StoryBench: A Multifaceted Benchmark for Continuous Story Visualization

Emanuele Bugliarello | Hernan Moraldo | Ruben Villegas | Mohammad Babaeizadeh | Mohammad Taghi Saffar
Han Zhang | Dumitru Erhan | Vittorio Ferrari | Pieter-Jan Kindermans | Paul Voigtlaender

Google Research
Google DeepMind
UNIVERSITY OF COPENHAGEN

We collect datasets that describe videos with a sequence of captions, one for each action, forming the story of the video; and their corresponding timestamps

We also (i) annotate each video segment with 34 labels; (ii) show the benefits of training on story-like data; (iii) establish human evaluation of video stories; and (iv) reaffirm the need for better automatic metrics for video generation

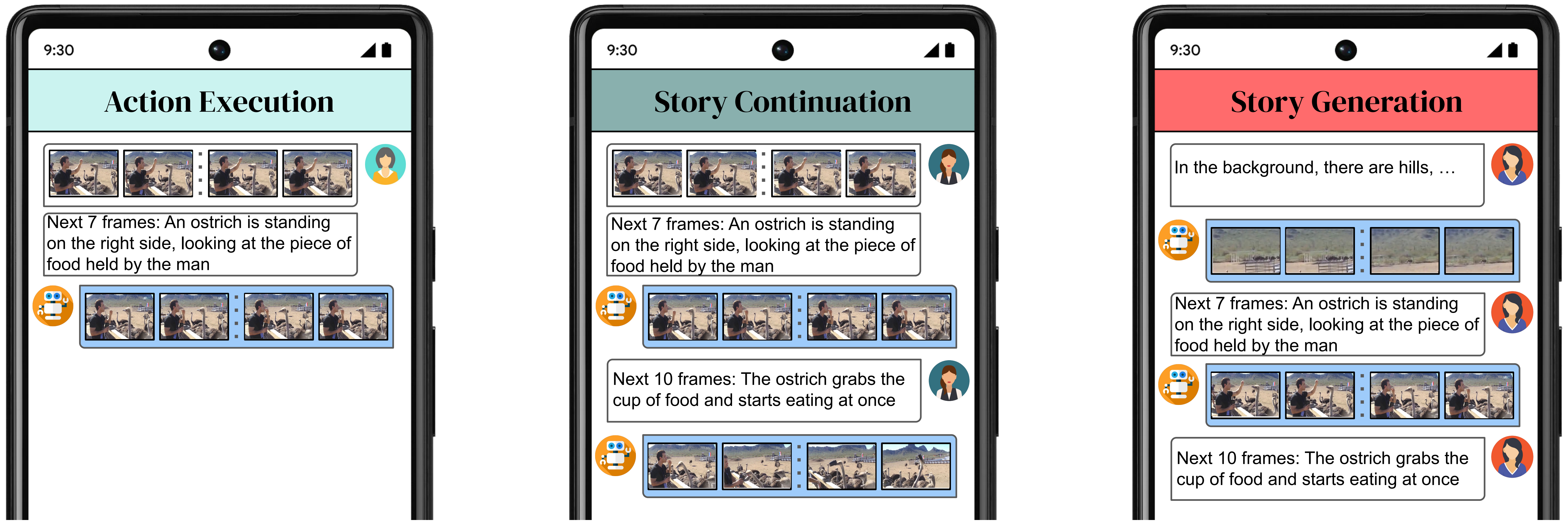
Evaluation Data Statistics

Dataset	# Videos	# Stories per video	# Segments per story
DiDeMo-CSV	1,399	1.00	3.52
Oops-CSV	1,888	1.72	2.22
UVO-CSV	2,917	1.46	1.46

Diagnostic Categories & Labels

Category	Labels
Camera Movements	static shot, pan, tilt, …
Foreground Entities	people, animals, …
Foreground Actions	humans moving, …
Background Actions	objects moving, …
Foreground Interactions	dialogues, direct, …
Foreground Transitions	new entities, …

Text-to-Video Tasks



Action Execution:

Next 7 frames: An ostrich is standing on the right side, looking at the piece of food held by the man

Story Continuation:

Next 7 frames: An ostrich is standing on the right side, looking at the piece of food held by the man

Next 10 frames: The ostrich grabs the cup of food and starts eating at once

Story Generation:

