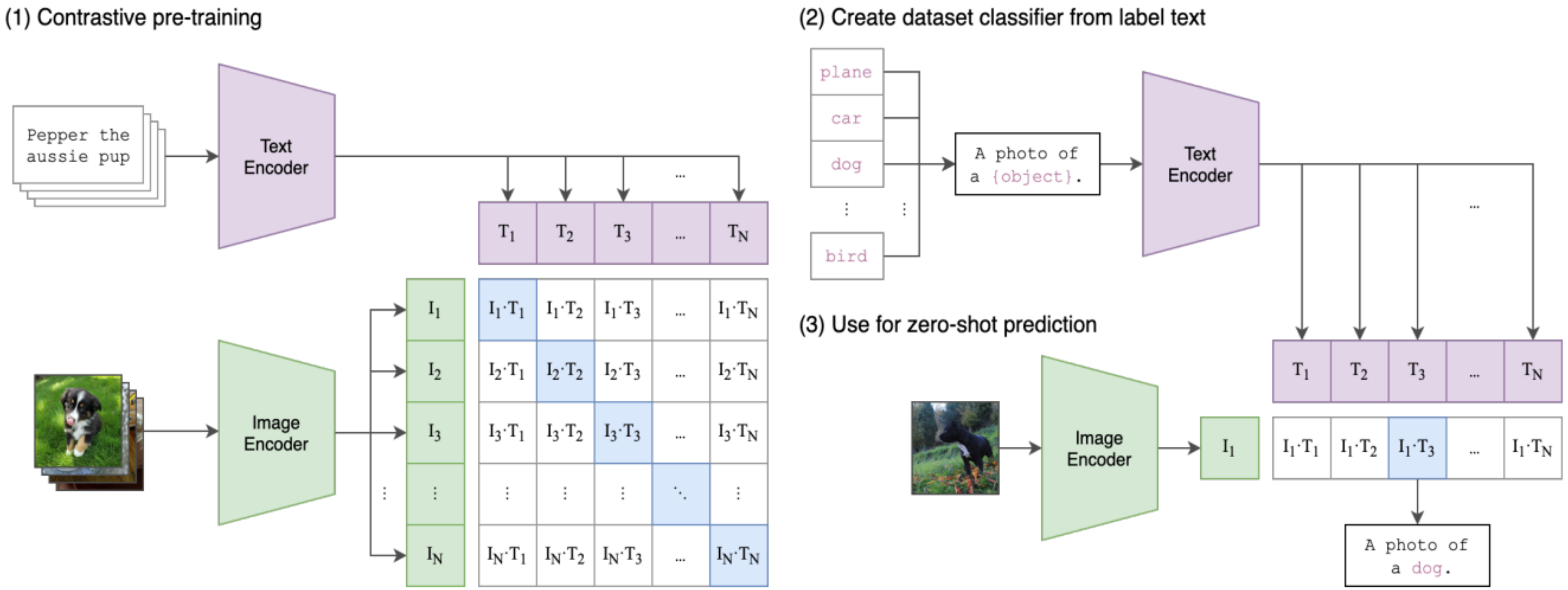
- **Vision & Language Tasks:** Visual Question Answering, Image Captioning, Video Captioning, Visual Storytelling.
- **Image Captioning (2014):** Roots on MSCOCO dataset by Microsoft. Not only *Image classification* or *object recognition* but aimed for understanding the broader context of the scene.
- **Vision & Language Models**: Generative models, such as Transformers, that receive visual and textual inputs and generate text outputs. *Multimodal Projector*: Aligns image and text representations.



Related Work

A FAMOUS V&L MODEL

- **CLIP (**Contrastive Language-Image Pretraining - *Radford et al.***):** A V&L model that is trained using contrastive learning, proposed by OPENAI. It is designed for understanding how well images and text (captions) fit together.
- During training: CLIP attempts to maximize the cosine similarity between correct pairs of (image-captions) and to minimize the cosine similarity of incorrect pairs.
- During inference: CLIP calculates the similarity scores between the vector of a single image with some possible caption vectors and gives a probability distribution indicating the most favorite caption (the one with the highest similarity).
- CLIP is NOT a captioning model.



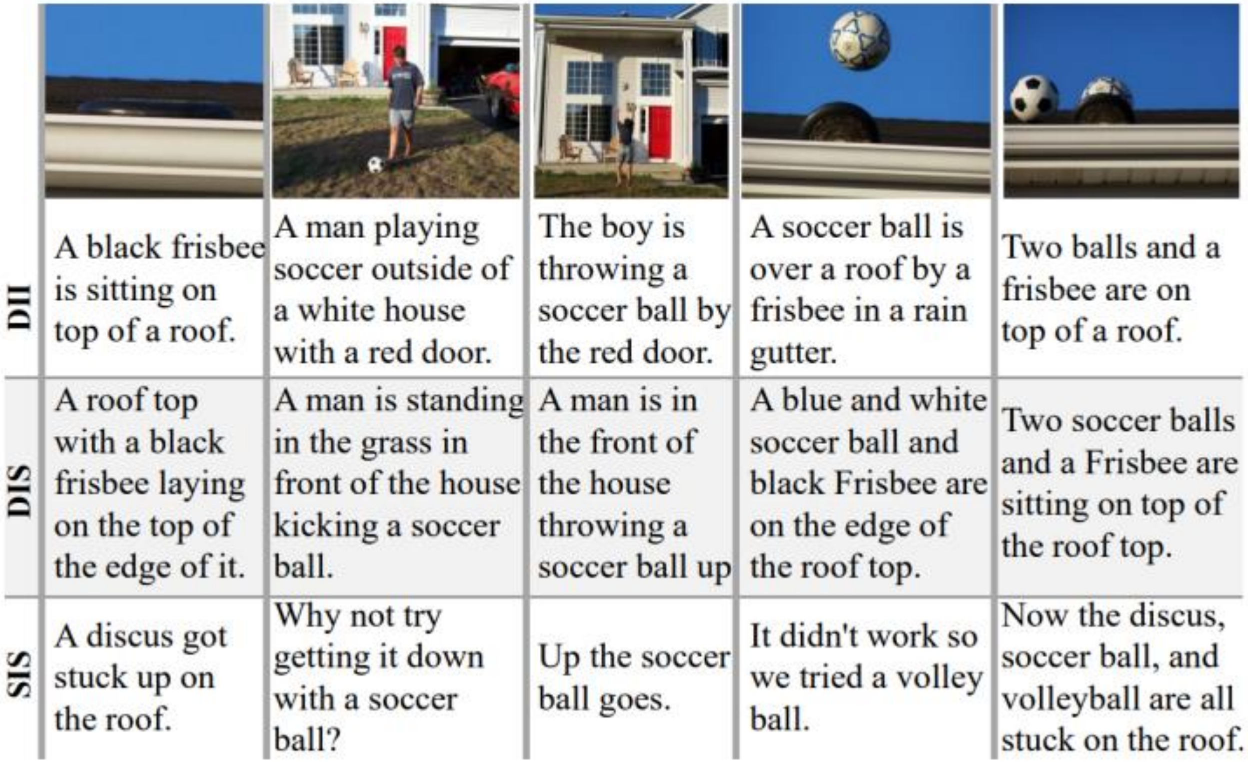
	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$

Related Work

DATASET

- **Visual Storytelling (2016) :** Roots on VIsual STorytelling (VIST) dataset by *Huang et al.*
- Three tiers: *Descriptions of images-in-isolation (DII); Descriptions of images-in-sequence (DIS); Stories for images-in-sequence (SIS).*
- For Captioning we use DII, for Storytelling we use SIS.
- <u>Visual Storytelling is significantly different task than Image captioning.</u>

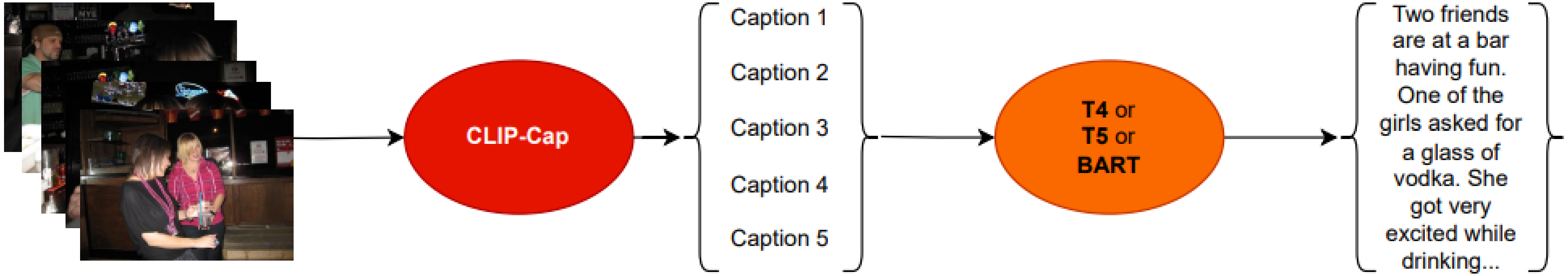


DII	A black frisbee is sitting on top of a roof.	A man playing soccer outside of a white house with a red door.	The boy is throwing a soccer ball by the red door.	A soccer ball is over a roof by a frisbee in a rain gutter.	Two balls and a frisbee are on top of a roof.
DIS	A roof top with a black frisbee laying on the top of the edge of it.	A man is standing in the grass in front of the house kicking a soccer ball.	A man is in the front of the house throwing a soccer ball up	A blue and white soccer ball and black Frisbee are on the edge of the roof top.	Two soccer balls and a Frisbee are sitting on top of the roof top.
SIS	A discus got stuck up on the roof.	Why not try getting it down with a soccer ball?	Up the soccer ball goes.	It didn't work so we tried a volley ball.	Now the discus, soccer ball, and volleyball are all stuck on the roof.

Methods

OVERALL FRAMEWORK

- *Image-to-caption* model: **Clip-Cap**
  - Generates captions for the series of images in the input.
- *Caption-to-story* model: **T4** or **T5** or **BART**
  - Reformulate the mere captions to a narrative story.

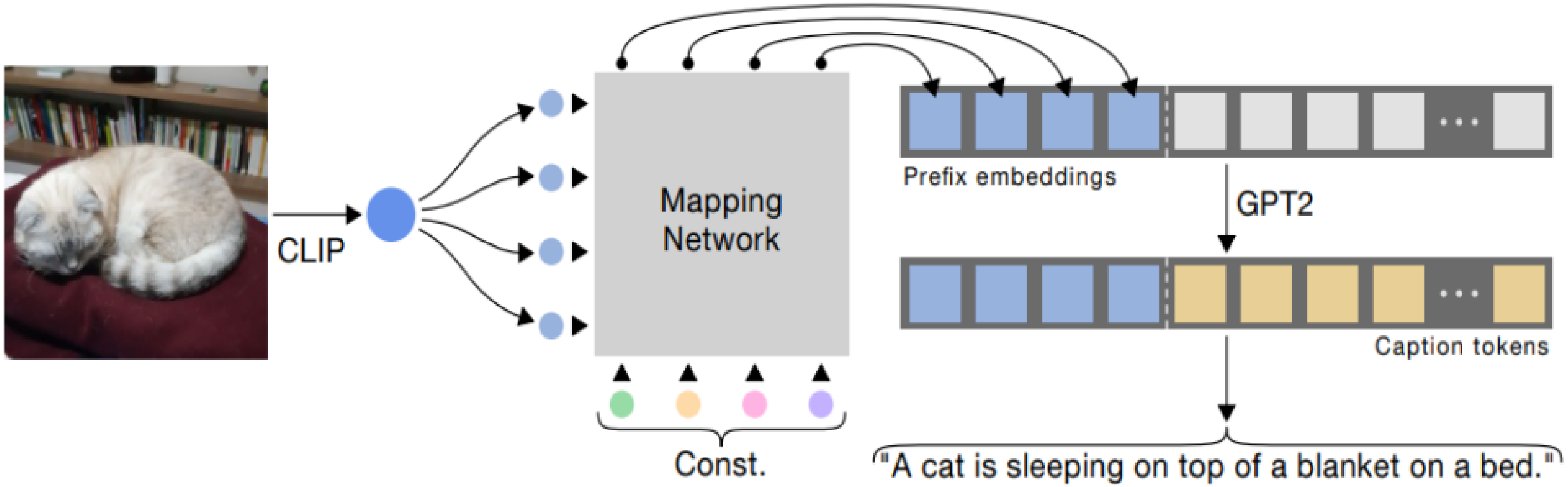


Two friends are at a bar having fun. One of the girls asked for a glass of vodka. She got very excited while drinking...

Methods

IMAGE-to-CAPTION MODEL

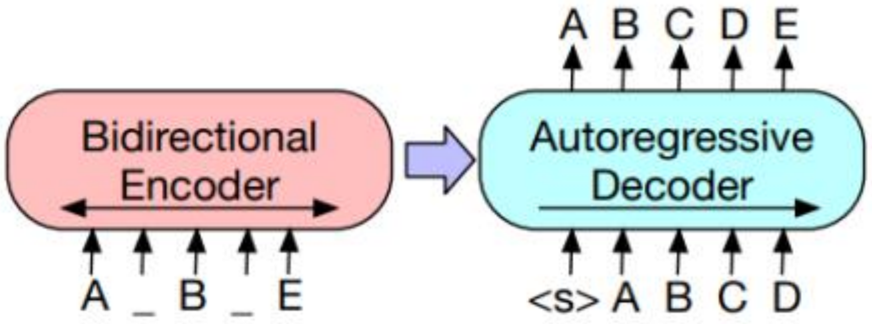
- **Clip-Cap (*Mokady et al.*):** Uses the encoded CLIP output as a prefix to the caption.
- Clip-Cap consists of three parts: 1) CLIP, 2) Mapping Network and 3) Language generator (GPT-2).
- It was fine-tuned and evaluated on three datasets: 1) MSCOCO, 2)Conceptual captions and 3) Nocaps.
- Mapping Network:
  - *Transformer*: Only the Mapping network is fine-tuned → Clip-Cap is extremely lightweight.
  - *MLP*: In that case GPT-2 is fine-tuned as well → Clip-Cap moderately heavier but still okay.



Methods

CAPTION-to-STORY MODEL

- **BART (**Bidirectional Auto-Regressive Transformer - *Lewis et al.***):** A denoising autoencoder transformer that maps a corrupted document back to the original document.
- Pre-trained Using Document Corruption Techniques: Token Masking, Token Deletion, Document Rotation, Sentence Permutation and Text Infilling.
- It combines BERT-like Encoders & GPT-like Decoders → Can be used for both understanding and generation tasks.

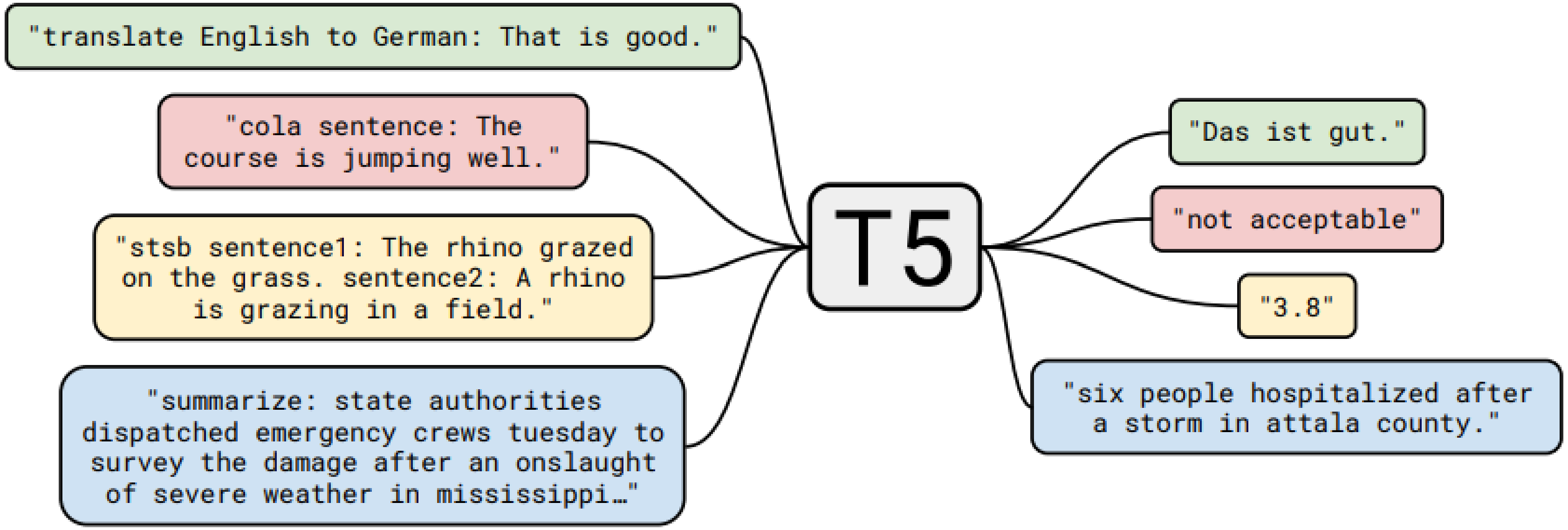


(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

Methods

CAPTION-to-STORY MODEL

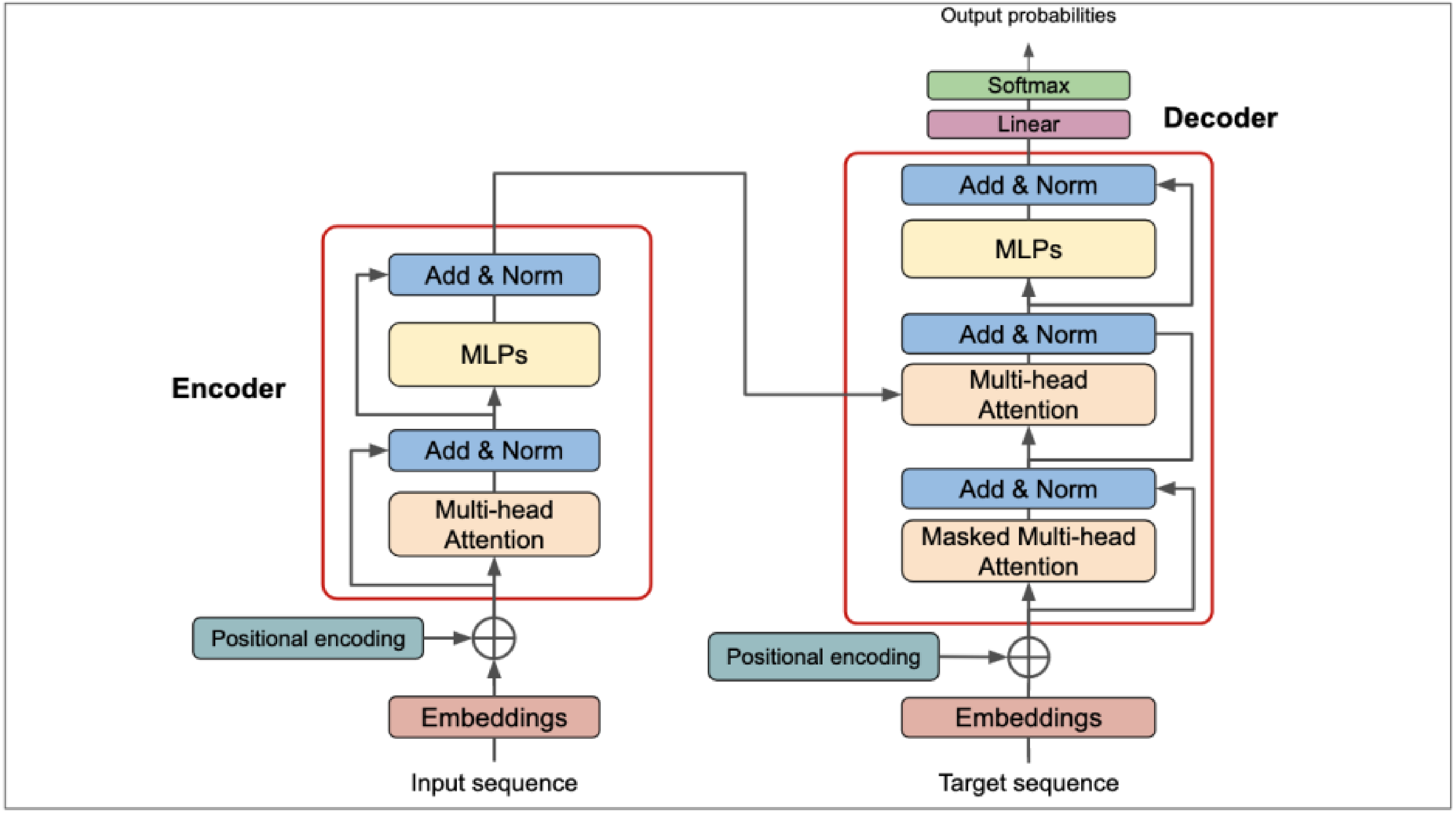
- **T5 (**Text-to-Text Transfer Transformer - *Lewis et al.***):** Follows the original transformer. Can handle a wide range of NLP *tasks by converting them into text-to-text format* and without change in its architecture.
- The conversion to text-to-text format is done by using prefixes in the original input by indicating the task.
- Pre-trained Using Document Corruption Techniques: Span Corruption, Token Deletion, Sentence Shuffling.
- Also combines BERT-like Encoders & GPT-like Decoders → Can be used for both understanding and generation tasks.


"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Methods

SELF MADE T4 TRANSFORMER

- **T4:** Our from-scratch architecture → Follows the original Encoder-Decoder transformer.
- Another **T**ext-**t**o-**T**ext **T**ransformer. (Small?) imitation of the T5 and BART models.



Utrecht University

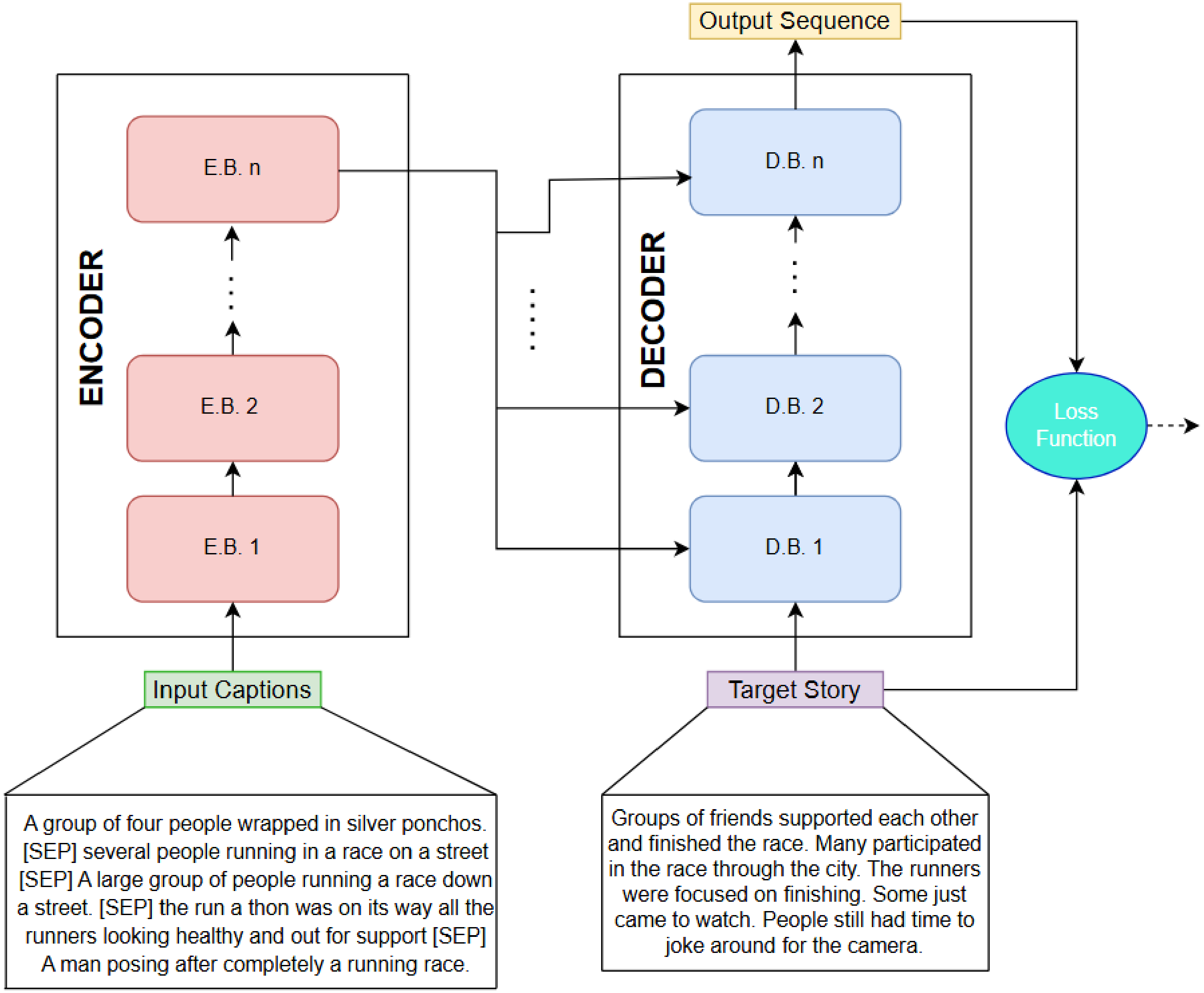
Methods

11-11-2024

HOW DO WE FINE-TUNE?

Direct Sequence-to-Sequence Training/Fine-tuning:

- Passing the input sequence (captions) and the target sequence (story) to the model at every batch iteration.
- The output sequence has size (*batch_size, |Y|, |V|*).
- The loss function is *Cross Entropy.*



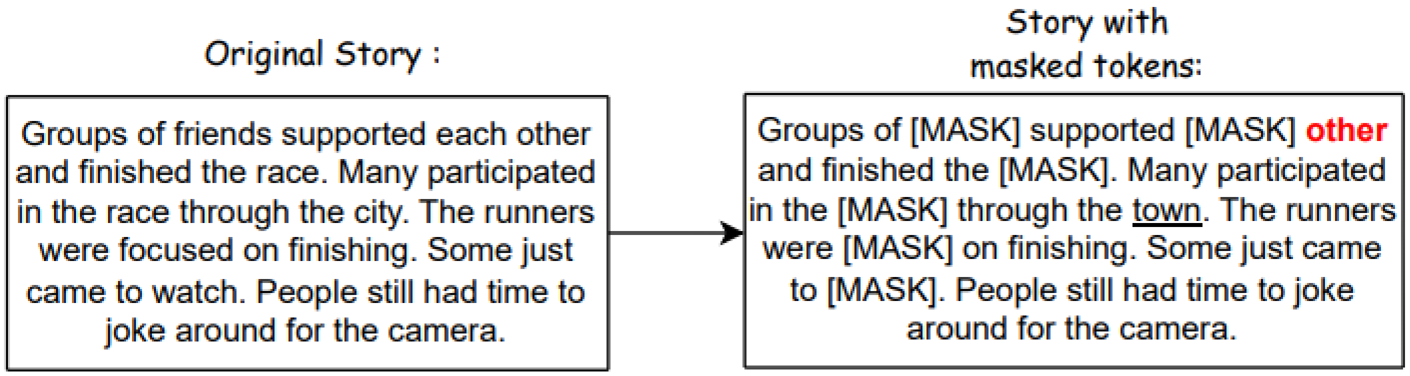
A group of four people wrapped in silver ponchos. [SEP] several people running in a race on a street [SEP] A large group of people running a race down a street. [SEP] the run a thon was on its way all the runners looking healthy and out for support [SEP] A man posing after completely a running race.

Groups of friends supported each other and finished the race. Many participated in the race through the city. The runners were focused on finishing. Some just came to watch. People still had time to joke around for the camera.

Methods

CORRUPTING OUR T4

- **MASK LANGUAGE MODELLING (MLM):**
  - A percentage of input tokens is masked at random → The model is trained to predict these tokens.
  - Most fundamental method of document corruption. Its bidirectional nature empowers the local contextual understanding of the model and increases its robustness and generalization abilities.
  - Attempt to be adapted to our Visual Storytelling setting.



Original Story :

Groups of friends supported each other and finished the race. Many participated in the race through the city. The runners were focused on finishing. Some just came to watch. People still had time to joke around for the camera.

Story with masked tokens:

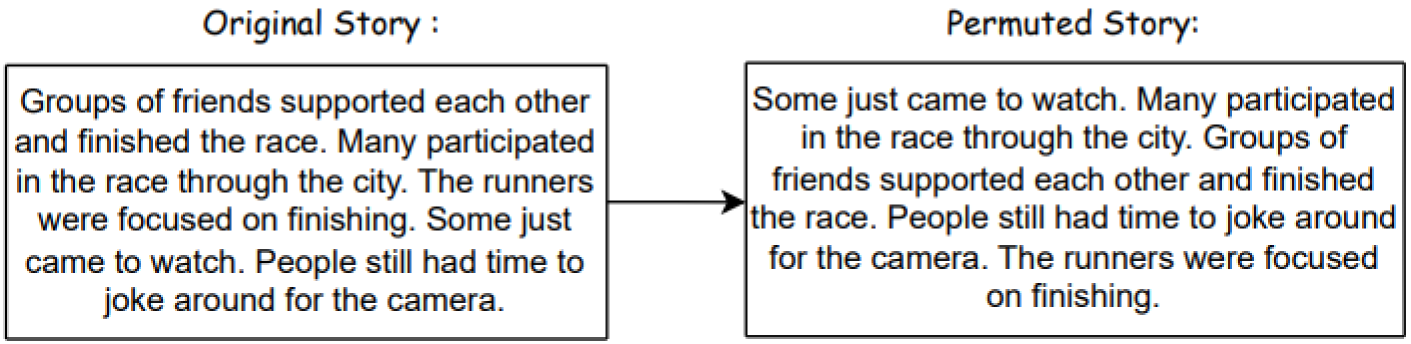
Groups of [MASK] supported [MASK] other and finished the [MASK]. Many participated in the [MASK] through the town. The runners were [MASK] on finishing. Some just came to [MASK]. People still had time to joke around for the camera.

Figure 20. *A representation of how stories are masked for the purpose of MLM. On the left, the original story. On the right, the story with 7 candidate masked tokens. In this case 5 tokens: "friends", "each", "race", "focused" and "watch" are replaced with the actual "[MASK]" token, while the red font word "other" is the one that remains unaltered. The underlined word "town" is the one that substituted the original token "city".*

Methods

CORRUPTING OUR T4

- **SENTENCE PERMUTATION (SP):**
  - Sentences within a document are shuffled randomly.
  - It also enhances the contextual understanding of the model, especially the long-range dependences that potentially can occur between two or more sentences that are further away →Increases the versatility of models making them adaptable to various NLP tasks.
  - Attempt to be adapted to our Visual Storytelling setting, as well.



Original Story :

Groups of friends supported each other and finished the race. Many participated in the race through the city. The runners were focused on finishing. Some just came to watch. People still had time to joke around for the camera.

Permuted Story:

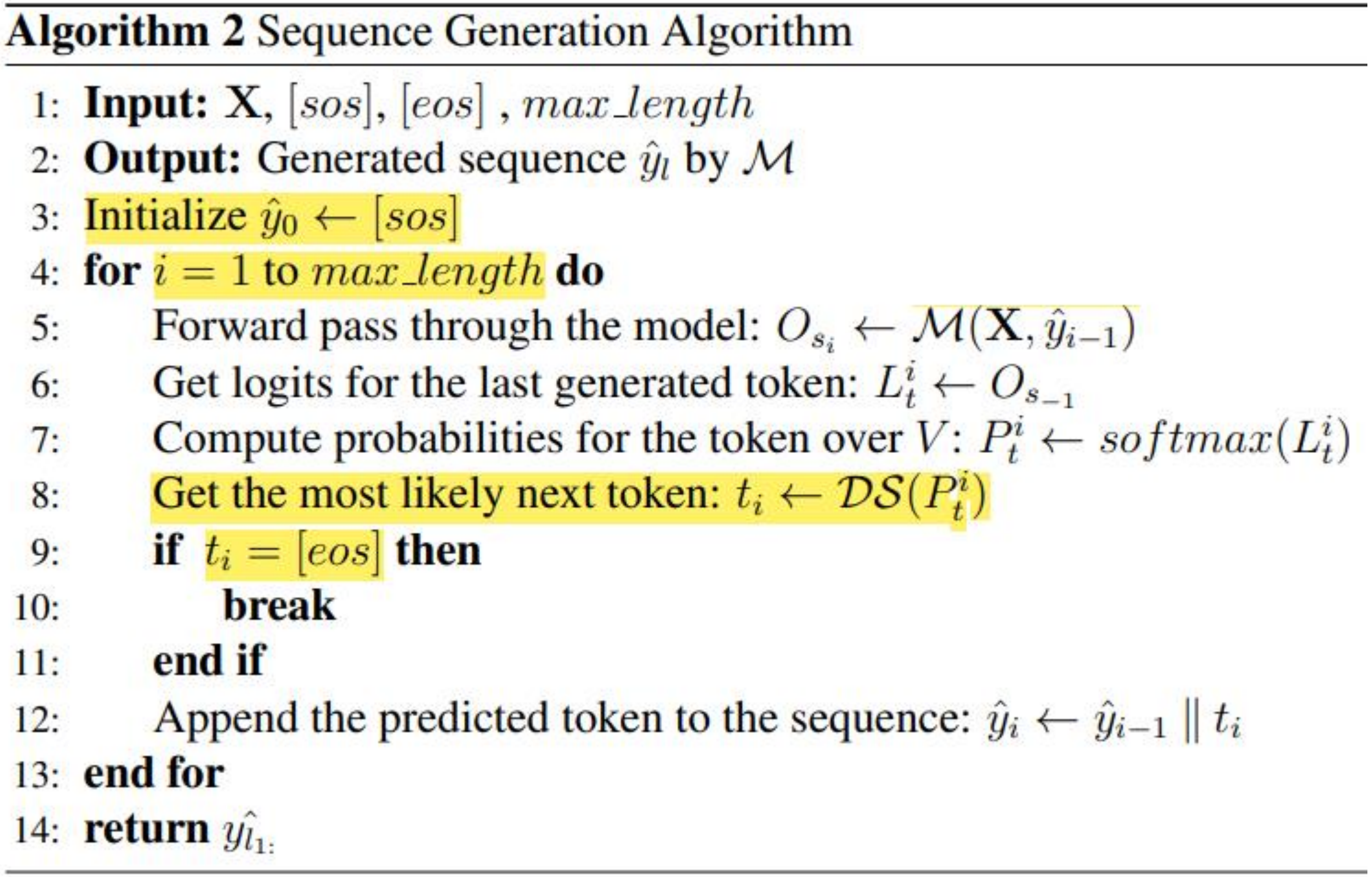
Some just came to watch. Many participated in the race through the city. Groups of friends supported each other and finished the race. People still had time to joke around for the camera. The runners were focused on finishing.