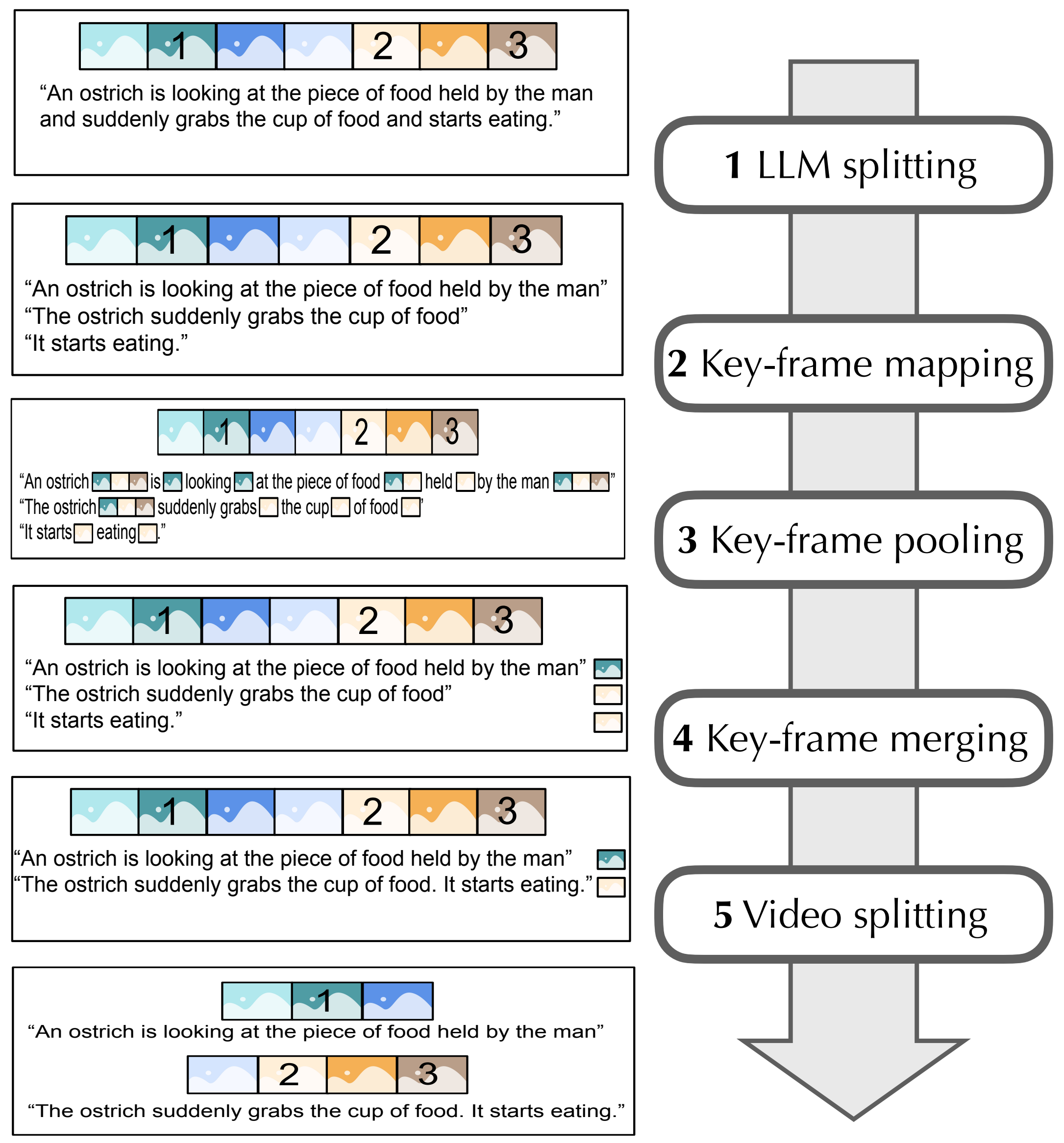
In the background, there are hills, …

Next 7 frames: An ostrich is standing on the right side, looking at the piece of food held by the man

Next 10 frames: The ostrich grabs the cup of food and starts eating at once

- **Action Execution**: Generate the next action specified in the input given the preceding ground-truth
- **Story Continuation**: Generate a video from a sequence of inputs given the first 0.5s of ground-truth video
- **Story Generation**: Generate a video from a sequence of inputs, given a video synthesized from a description of the background

Training Data Challenge



"An ostrich is looking at the piece of food held by the man and suddenly grabs the cup of food and starts eating."

"An ostrich is looking at the piece of food held by the man"
"The ostrich suddenly grabs the cup of food"
"It starts eating."