Figure 21. *A representation of how stories are altered for the purpose of Sentence Permutation technique. On the left, the original five-sentences story. On the right, the permuted story, where all five sentences have been randomly shuffled.*

Methods

IS T4 ABLE TO SPEAK?

- **Autoregressive Generation**



Algorithm 2 Sequence Generation Algorithm

```
Input: $\mathbf{X}$, $[sos]$, $[eos]$ , $max\_length$
Output: Generated sequence $\hat{y}_l$ by $\mathcal{M}$
Initialize $\hat{y}_0 \leftarrow [sos]$
for $i = 1$ to $max\_length$ do
    Forward pass through the model: $O_{s_i} \leftarrow \mathcal{M}(\mathbf{X}, \hat{y}_{i-1})$
    Get logits for the last generated token: $L_t^i \leftarrow O_{s_{-1}}$
    Compute probabilities for the token over $V$: $P_t^i \leftarrow softmax(L_t^i)$
    Get the most likely next token: $t_i \leftarrow \mathcal{DS}(P_t^i)$
    if $t_i = [eos]$ then
        break
    end if
    Append the predicted token to the sequence: $\hat{y}_i \leftarrow \hat{y}_{i-1} \mathbin{||} t_i$
end for
return $\hat{y_{l_{1:}}}$
```

- **Decoding Strategies**

1. *Greedy Search*
2. *Multinomial Sampling*
3. *Nucleus or P-sampling*
4. *Beam Search*

Methods

EVALUATION PROCEDURES

→ **Automatic Evaluation Metrics:**

- *BLEU*: Measures co-occurrences of exact n-grams between the candidate and reference sentences.
- *METEOR:* Find similarity between a candidate sentence and reference sentences, including exact matches, synonyms, stemmed words etc.
- *ROUGE_L:* Computes the F-measure (Precision & Recall) based on the Longest Common Sub-sequence between the candidate and reference sentences.
- *CIDEr:* Measures consensus between candidate and reference sentences by performing a TF-IDF weighting for each n-gram.
- *SPICE:* A *Dependency Parser* creates a scene graph for the words for both the candidate and reference sentences establishing also semantic relations between the words.
- *SPIDER: A Linear combination of SPICE & CIDEr.*

→ **Human Evaluation:**

- *According to Criteria*: 1)Relevance with the Input Images, 2) Coherence & Flow, 3) Narrative Depth, 4) Imagination & Creativity, 5) Engagement & Interest, 6) Language & Style.
- *Guessing:* Guess among the five-given stories (4 machine, 1 human) which is the human written.

→ **Artificial Evaluation with an LLM:**

- Recruiting GPT-4o to judge the transition between the generated sentences in terms of coherence, temporal continuity and logical flow.

Experimental Set-up

AN ARMY OF MODELS

- Following the sub-research questions → Separate the experiments in **3 Phases**
- **Phase 1** → Train/Fine-tune and test different variants of Clip-Cap on VIST images → *Only Automatic Evaluation.*
- **Phase 2** → Train/Fine-tune and test different variants of text-to-text LMs (T4, T5, BART) on VIST DII captions → *Only Automatic Evaluation.*
- **Ultimate Phase** → Use the whole framework with the best fine-tuned models from the previous 2 phases for Visual Storytelling, only for inference time (no training/fine-tuning) → All types of evaluation processes.

Phase 1:

Three families of models (Clip-Cap variants):

- Zero-shot models from MSCOCO (4 in total).
- Models taken from MSCOCO & fine-tuned on VIST (8 in total).
- From-scratch models trained solely on VIST. (16 in total)

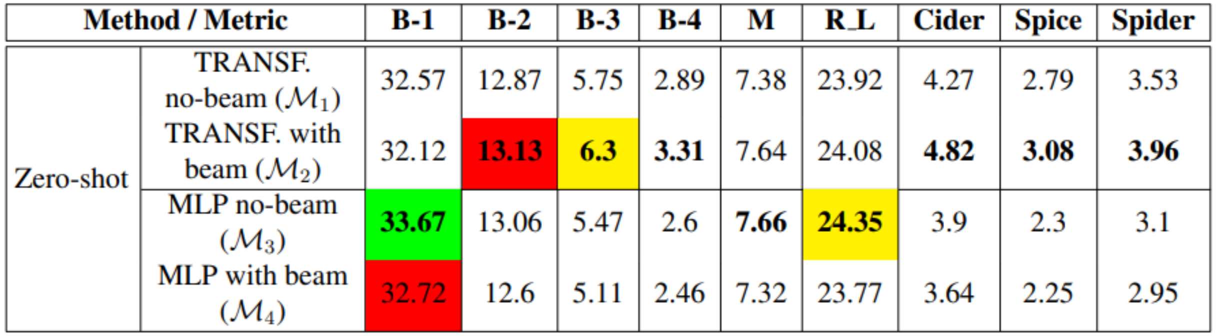
Phase 2:

Four families of models (but three types):

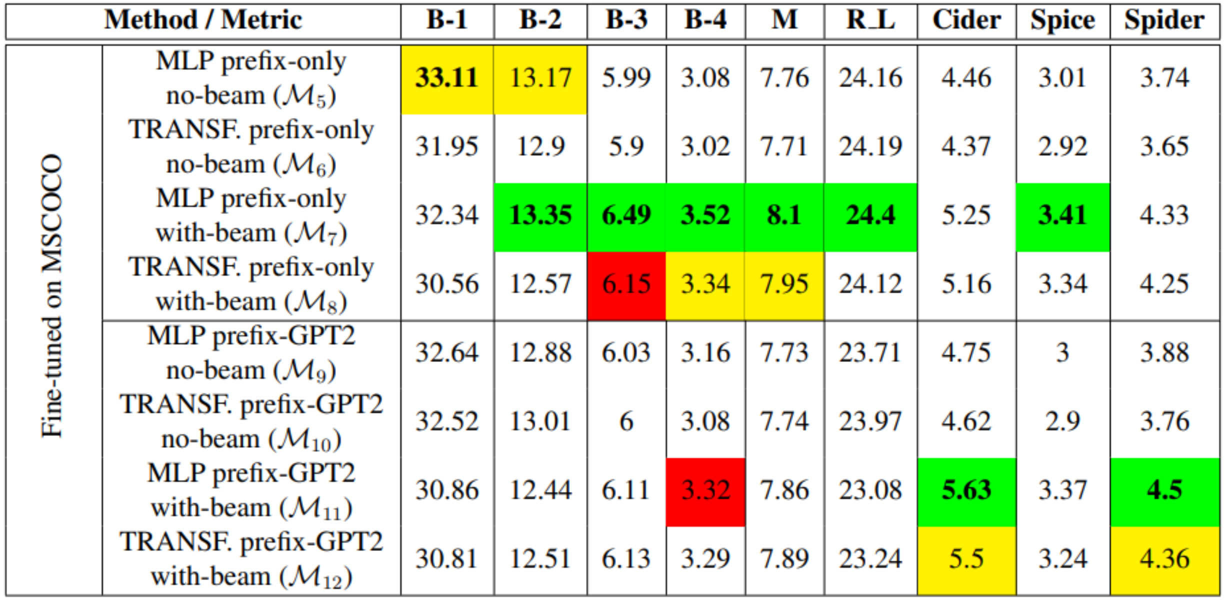
- T4 only trained with sequence-to-sequence on VIST/DII-SIS(4 in total).
- T4 pre-trained with MLM & SP and then seq-to-seq on VIST/DII-SIS (4 in total).
- T5 fine-tuned (seq-to-seq) on VIST/DII-SIS (3 in total).
- BART fine-tuned (seq-to-seq) on VIST/DII-SIS . (3 in total)

Results

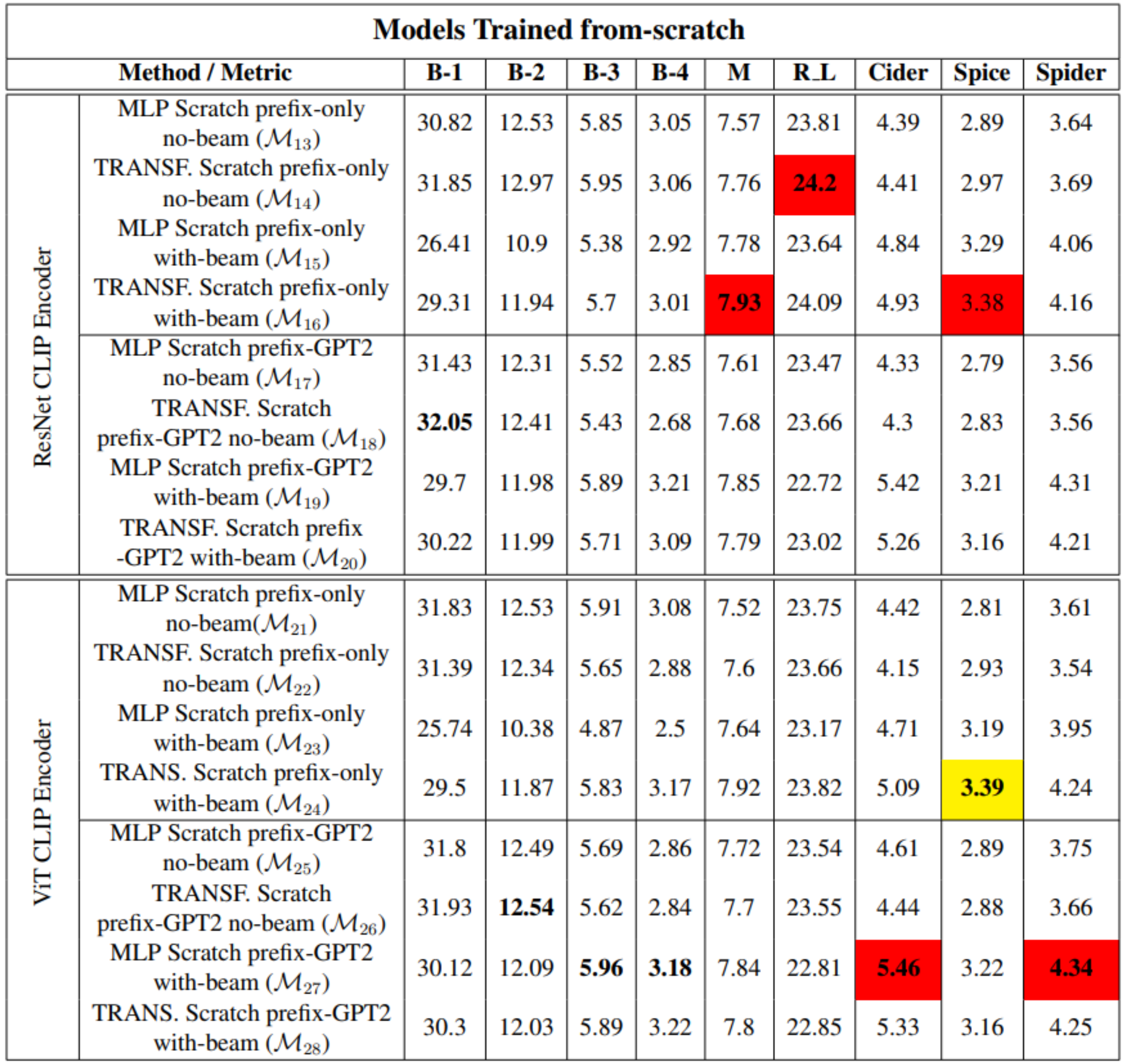
PHASE 1



Method / Metric		B-1	B-2	B-3	B-4	M	R_L	Cider	Spice	Spider
Zero-shot	TRANSF. no-beam ($\mathcal{M}_1$)	32.57	12.87	5.75	2.89	7.38	23.92	4.27	2.79	3.53
	TRANSF. with beam ($\mathcal{M}_2$)	32.12	**13.13**	**6.3**	**3.31**	7.64	24.08	**4.82**	**3.08**	**3.96**
	MLP no-beam ($\mathcal{M}_3$)	**33.67**	13.06	5.47	2.6	**7.66**	**24.35**	3.9	2.3	3.1
	MLP with beam ($\mathcal{M}_4$)	32.72	12.6	5.11	2.46	7.32	23.77	3.64	2.25	2.95



Method / Metric		B-1	B-2	B-3	B-4	M	R_L	Cider	Spice	Spider
Fine-tuned on MSCOCO	MLP prefix-only no-beam ($\mathcal{M}_5$)	**33.11**	**13.17**	5.99	3.08	7.76	24.16	4.46	3.01	3.74
	TRANSF. prefix-only no-beam ($\mathcal{M}_6$)	31.95	12.9	5.9	3.02	7.71	24.19	4.37	2.92	3.65
	MLP prefix-only with-beam ($\mathcal{M}_7$)	32.34	**13.35**	**6.49**	**3.52**	**8.1**	**24.4**	5.25	**3.41**	4.33
	TRANSF. prefix-only with-beam ($\mathcal{M}_8$)	30.56	12.57	6.15	3.34	7.95	24.12	5.16	3.34	4.25
	MLP prefix-GPT2 no-beam ($\mathcal{M}_9$)	32.64	12.88	6.03	3.16	7.73	23.71	4.75	3	3.88
	TRANSF. prefix-GPT2 no-beam ($\mathcal{M}_{10}$)	32.52	13.01	6	3.08	7.74	23.97	4.62	2.9	3.76
	MLP prefix-GPT2 with-beam ($\mathcal{M}_{11}$)	30.86	12.44	6.11	3.32	7.86	23.08	**5.63**	3.37	**4.5**
	TRANSF. prefix-GPT2 with-beam ($\mathcal{M}_{12}$)	30.81	12.51	6.13	3.29	7.89	23.24	5.5	3.24	4.36



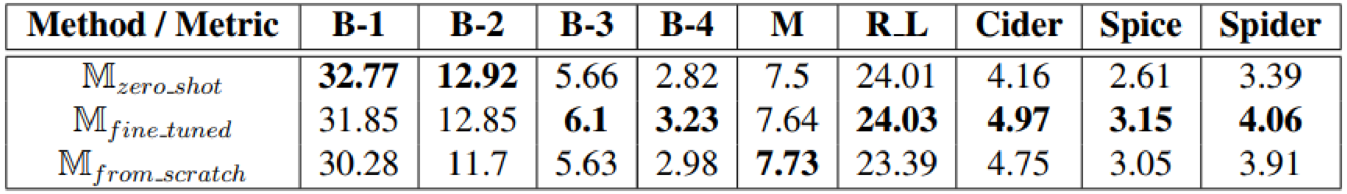
Models Trained from-scratch

Method / Metric		B-1	B-2	B-3	B-4	M	R_L	Cider	Spice	Spider
ResNet CLIP Encoder	MLP Scratch prefix-only no-beam ($\mathcal{M}_{13}$)	30.82	12.53	5.85	3.05	7.57	23.81	4.39	2.89	3.64
	TRANSF. Scratch prefix-only no-beam ($\mathcal{M}_{14}$)	31.85	12.97	5.95	3.06	7.76	**24.2**	4.41	2.97	3.69
	MLP Scratch prefix-only with-beam ($\mathcal{M}_{15}$)	26.41	10.9	5.38	2.92	7.78	23.64	4.84	3.29	4.06
	TRANSF. Scratch prefix-only with-beam ($\mathcal{M}_{16}$)	29.31	11.94	5.7	3.01	**7.93**	24.09	4.93	3.38	4.16
	MLP Scratch prefix-GPT2 no-beam ($\mathcal{M}_{17}$)	31.43	12.31	5.52	2.85	7.61	23.47	4.33	2.79	3.56
	TRANSF. Scratch prefix-GPT2 no-beam ($\mathcal{M}_{18}$)	**32.05**	12.41	5.43	2.68	7.68	23.66	4.3	2.83	3.56
	MLP Scratch prefix-GPT2 with-beam ($\mathcal{M}_{19}$)	29.7	11.98	5.89	3.21	7.85	22.72	5.42	3.21	4.31
	TRANSF. Scratch prefix -GPT2 with-beam ($\mathcal{M}_{20}$)	30.22	11.99	5.71	3.09	7.79	23.02	5.26	3.16	4.21
ViT CLIP Encoder	MLP Scratch prefix-only no-beam($\mathcal{M}_{21}$)	31.83	12.53	5.91	3.08	7.52	23.75	4.42	2.81	3.61
	TRANSF. Scratch prefix-only no-beam ($\mathcal{M}_{22}$)	31.39	12.34	5.65	2.88	7.6	23.66	4.15	2.93	3.54
	MLP Scratch prefix-only with-beam ($\mathcal{M}_{23}$)	25.74	10.38	4.87	2.5	7.64	23.17	4.71	3.19	3.95
	TRANS. Scratch prefix-only with-beam ($\mathcal{M}_{24}$)	29.5	11.87	5.83	3.17	7.92	23.82	5.09	**3.39**	4.24
	MLP Scratch prefix-GPT2 no-beam ($\mathcal{M}_{25}$)	31.8	12.49	5.69	2.86	7.72	23.54	4.61	2.89	3.75
	TRANSF. Scratch prefix-GPT2 no-beam ($\mathcal{M}_{26}$)	31.93	**12.54**	5.62	2.84	7.7	23.55	4.44	2.88	3.66
	MLP Scratch prefix-GPT2 with-beam ($\mathcal{M}_{27}$)	30.12	12.09	**5.96**	**3.18**	7.84	22.81	**5.46**	3.22	4.34
	TRANS. Scratch prefix-GPT2 with-beam ($\mathcal{M}_{28}$)	30.3	12.03	5.89	3.22	7.8	22.85	5.33	3.16	4.25

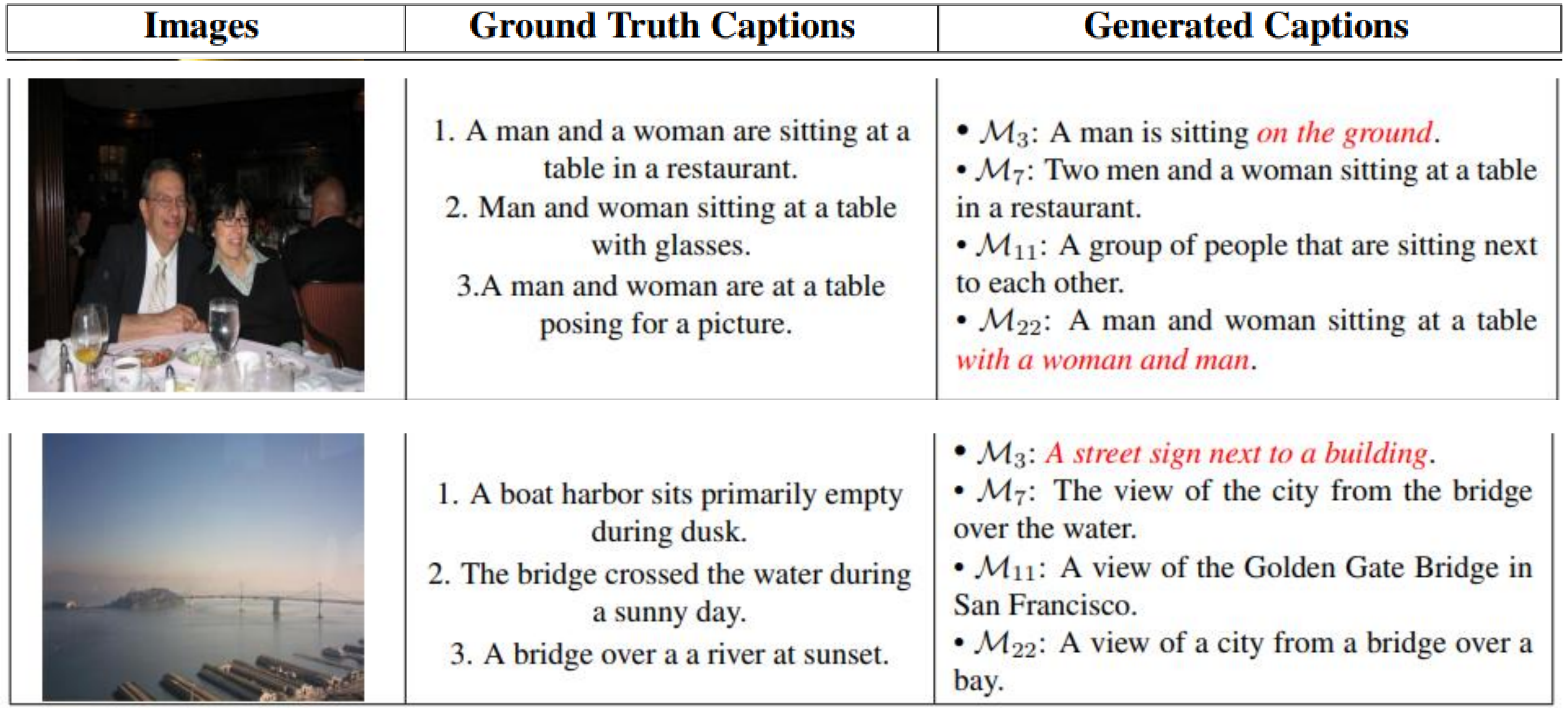
Results

PHASE 1

- Fine tuned models perform better on most metrics which essentially makes sense since they are doubly trained.



Method / Metric	B-1	B-2	B-3	B-4	M	R_L	Cider	Spice	Spider
$\mathbb{M}_{zero\_shot}$	**32.77**	**12.92**	5.66	2.82	7.5	24.01	4.16	2.61	3.39
$\mathbb{M}_{fine\_tuned}$	31.85	12.85	**6.1**	**3.23**	7.64	**24.03**	**4.97**	**3.15**	**4.06**
$\mathbb{M}_{from\_scratch}$	30.28	11.7	5.63	2.98	**7.73**	23.39	4.75	3.05	3.91

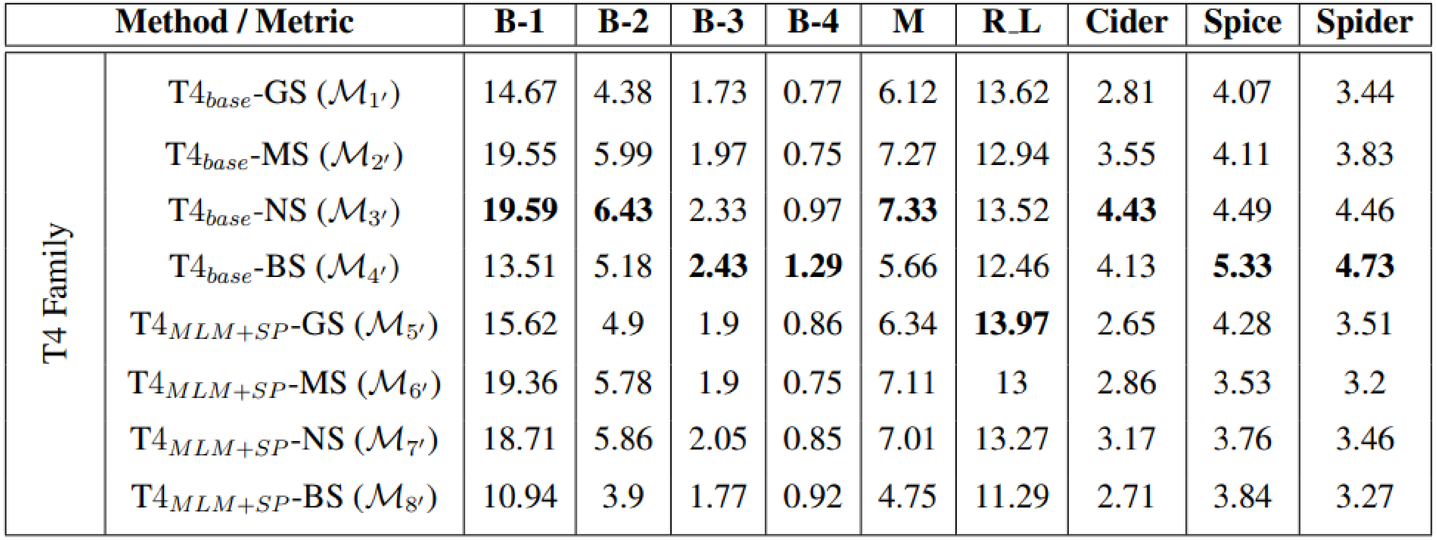


Images	Ground Truth Captions	Generated Captions
	1. A man and a woman are sitting at a table in a restaurant. 2. Man and woman sitting at a table with glasses. 3.A man and woman are at a table posing for a picture.	• $\mathcal{M}_3$: A man is sitting *on the ground*. • $\mathcal{M}_7$: Two men and a woman sitting at a table in a restaurant. • $\mathcal{M}_{11}$: A group of people that are sitting next to each other. • $\mathcal{M}_{22}$: A man and woman sitting at a table *with a woman and man*.
	1. A boat harbor sits primarily empty during dusk. 2. The bridge crossed the water during a sunny day. 3. A bridge over a a river at sunset.	• $\mathcal{M}_3$: *A street sign next to a building*. • $\mathcal{M}_7$: The view of the city from the bridge over the water. • $\mathcal{M}_{11}$: A view of the Golden Gate Bridge in San Francisco. • $\mathcal{M}_{22}$: A view of a city from a bridge over a bay.

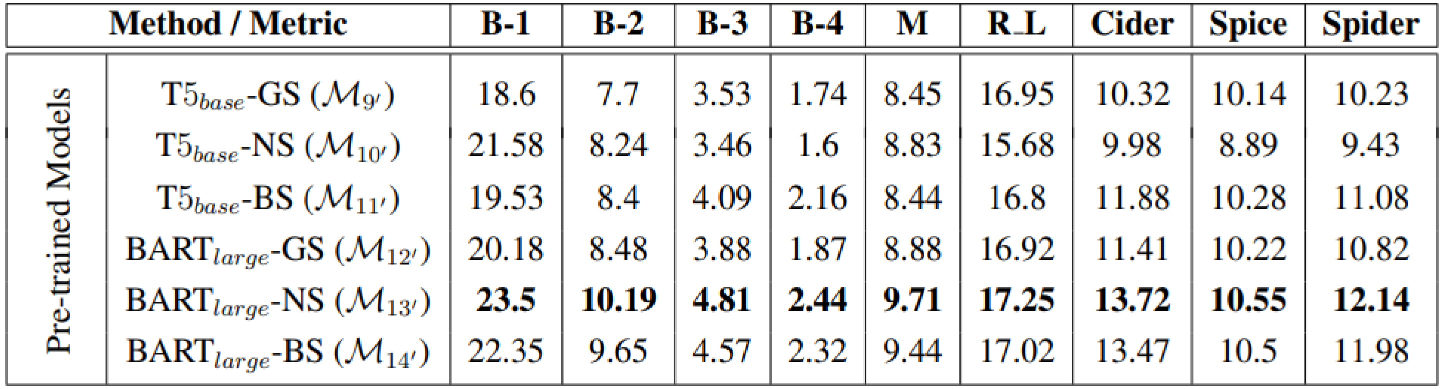
Results

PHASE 2

- BART models seem to perform better according to automatic metrics, but BART is … large.



	Method / Metric	B-1	B-2	B-3	B-4	M	R_L	Cider	Spice	Spider
T4 Family	T4$_{base}$-GS ($\mathcal{M}_{1'}$)	14.67	4.38	1.73	0.77	6.12	13.62	2.81	4.07	3.44
	T4$_{base}$-MS ($\mathcal{M}_{2'}$)	19.55	5.99	1.97	0.75	7.27	12.94	3.55	4.11	3.83
	T4$_{base}$-NS ($\mathcal{M}_{3'}$)	**19.59**	**6.43**	2.33	0.97	**7.33**	13.52	**4.43**	4.49	4.46
	T4$_{base}$-BS ($\mathcal{M}_{4'}$)	13.51	5.18	**2.43**	**1.29**	5.66	12.46	4.13	**5.33**	**4.73**
	T4$_{MLM+SP}$-GS ($\mathcal{M}_{5'}$)	15.62	4.9	1.9	0.86	6.34	**13.97**	2.65	4.28	3.51
	T4$_{MLM+SP}$-MS ($\mathcal{M}_{6'}$)	19.36	5.78	1.9	0.75	7.11	13	2.86	3.53	3.2
	T4$_{MLM+SP}$-NS ($\mathcal{M}_{7'}$)	18.71	5.86	2.05	0.85	7.01	13.27	3.17	3.76	3.46
	T4$_{MLM+SP}$-BS ($\mathcal{M}_{8'}$)	10.94	3.9	1.77	0.92	4.75	11.29	2.71	3.84	3.27

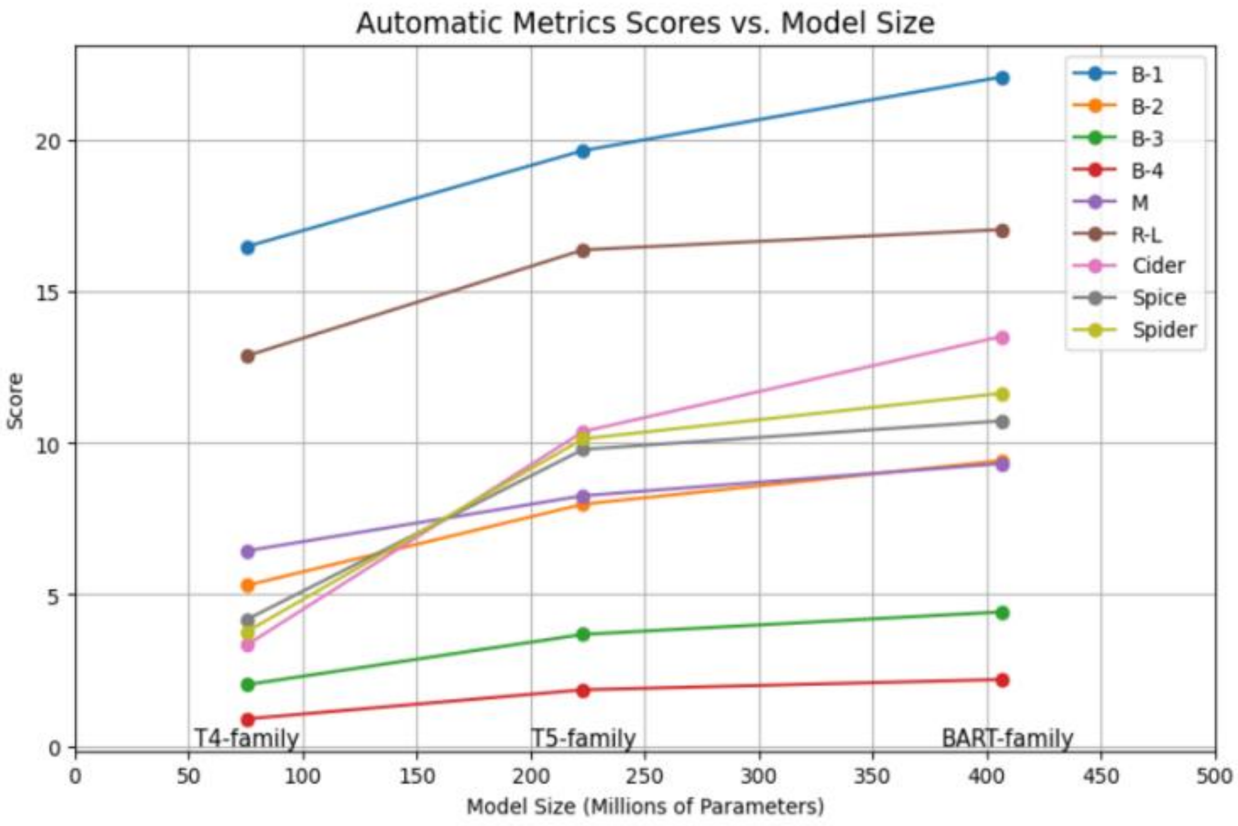


	Method / Metric	B-1	B-2	B-3	B-4	M	R_L	Cider	Spice	Spider
Pre-trained Models	T5$_{base}$-GS ($\mathcal{M}_{9'}$)	18.6	7.7	3.53	1.74	8.45	16.95	10.32	10.14	10.23
	T5$_{base}$-NS ($\mathcal{M}_{10'}$)	21.58	8.24	3.46	1.6	8.83	15.68	9.98	8.89	9.43
	T5$_{base}$-BS ($\mathcal{M}_{11'}$)	19.53	8.4	4.09	2.16	8.44	16.8	11.88	10.28	11.08
	BART$_{large}$-GS ($\mathcal{M}_{12'}$)	20.18	8.48	3.88	1.87	8.88	16.92	11.41	10.22	10.82
	BART$_{large}$-NS ($\mathcal{M}_{13'}$)	**23.5**	**10.19**	**4.81**	**2.44**	**9.71**	**17.25**	**13.72**	**10.55**	**12.14**
	BART$_{large}$-BS ($\mathcal{M}_{14'}$)	22.35	9.65	4.57	2.32	9.44	17.02	13.47	10.5	11.98

Results

PHASE 2

- Average results for all metrics per family of LMs.
- As a result, the efficiency on automatic metrics is proportional to the size (in terms of trainable params) of the LM.