"An ostrich is looking at the piece of food held by the man"
"The ostrich suddenly grabs the cup of food"
"It starts eating."

"An ostrich is looking at the piece of food held by the man"
"The ostrich suddenly grabs the cup of food"
"It starts eating."

"An ostrich is looking at the piece of food held by the man"
"The ostrich suddenly grabs the cup of food. It starts eating."

"An ostrich is looking at the piece of food held by the man"
"The ostrich suddenly grabs the cup of food. It starts eating."

Experimental Setup

PHENAKI (345M)

Training

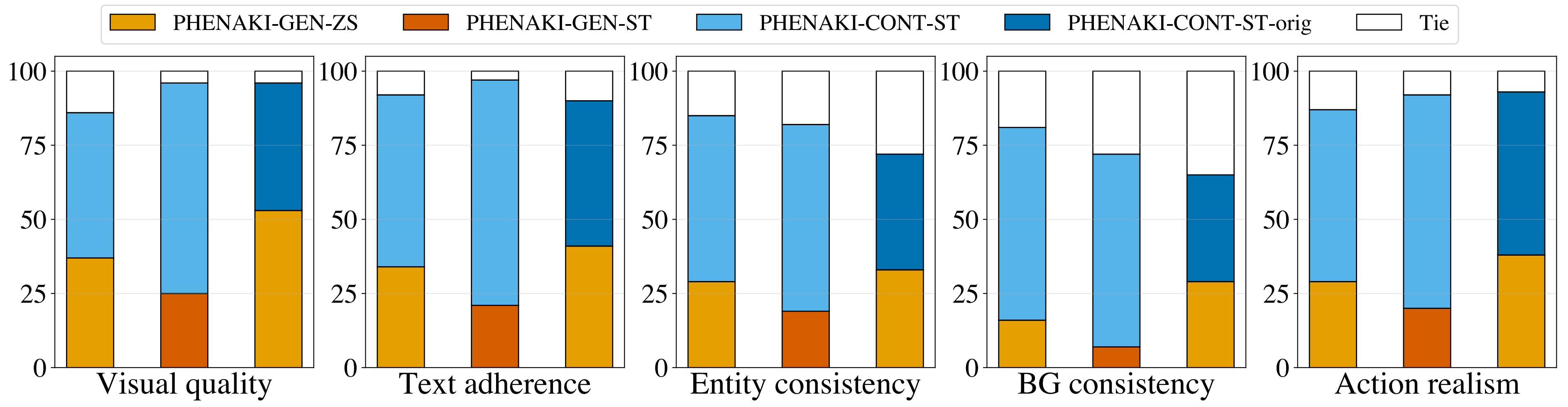
- **-GEN:** generation mode training
- **-CONT:** continuation mode fine-tuning

Evaluation

- **-ZS:** zero-shot evaluation
- **-ST:** single-task fine-tuning
- **-ST-orig:** single-task fine-tuning on original training data (no story-like pipeline applied)
- **-MT:** multi-task fine-tuning

Human Evaluation

Story Continuation on Oops-CSV



- **Action Execution & Story Continuation**
  - Fine-tuning in continuation mode is effective
  - Fine-tuning on the original data brings less benefits than using story-like data
- **Story Generation**: The zero-shot model is generally preferred

Automatic Evaluation

Oops-CSV	Action Execution					Story Continuation					Story Generation				
Phenaki	FID	FVD	SIM	PQA	VTM	FID	FVD	SIM	PQA	VTM	FID	FVD	SIM	PQA	VTM
Zero-Shot															
-Gen-ZS	167	416	72.8	**5.8**	**22.1**	277	623	70.3	**7.2**	**21.7**	310	933	N/A	**8.1**	21.0
Single-Task															
-Gen-ST	177	446	72.3	4.0	21.5	250	589	70.0	4.3	21.3	246	**614**	N/A	4.3	**21.1**
-Cont-ST	**114**	**350**	**73.2**	4.9	21.5	**155**	**488**	**71.1**	5.3	21.2	**171**	711	N/A	5.4	19.4
Multi-Task															
-Cont-MT	140	353	72.8	4.7	21.7	198	511	70.6	5.1	21.4	201	860	N/A	5.0	19.4

- **Action Execution & Story Continuation**
  - Phenaki-Cont-ST performs the best w.r.t. FID, FVD, and SIM
- Gen-ZS achieves higher **PQA**, but humans found all models have similar ***visual quality***
- While humans prefer CONT-ST over GEN-ZS, **SIM**, **PQA** and **VTM** metrics do **not** reflect this