Results

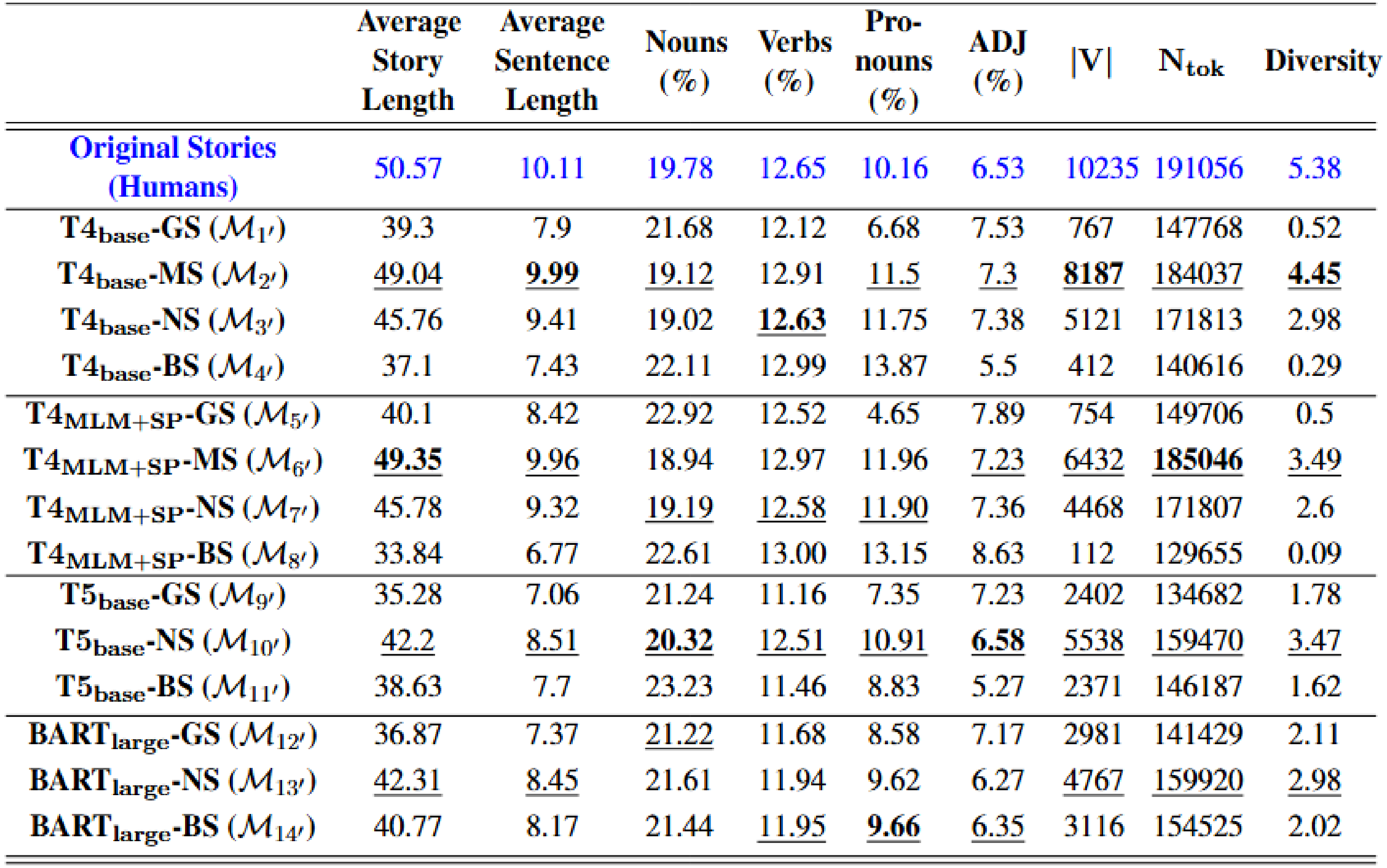
DID WE ACTUALLY LOSE?

- According to many linguistically related metrics T4 models surpass T5 and BART (sometimes even on the same strategy) → Indication for more diverse sentences (stories), although depends on the strategy as well.

- We propose the metric/tool of *Ideality.*
- *Ideality* essentially measures how far are certain (lexical) characteristics of our machine generated stories, from the respective traits of some golden (human written) stories.

$$ideality = \sum_{ms,hs \in \mathbb{S}} \frac{1}{\sigma(|ms - hs|)},$$

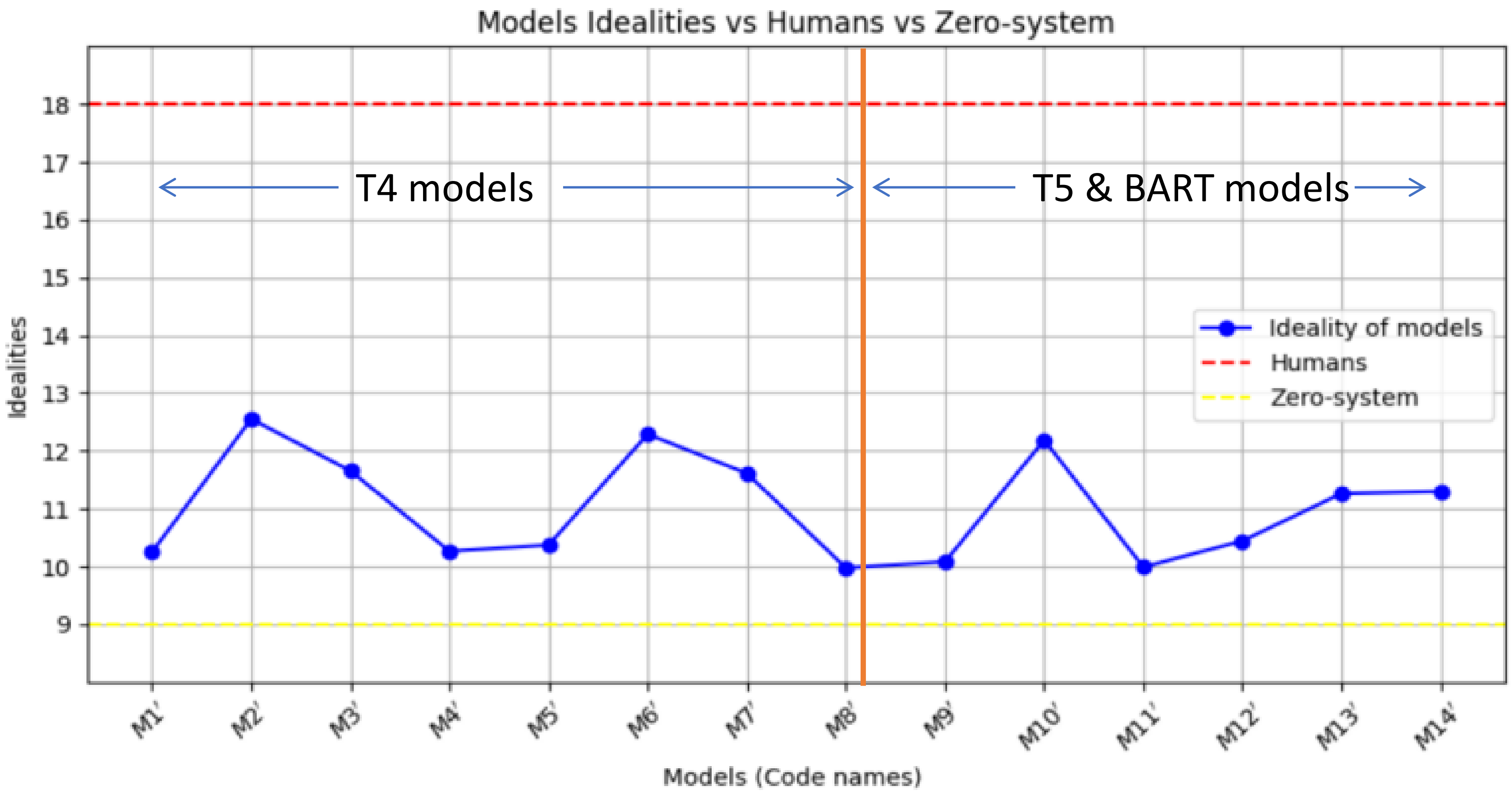
$$ideality = I(ms) : \mathbb{R} \rightarrow (n, 2n]$$



	Average Story Length	Average Sentence Length	Nouns (%)	Verbs (%)	Pro-nouns (%)	ADJ (%)	|V|	N_{tok}	Diversity
Original Stories (Humans)	50.57	10.11	19.78	12.65	10.16	6.53	10235	191056	5.38
$\mathbf{T4_{base}}$-GS ($\mathcal{M}_{1'}$)	39.3	7.9	21.68	12.12	6.68	7.53	767	147768	0.52
$\mathbf{T4_{base}}$-MS ($\mathcal{M}_{2'}$)	49.04	**9.99**	19.12	12.91	11.5	7.3	**8187**	184037	**4.45**
$\mathbf{T4_{base}}$-NS ($\mathcal{M}_{3'}$)	45.76	9.41	19.02	**12.63**	11.75	7.38	5121	171813	2.98
$\mathbf{T4_{base}}$-BS ($\mathcal{M}_{4'}$)	37.1	7.43	22.11	12.99	13.87	5.5	412	140616	0.29
$\mathbf{T4_{MLM+SP}}$-GS ($\mathcal{M}_{5'}$)	40.1	8.42	22.92	12.52	4.65	7.89	754	149706	0.5
$\mathbf{T4_{MLM+SP}}$-MS ($\mathcal{M}_{6'}$)	**49.35**	9.96	18.94	12.97	11.96	7.23	6432	**185046**	3.49
$\mathbf{T4_{MLM+SP}}$-NS ($\mathcal{M}_{7'}$)	45.78	9.32	19.19	12.58	11.90	7.36	4468	171807	2.6
$\mathbf{T4_{MLM+SP}}$-BS ($\mathcal{M}_{8'}$)	33.84	6.77	22.61	13.00	13.15	8.63	112	129655	0.09
$\mathbf{T5_{base}}$-GS ($\mathcal{M}_{9'}$)	35.28	7.06	21.24	11.16	7.35	7.23	2402	134682	1.78
$\mathbf{T5_{base}}$-NS ($\mathcal{M}_{10'}$)	42.2	8.51	**20.32**	12.51	10.91	**6.58**	5538	159470	3.47
$\mathbf{T5_{base}}$-BS ($\mathcal{M}_{11'}$)	38.63	7.7	23.23	11.46	8.83	5.27	2371	146187	1.62
$\mathbf{BART_{large}}$-GS ($\mathcal{M}_{12'}$)	36.87	7.37	21.22	11.68	8.58	7.17	2981	141429	2.11
$\mathbf{BART_{large}}$-NS ($\mathcal{M}_{13'}$)	42.31	8.45	21.61	11.94	9.62	6.27	4767	159920	2.98
$\mathbf{BART_{large}}$-BS ($\mathcal{M}_{14'}$)	40.77	8.17	21.44	11.95	**9.66**	6.35	3116	154525	2.02

Results

DID WE ACTUALLY LOSE?

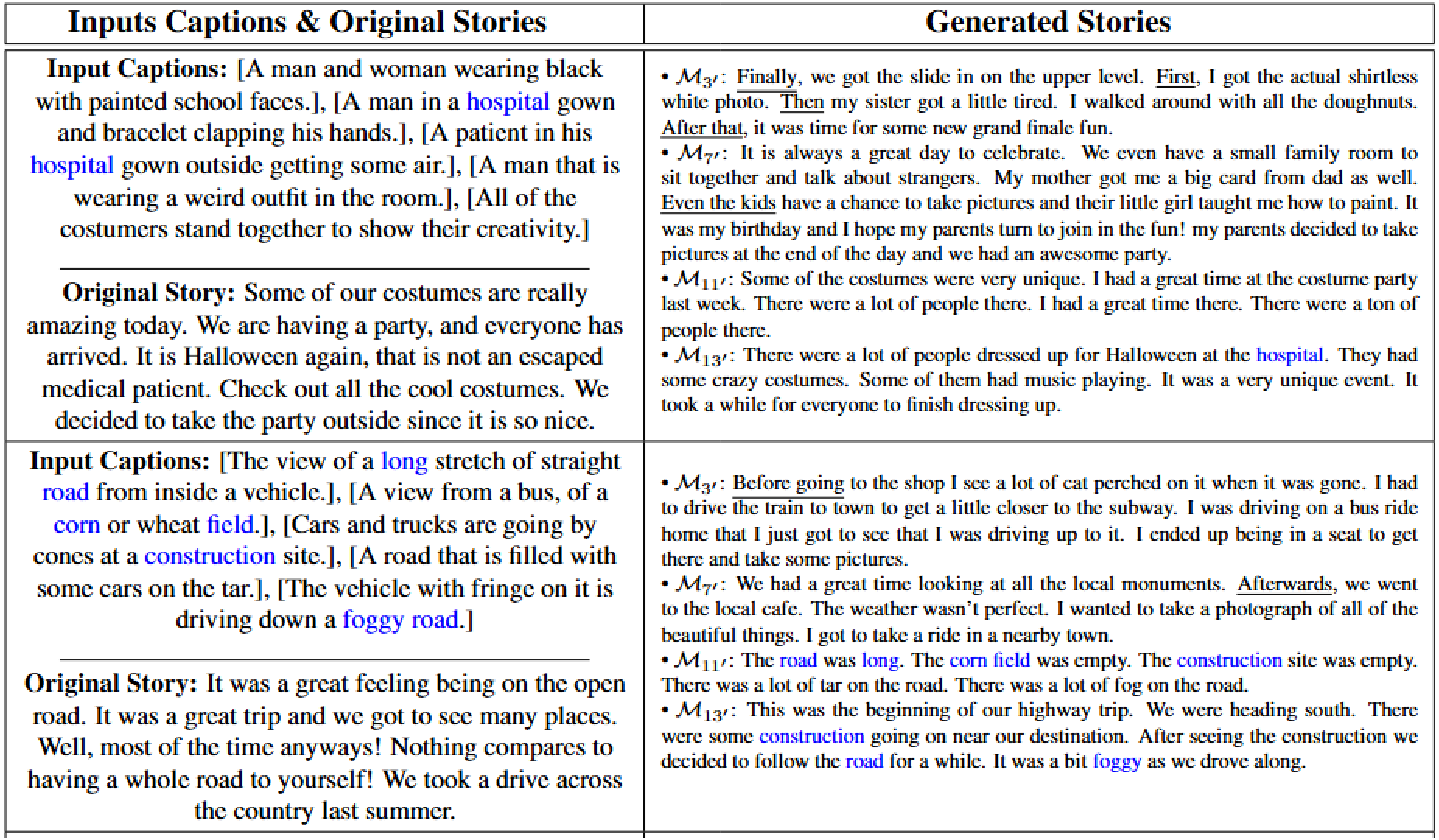


T4 models

T5 & BART models

Results

EXAMPLES PHASE 2

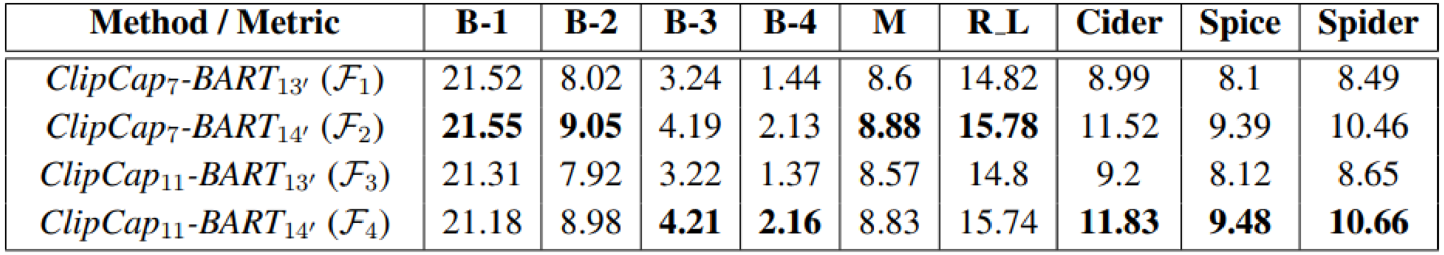


Inputs Captions & Original Stories	Generated Stories
Input Captions: [A man and woman wearing black with painted school faces.], [A man in a hospital gown and bracelet clapping his hands.], [A patient in his hospital gown outside getting some air.], [A man that is wearing a weird outfit in the room.], [All of the costumers stand together to show their creativity.] **Original Story:** Some of our costumes are really amazing today. We are having a party, and everyone has arrived. It is Halloween again, that is not an escaped medical patient. Check out all the cool costumes. We decided to take the party outside since it is so nice.	• $\mathcal{M}_{3'}$: Finally, we got the slide in on the upper level. First, I got the actual shirtless white photo. Then my sister got a little tired. I walked around with all the doughnuts. After that, it was time for some new grand finale fun. • $\mathcal{M}_{7'}$: It is always a great day to celebrate. We even have a small family room to sit together and talk about strangers. My mother got me a big card from dad as well. Even the kids have a chance to take pictures and their little girl taught me how to paint. It was my birthday and I hope my parents turn to join in the fun! my parents decided to take pictures at the end of the day and we had an awesome party. • $\mathcal{M}_{11'}$: Some of the costumes were very unique. I had a great time at the costume party last week. There were a lot of people there. I had a great time there. There were a ton of people there. • $\mathcal{M}_{13'}$: There were a lot of people dressed up for Halloween at the hospital. They had some crazy costumes. Some of them had music playing. It was a very unique event. It took a while for everyone to finish dressing up.
Input Captions: [The view of a long stretch of straight road from inside a vehicle.], [A view from a bus, of a corn or wheat field.], [Cars and trucks are going by cones at a construction site.], [A road that is filled with some cars on the tar.], [The vehicle with fringe on it is driving down a foggy road.] **Original Story:** It was a great feeling being on the open road. It was a great trip and we got to see many places. Well, most of the time anyways! Nothing compares to having a whole road to yourself! We took a drive across the country last summer.	• $\mathcal{M}_{3'}$: Before going to the shop I see a lot of cat perched on it when it was gone. I had to drive the train to town to get a little closer to the subway. I was driving on a bus ride home that I just got to see that I was driving up to it. I ended up being in a seat to get there and take some pictures. • $\mathcal{M}_{7'}$: We had a great time looking at all the local monuments. Afterwards, we went to the local cafe. The weather wasn't perfect. I wanted to take a photograph of all the beautiful things. I got to take a ride in a nearby town. • $\mathcal{M}_{11'}$: The road was long. The corn field was empty. The construction site was empty. There was a lot of tar on the road. There was a lot of fog on the road. • $\mathcal{M}_{13'}$: This was the beginning of our highway trip. We were heading south. There were some construction going on near our destination. After seeing the construction we decided to follow the road for a while. It was a bit foggy as we drove along.

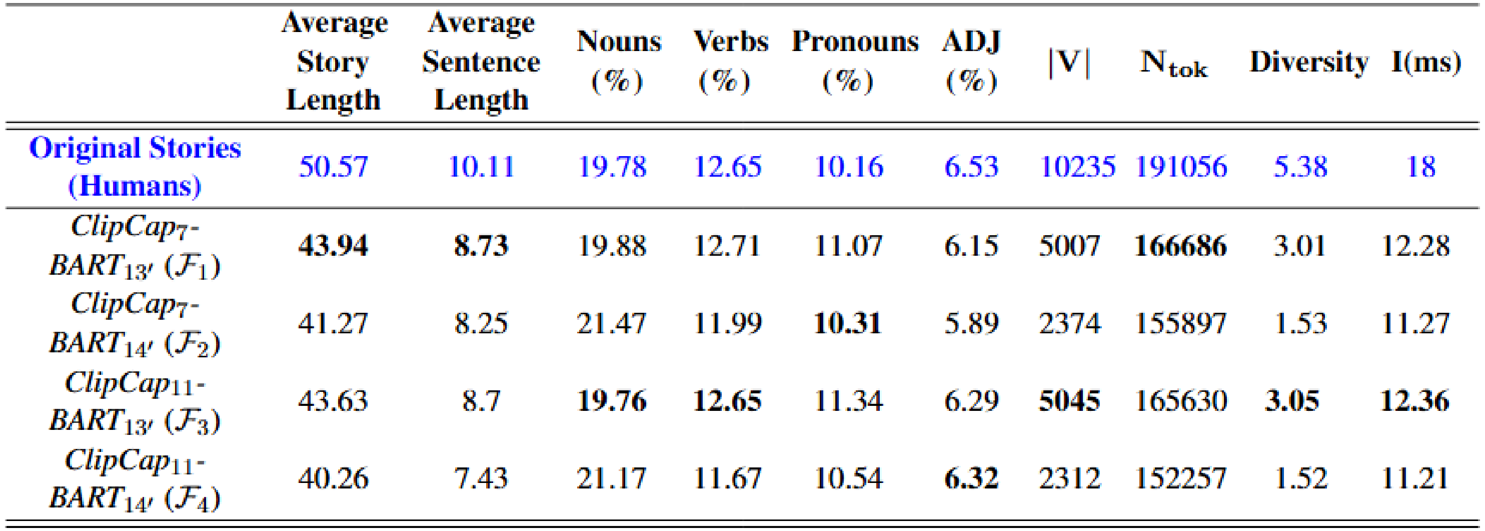
Results

ULTIMATE PHASE

- Two best captioners, two best storytellers → 4 combinations.
- Frameworks F2 & F4 with BART 14’ (beam search) > Frameworks F1 & F3 with BART 13’ (p-sampling).
- However, BART 13’ models accomplish higher scores of ideality, diversity → Greater variety of words and longer sentences are generated by these models → Indication for richer language.



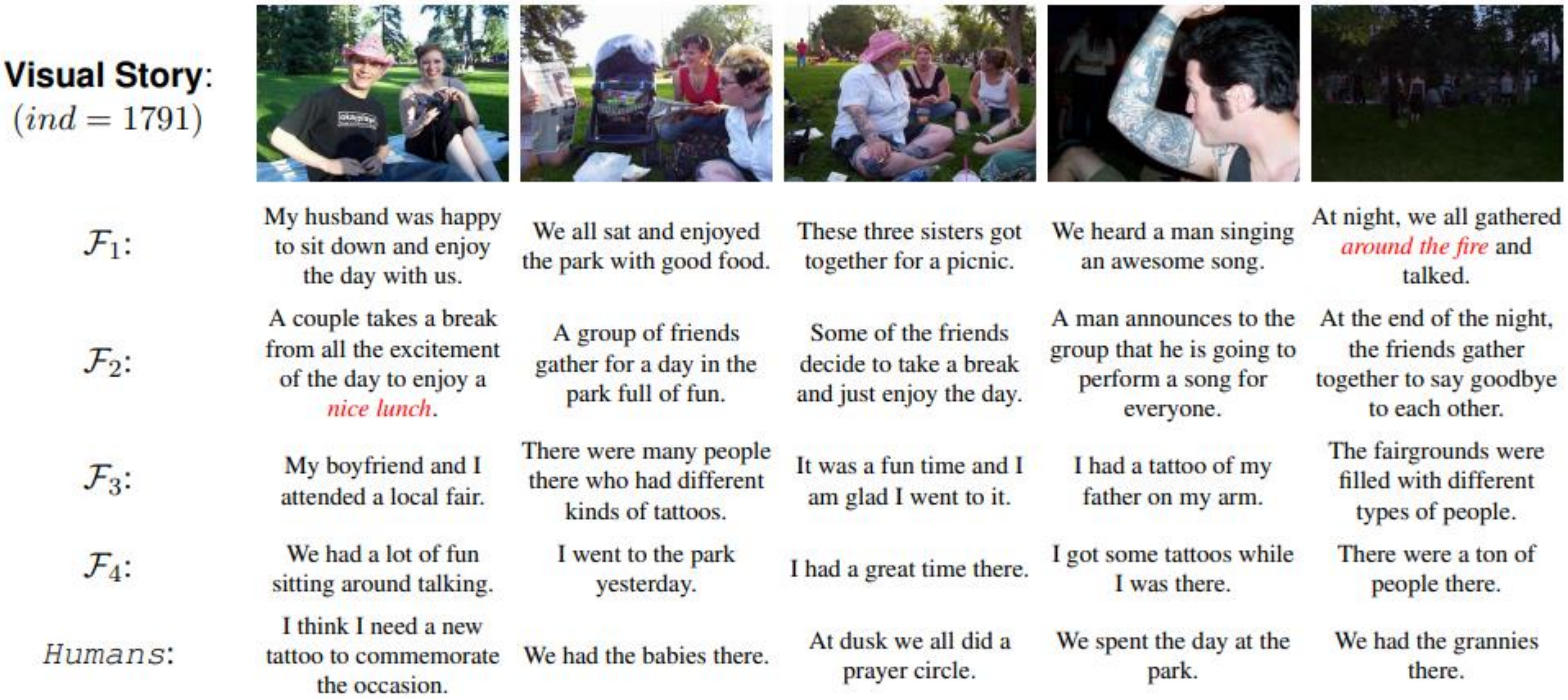
Method / Metric	B-1	B-2	B-3	B-4	M	R_L	Cider	Spice	Spider
ClipCap$_7$-*BART*$_{13'}$ ($\mathcal{F}_1$)	21.52	8.02	3.24	1.44	8.6	14.82	8.99	8.1	8.49
ClipCap$_7$-*BART*$_{14'}$ ($\mathcal{F}_2$)	**21.55**	**9.05**	4.19	2.13	**8.88**	**15.78**	11.52	9.39	10.46
ClipCap$_{11}$-*BART*$_{13'}$ ($\mathcal{F}_3$)	21.31	7.92	3.22	1.37	8.57	14.8	9.2	8.12	8.65
ClipCap$_{11}$-*BART*$_{14'}$ ($\mathcal{F}_4$)	21.18	8.98	**4.21**	**2.16**	8.83	15.74	**11.83**	**9.48**	**10.66**



	Average Story Length	Average Sentence Length	Nouns (%)	Verbs (%)	Pronouns (%)	ADJ (%)	|V|	N$_{tok}$	Diversity	I(ms)
Original Stories (Humans)	50.57	10.11	19.78	12.65	10.16	6.53	10235	191056	5.38	18
ClipCap$_7$-*BART*$_{13'}$ ($\mathcal{F}_1$)	**43.94**	**8.73**	19.88	12.71	11.07	6.15	5007	**166686**	3.01	12.28
ClipCap$_7$-*BART*$_{14'}$ ($\mathcal{F}_2$)	41.27	8.25	21.47	11.99	**10.31**	5.89	2374	155897	1.53	11.27
ClipCap$_{11}$-*BART*$_{13'}$ ($\mathcal{F}_3$)	43.63	8.7	**19.76**	**12.65**	11.34	6.29	**5045**	165630	**3.05**	**12.36**
ClipCap$_{11}$-*BART*$_{14'}$ ($\mathcal{F}_4$)	40.26	7.43	21.17	11.67	10.54	**6.32**	2312	152257	1.52	11.21

Results

EXAMPLES ULTIMATE PHASE



Visual Story: $(ind = 1791)$					
$\mathcal{F}_1$:	My husband was happy to sit down and enjoy the day with us.	We all sat and enjoyed the park with good food.	These three sisters got together for a picnic.	We heard a man singing an awesome song.	At night, we all gathered *around the fire* and talked.
$\mathcal{F}_2$:	A couple takes a break from all the excitement of the day to enjoy a *nice lunch*.	A group of friends gather for a day in the park full of fun.	Some of the friends decide to take a break and just enjoy the day.	A man announces to the group that he is going to perform a song for everyone.	At the end of the night, the friends gather together to say goodbye to each other.
$\mathcal{F}_3$:	My boyfriend and I attended a local fair.	There were many people there who had different kinds of tattoos.	It was a fun time and I am glad I went to it.	I had a tattoo of my father on my arm.	The fairgrounds were filled with different types of people.
$\mathcal{F}_4$:	We had a lot of fun sitting around talking.	I went to the park yesterday.	I had a great time there.	I got some tattoos while I was there.	There were a ton of people there.
`Humans`:	I think I need a new tattoo to commemorate the occasion.	We had the babies there.	At dusk we all did a prayer circle.	We spent the day at the park.	We had the grannies there.

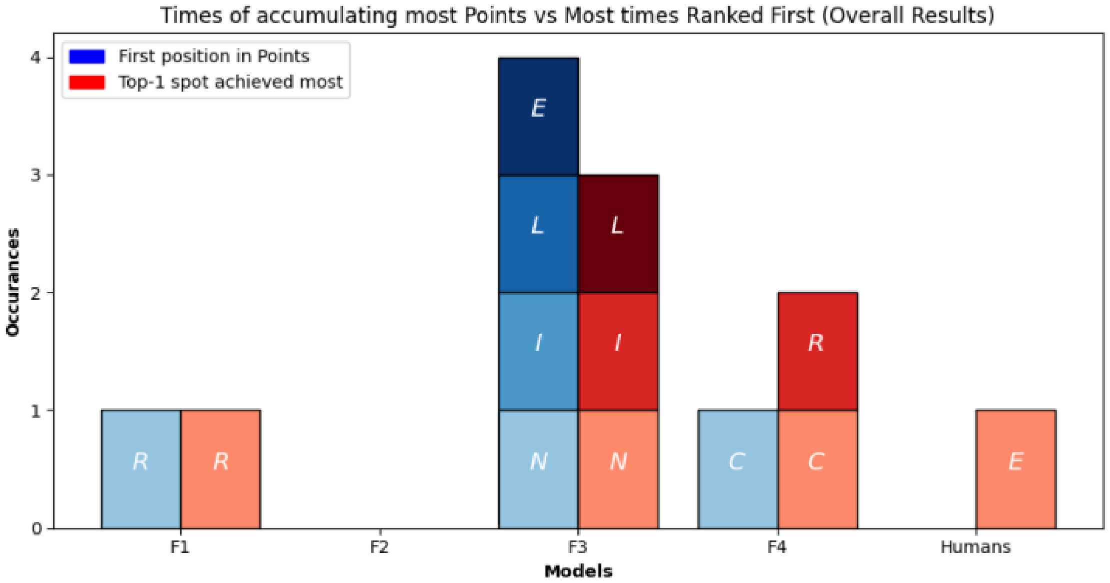
Utrecht University

11-11-2024

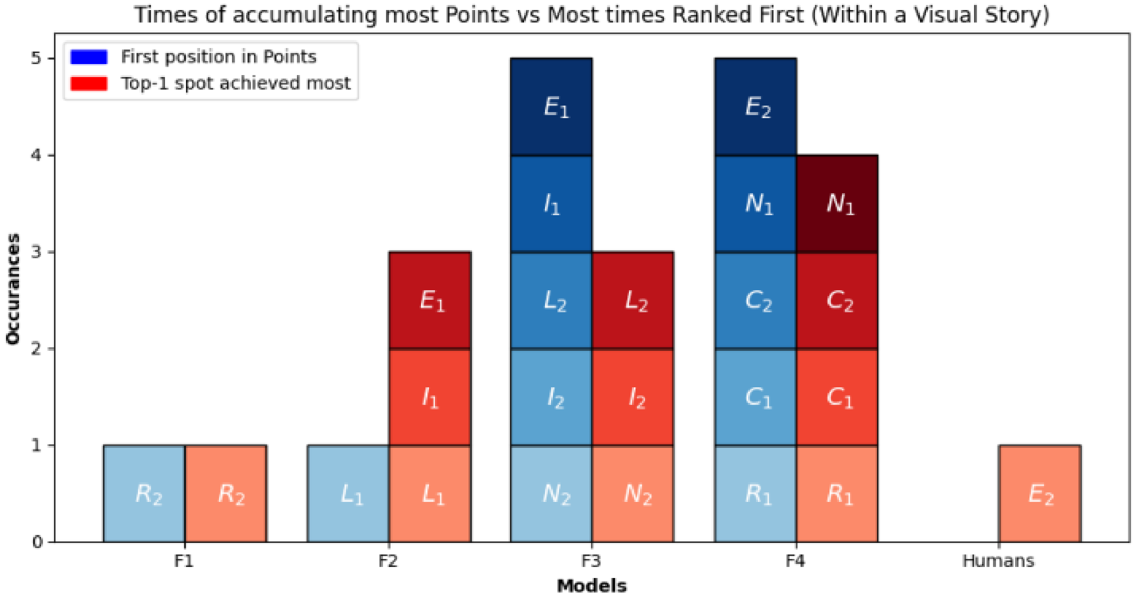
Results

HUMAN EVALUATION ON CRITERIA

- Rank the generated stories according to criteria. Based on the models' position we give points.
- Per framework we care about: 1) the total sum of gathered points, 2) the number of times it got the first position.
- 25 participants & each criterion was evaluated upon 2 different visual stories →50 ranks per criterion.
- Overall: F3 dominates mainly because it gets the most points (& top-1 spots) in I, E and L.
- Within stories: F3 & F4 are both very well. F4 prevails regarding R and C, while N is in between.



(a) Total times that each model either obtained: 1) the most points within a criterion or 2) most times the top-1 spot within a criterion, accounting for both visual stories utilized.

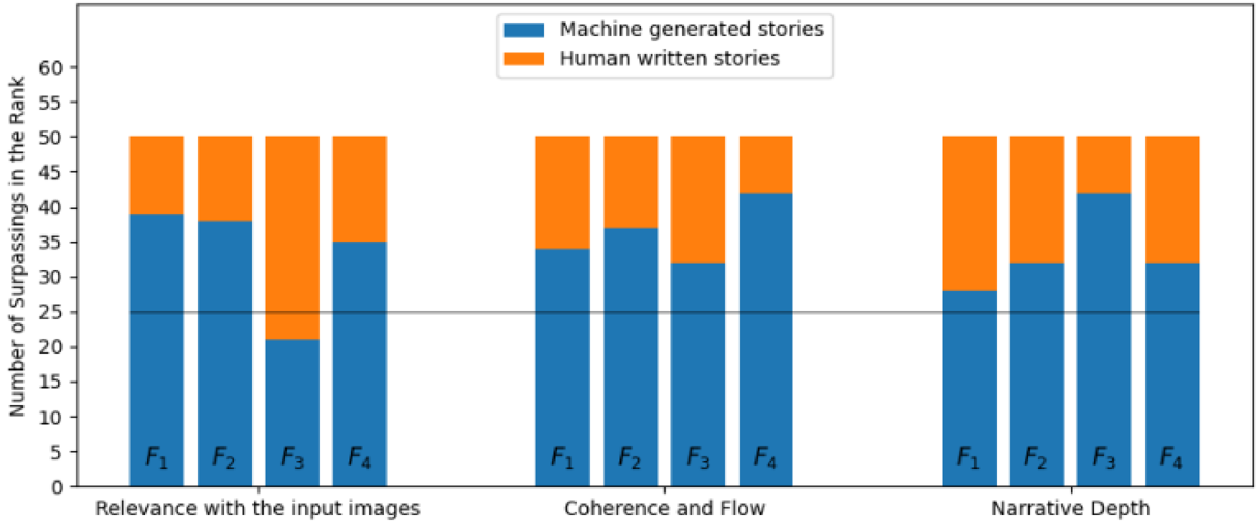


(b) Times that each model, within one of the two tested visual stories, obtained either: 1) the most total points within a criterion or 2) most times the top-1 spot within a criterion.

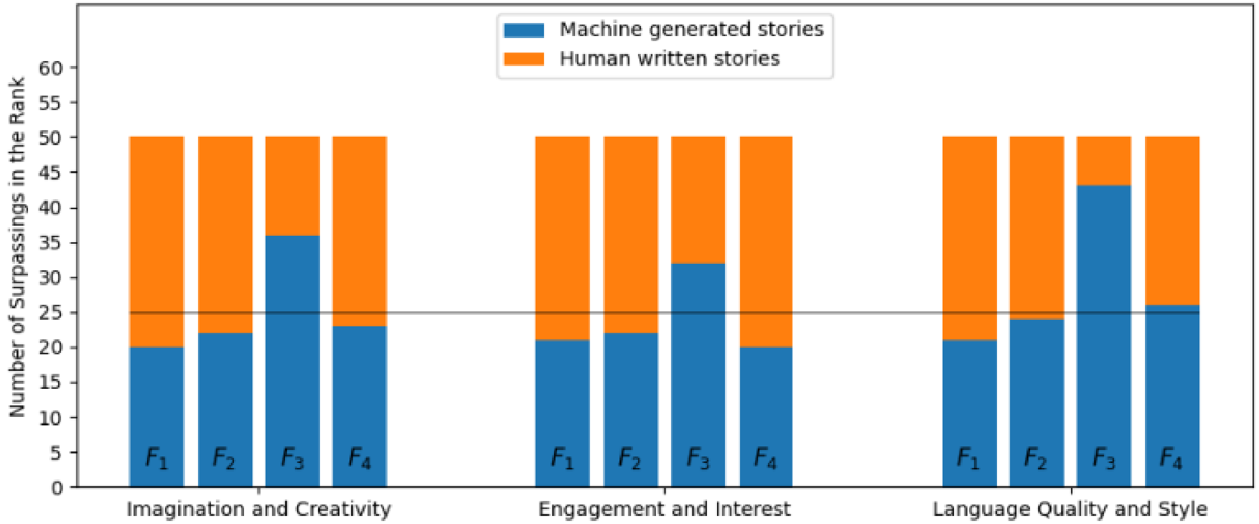
Results

HUMANS VS MACHINES

- Here we care about the number of times that a machine generated story by any framework was placed above the respective human stories.
- We have 50 ranks → Majority is > 25.
- On simpler criteria (R, C), our models outperform the human stories.
- In more abstract terms such as I & E humans deliver better possibly because the L is better → engages more the annotators.



(a) *Head-to-head comparison for Relevance, Coherence and Narration.*

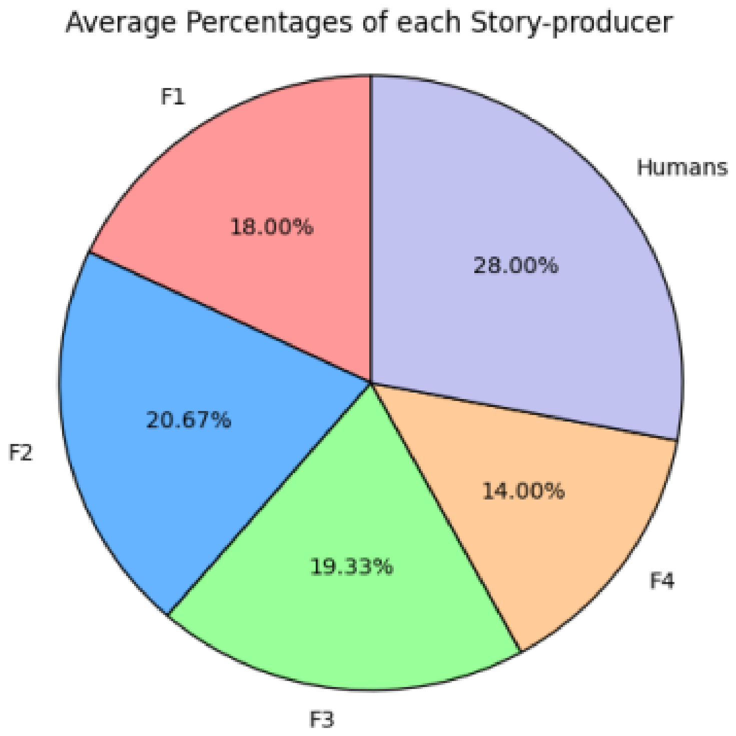


(b) *Head-to-head comparison for Imagination, Engagement and Language.*

Results

HUMANS VS MACHINES

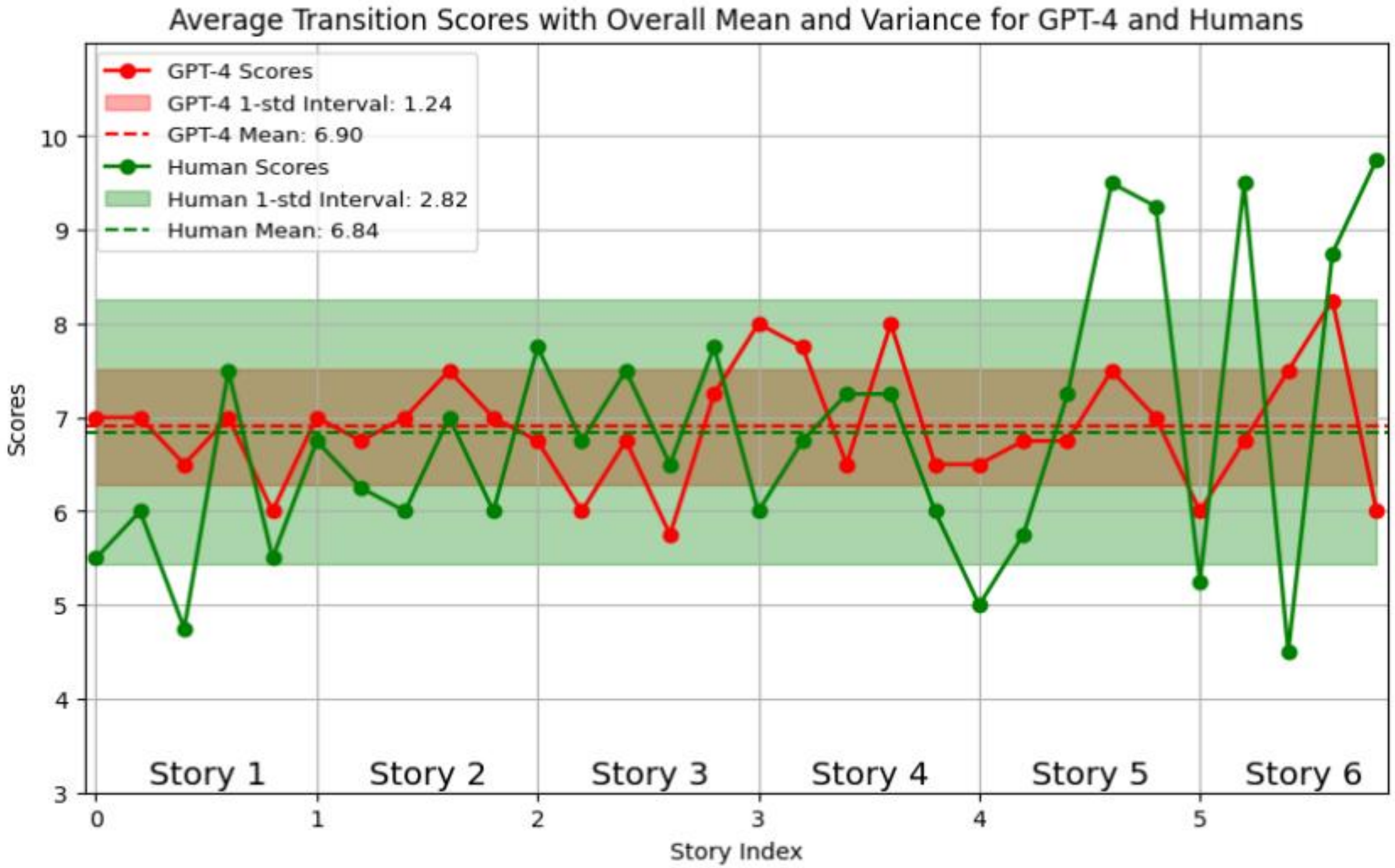
- Among the five given storylines, find the (one) human written.
- Indeed, on average the participants guessed the actual human story as the one that was not machine generated.
- Still weak majority → Lot of uncertainty of what is human written/not.



Results

HUMANS VS (STRONG) MACHINES

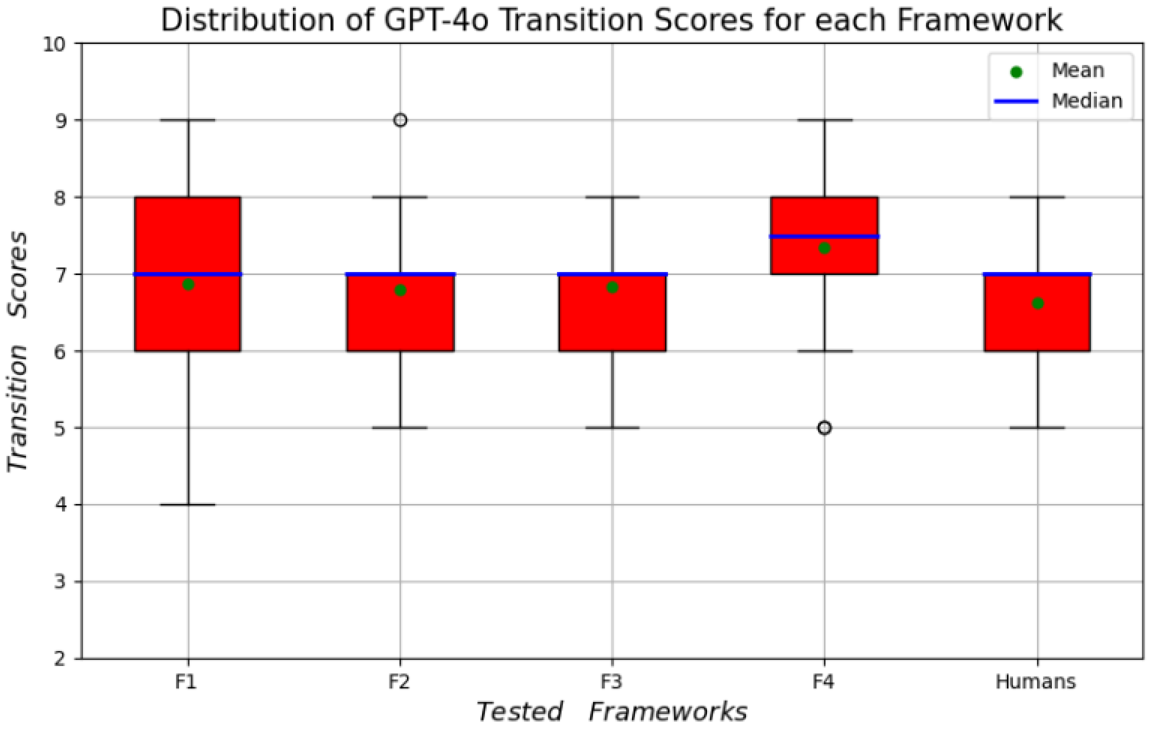
- We asked GPT-4o to judge the transition between the generated sentences in terms of coherence and logical flow giving a mark from 1 to 10.
- Recruiting two workers to do the precisely the same for cross-validating the results.
- Both kinds of evaluation agree on average scores.
- Some similar grading patterns can be discerned.
- Human scores maintain much greater variance → Although depends on the evaluator.

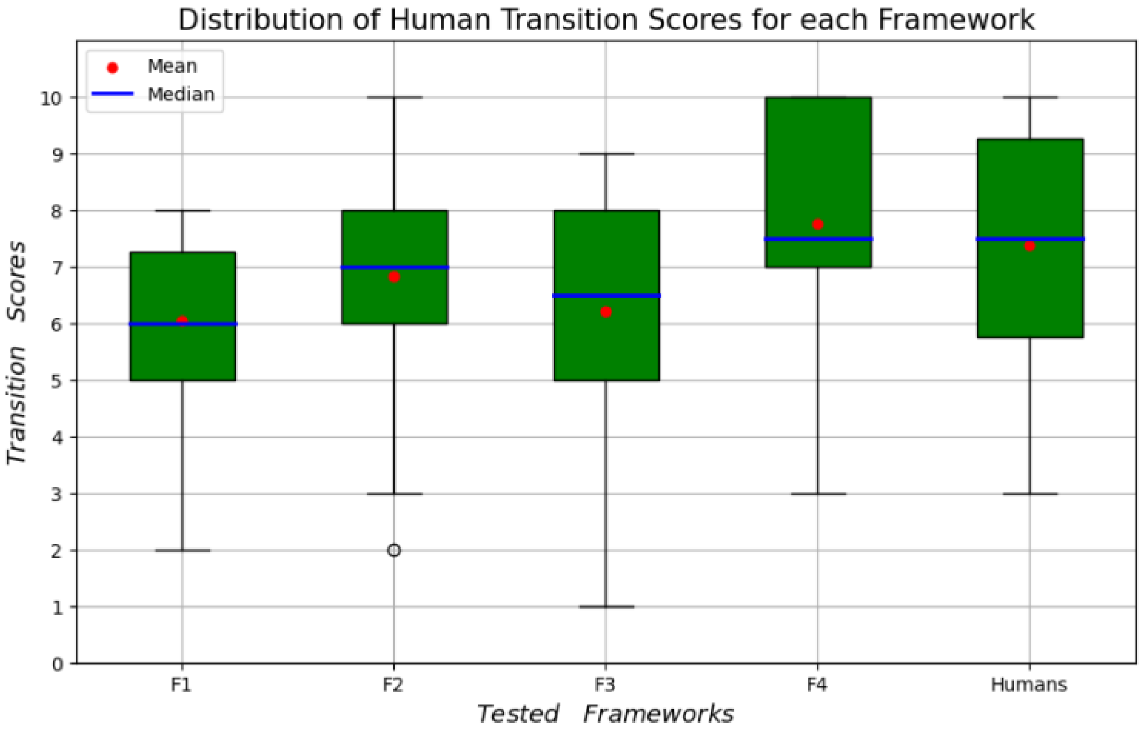


Results

HUMANS VS (STRONG) MACHINES

- Per Model Comparison between Human & GPT-4o judgment → Both agree that F4 produced the most coherent stories, while F1 and F3 the least coherent ones.





Conclusions

MAIN TAKEAWAYS & LIMITATIONS

Per (sub)-research question:

- Sub-Q1: Fine-tuned Clip-Cap models are quite accurate on VIST dataset for captioning.
- Sub-Q2a: Unfortunately, T5 & BART > T4 in terms of automatic evaluation. However, the are immensely more pretrained and much more capacious in terms of parameters.
- Out of the box: In terms of general lexical traits, T4 is matching against T5 & BART and obtains greater diversity and ideality on some occasions.
- Problem of T4: It loses correspondence with its input (in this case the captions).
- Sub-Q2b: Undoubtedly, BART & T5 can be fine-tuned appropriately for the task of storytelling, maintaining high attention to the input and delivering diversity as well.
- Sub-Q3: All evaluation procedures showed that our proposed framework is capable of visual storytelling.
- Best framework?:

  1) Both automatic & human evaluations showed that F4 is particularly strong.
  Nevertheless, it depends on the criteria we look upon (F3 is also strong).
  2) LLM and human evaluation showed that F4 produces more cohesive stories.
- Humans vs Machines: Our stories pose a serious challenge for recognizing which story is human written or not, especially regarding R, C and N → *We are approaching the human baseline for VIST dataset.*
- HOWEVER…, due to the moderate quality of human stories in VIST, we can’t claim that our framework is on general human level of storytelling

Conclusions

EXPANSIONS

- Our proposed framework is not trained end-to-end → It's promising to unify the model and subject it to fine-tuning under a common loss.
- Changing the way that we exploit the captions → Instead of "Serialization", we could use "Parallelization" along with the image features.

Serialization:

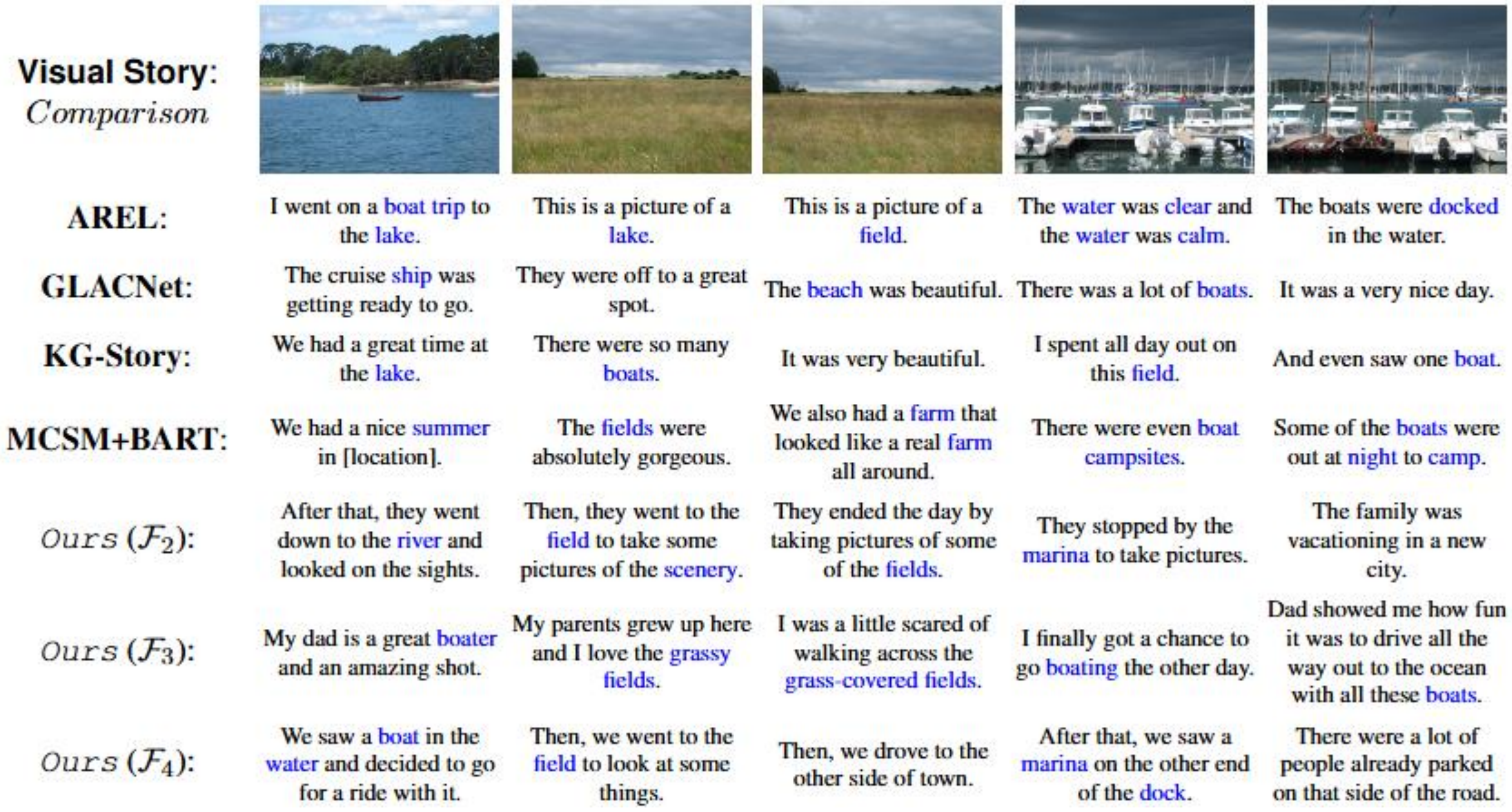
$$images \rightarrow captions \rightarrow story$$

Parallelization:

$$images + captions \rightarrow story$$

Comparison with others works

CAN WE STAND AGAINST SOTA ?

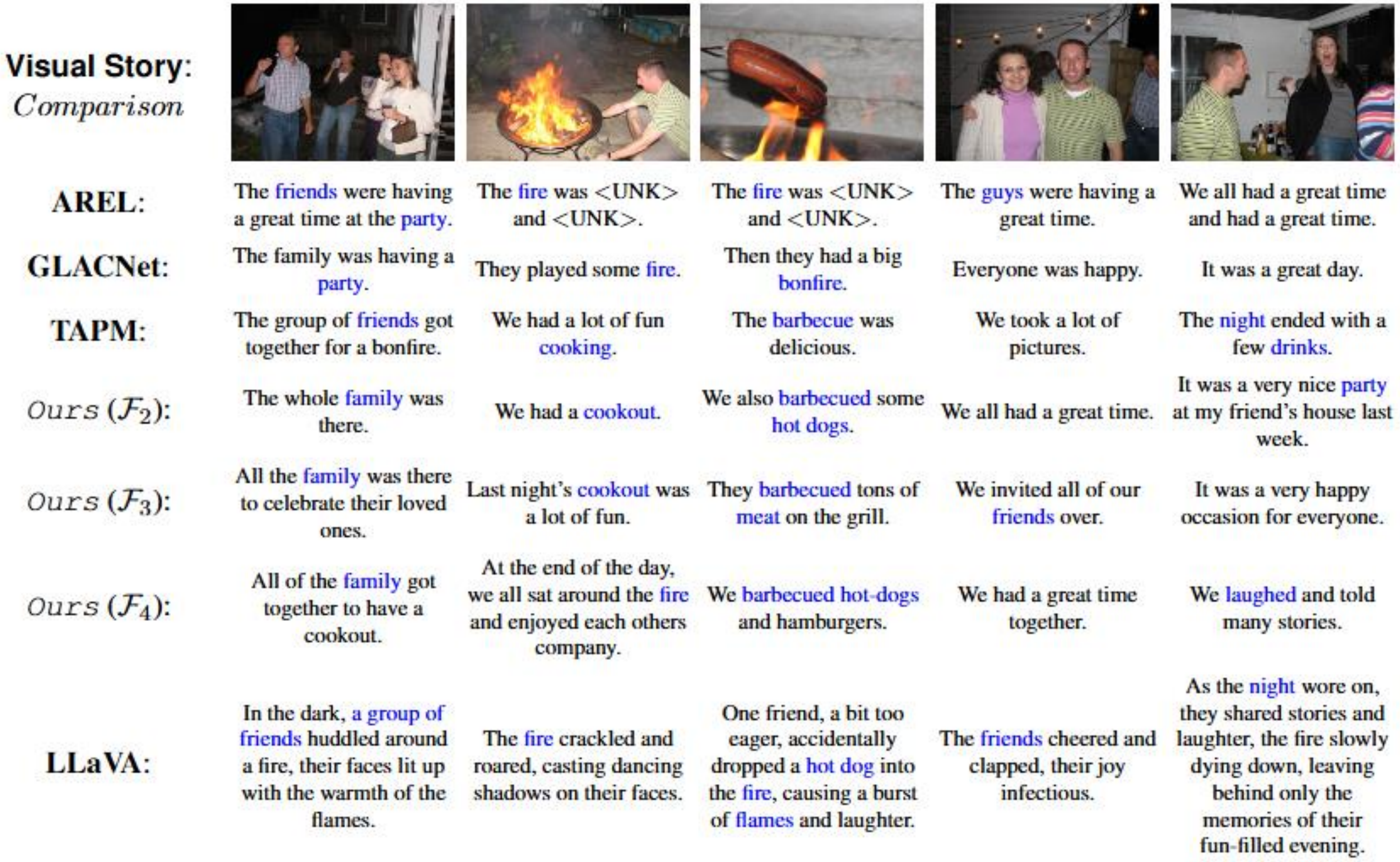


Visual Story: *Comparison*					
AREL:	I went on a boat trip to the lake.	This is a picture of a lake.	This is a picture of a field.	The water was clear and the water was calm.	The boats were docked in the water.
GLACNet:	The cruise ship was getting ready to go.	They were off to a great spot.	The beach was beautiful.	There was a lot of boats.	It was a very nice day.
KG-Story:	We had a great time at the lake.	There were so many boats.	It was very beautiful.	I spent all day out on this field.	And even saw one boat.
MCSM+BART:	We had a nice summer in [location].	The fields were absolutely gorgeous.	We also had a farm that looked like a real farm all around.	There were even boat campsites.	Some of the boats were out at night to camp.
Ours ($\mathcal{F}_2$):	After that, they went down to the river and looked on the sights.	Then, they went to the field to take some pictures of the scenery.	They ended the day by taking pictures of some of the fields.	They stopped by the marina to take pictures.	The family was vacationing in a new city.
Ours ($\mathcal{F}_3$):	My dad is a great boater and an amazing shot.	My parents grew up here and I love the grassy fields.	I was a little scared of walking across the grass-covered fields.	I finally got a chance to go boating the other day.	Dad showed me how fun it was to drive all the way out to the ocean with all these boats.
Ours ($\mathcal{F}_4$):	We saw a boat in the water and decided to go for a ride with it.	Then, we went to the field to look at some things.	Then, we drove to the other side of town.	After that, we saw a marina on the other end of the dock.	There were a lot of people already parked on that side of the road.

Example of visual generated stories by four state-of-the-art Visual Storytelling models, along with the respective generated stories from three variants of our proposed framework. Highlighted blue words visually relate to the input images.

Comparison with others works

CAN WE STAND AGAINST SOTA ?



Visual Story: *Comparison*					
AREL:	The friends were having a great time at the party.	The fire was <UNK> and <UNK>.	The fire was <UNK> and <UNK>.	The guys were having a great time.	We all had a great time and had a great time.
GLACNet:	The family was having a party.	They played some fire.	Then they had a big bonfire.	Everyone was happy.	It was a great day.
TAPM:	The group of friends got together for a bonfire.	We had a lot of fun cooking.	The barbecue was delicious.	We took a lot of pictures.	The night ended with a few drinks.
Ours ($\mathcal{F}_2$):	The whole family was there.	We had a cookout.	We also barbecued some hot dogs.	We all had a great time.	It was a very nice party at my friend's house last week.
Ours ($\mathcal{F}_3$):	All the family was there to celebrate their loved ones.	Last night's cookout was a lot of fun.	They barbecued tons of meat on the grill.	We invited all of our friends over.	It was a very happy occasion for everyone.
Ours ($\mathcal{F}_4$):	All of the family got together to have a cookout.	At the end of the day, we all sat around the fire and enjoyed each others company.	We barbecued hot-dogs and hamburgers.	We had a great time together.	We laughed and told many stories.
LLaVA:	In the dark, a group of friends huddled around a fire, their faces lit up with the warmth of the flames.	The fire crackled and roared, casting dancing shadows on their faces.	One friend, a bit too eager, accidentally dropped a hot dog into the fire, causing a burst of flames and laughter.	The friends cheered and clapped, their joy infectious.	As the night wore on, they shared stories and laughter, the fire slowly dying down, leaving behind only the memories of their fun-filled evening.

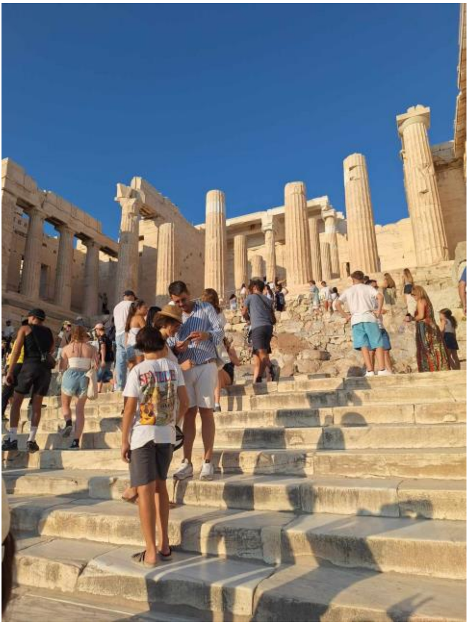
Example of visual generated stories by four state-of-the-art Visual Storytelling models, along with the respective generated stories from three variants of our proposed framework. Highlighted blue words visually relate to the input images.

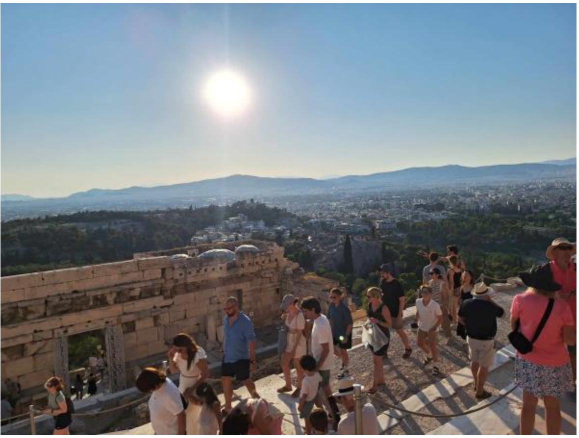
Utrecht University

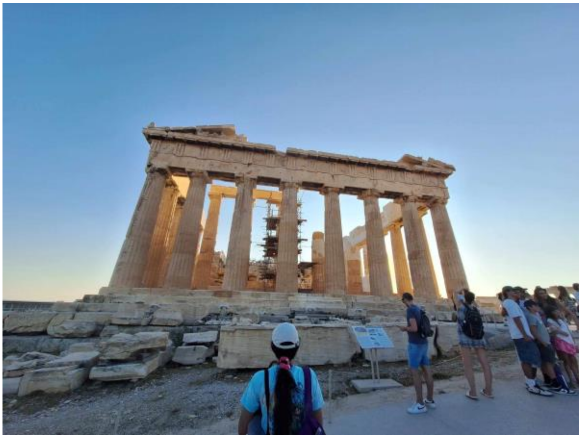
My Story

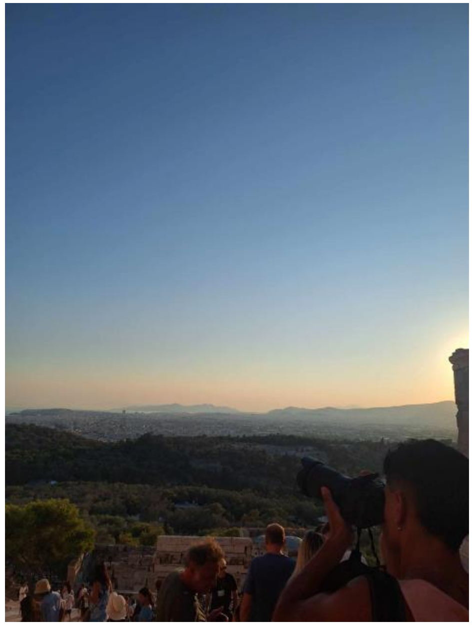
11-11-2024

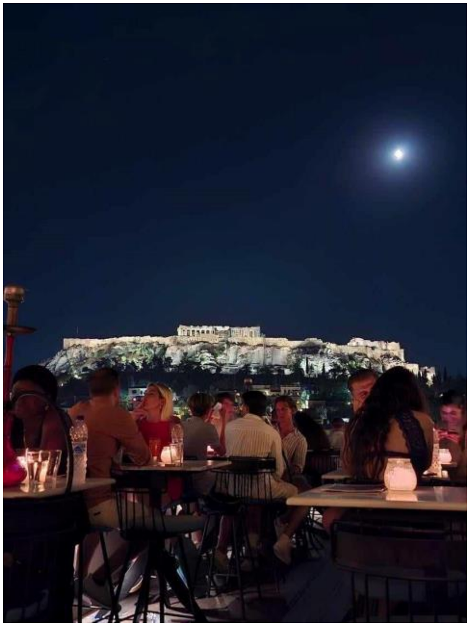
A VISIT TO ACROPOLIS











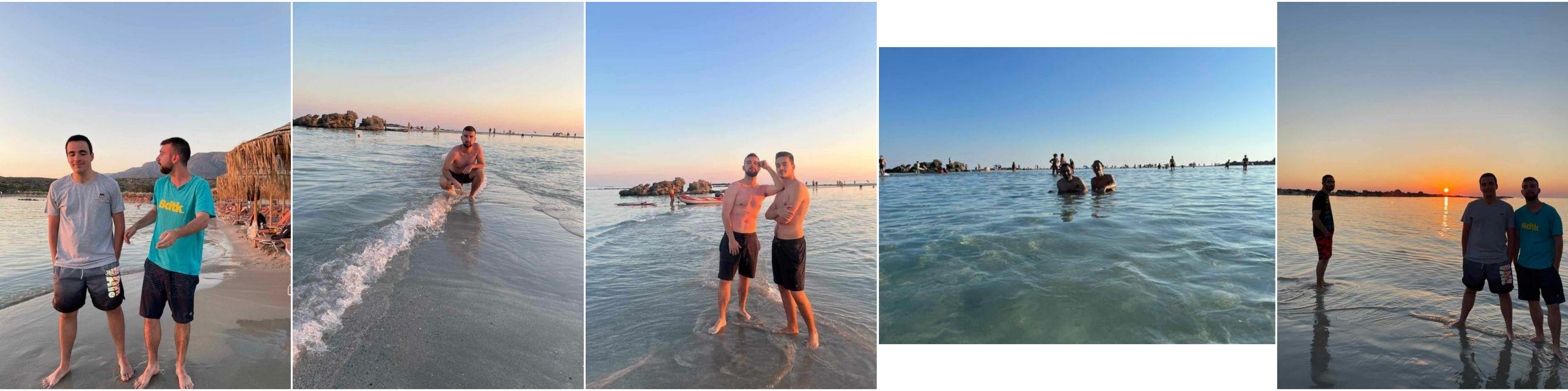
STORY 1: They checked out the museum before leaving. They stood on the steps for a while. The group visited the museum in Greece. As the sun set, they watched down the city. Finally, they ate dinner .

STORY 2: We are just getting to the gate and ready to go in. Oh, my goodness! This is just huge. Look at all the people standing here. Wow, this building is so huge. We better go and get a picture before we leave. The view from our table tonight, awesome.

STORY 3: The tour had begun on time; we could see the group gathering. We gathered in front of the columned building on the outskirts of the city. My friends and I decided to take a trip to a historic site that was under construction. Once inside, I took pictures of some of the more important areas. Finally, we went to dinner at a nearby pub to end our tour.

My Story

IN THE ELAFONISI BEACH



STORY 1: *Two friends decided to go to the beach for the weekend. It was a little chilly, so they wore wetsuits. The two friends were glad they wore sunscreen. They stayed until sundown and then they went home. Their friend joked around before going into the water.*

STORY 2: *Jason and Steve made a pact to never go in the ocean together. Then Jason realized he'd never gone in the water before, and he jumped in immediately. He grabbed his friend, and they ran as fast as they could away from the waves. They waded into the ocean and had a blast. After they were done, they posed together for a photo to show everyone they'd gone swimming.*

THANK YOU VERY MUCH FOR YOUR ATTENTION

Utrecht University

Sharing science,
shaping tomorrow

QUESTIONS ?