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

NeurIPS 2023

Google Research
Google DeepMind
UNIVERSITY OF COPENHAGEN

Generative AI for Human Creativity

Text

Sound

Image

A small cactus wearing a straw hat and neon sunglasses in the Sahara desert.

What's next? Movies 🎬

Challenges of video generation

- Coherent over time
- Smooth transitions
- Reflect the actions described in text prompts
- Computationally expensive
- Smaller video–text datasets

StoryBench: Overview

Datasets that describe the story of a video

- With a sequence of captions (one for each action)
- And their corresponding timestamps

StoryBench: Example

PROMPT: A man wearing white shorts is jumping on a trampoline.

StoryBench: Example

PROMPT: A man wearing white shorts is jumping on a trampoline.

PROMPT: The man performing a flip.

StoryBench: Example

PROMPT: A man wearing white shorts is jumping on a trampoline.

PROMPT: The man performing a flip.

PROMPT: The man falls when the trampoline falls on the ground.

StoryBench: Example & Stats

PROMPT: A man wearing white shorts is jumping on a trampoline.

PROMPT: The man performing a flip.

PROMPT: The man falls when the trampoline falls on the ground.

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

StoryBench: Example & Stats & Labels

PROMPT: A man wearing white shorts is jumping on a trampoline.

PROMPT: The man performing a flip.

PROMPT: The man falls when the trampoline falls on the ground.

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

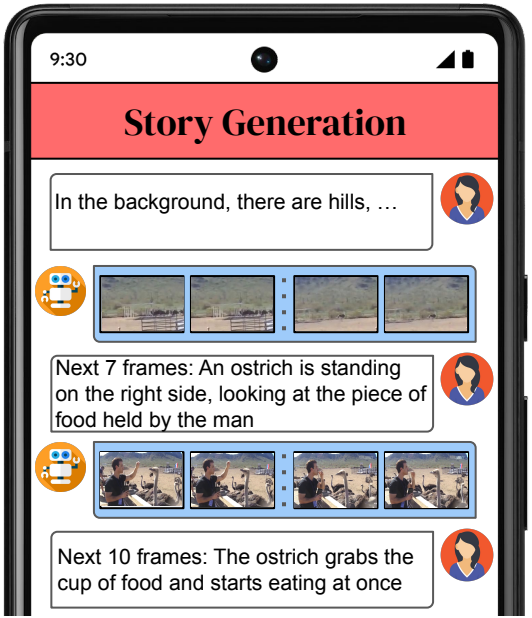
Diagnostic Categories & Labels

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, …

StoryBench: Tasks

StoryBench: Tasks

StoryBench: Tasks



Training Data Challenge

- Lack of large-scale high-quality data limits text-to-video models
- We define the challenge of training data curation for StoryBench

Training Data Challenge

- Lack of large-scale high-quality data limits text-to-video models
- We define the challenge of training data curation for StoryBench
- A first approach to transform the captions for VidLN videos into stories

Experimental Setup

Baseline

- **Phenaki-Gen: A 345M Phenaki model**

Experimental Setup

Baseline

- Phenaki-Gen: A 345M Phenaki model

Fine-tuning

- -GEN: generation mode
- -CONT: continuation mode

Experimental Setup

Baseline

- Phenaki-Gen: A 345M Phenaki model

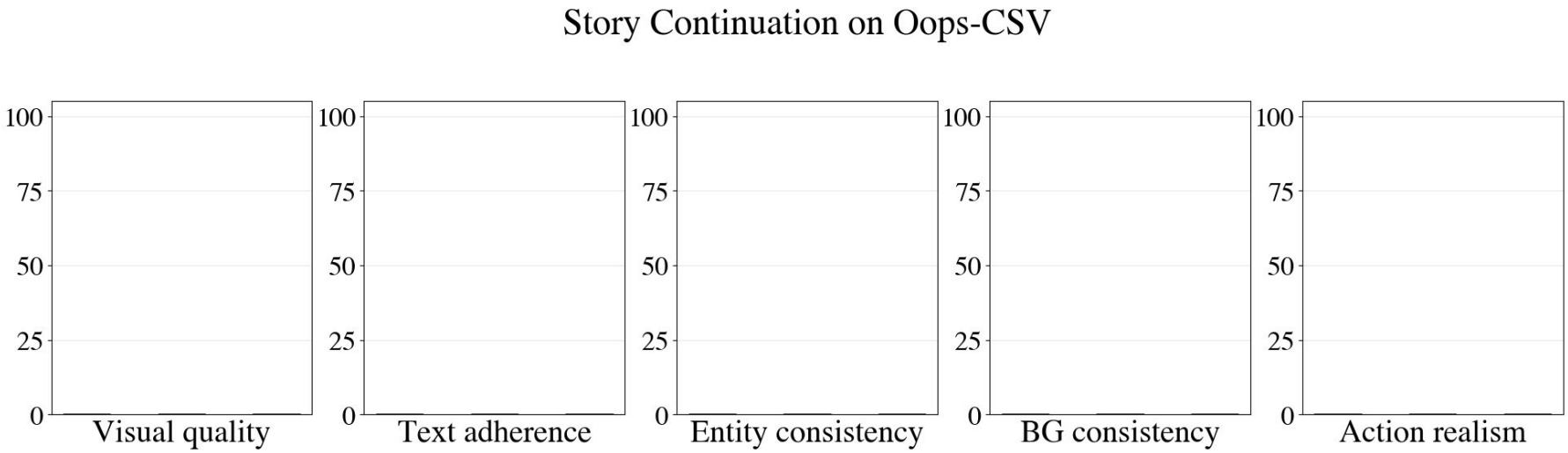
Fine-tuning

- -GEN: generation mode
- -CONT: continuation mode

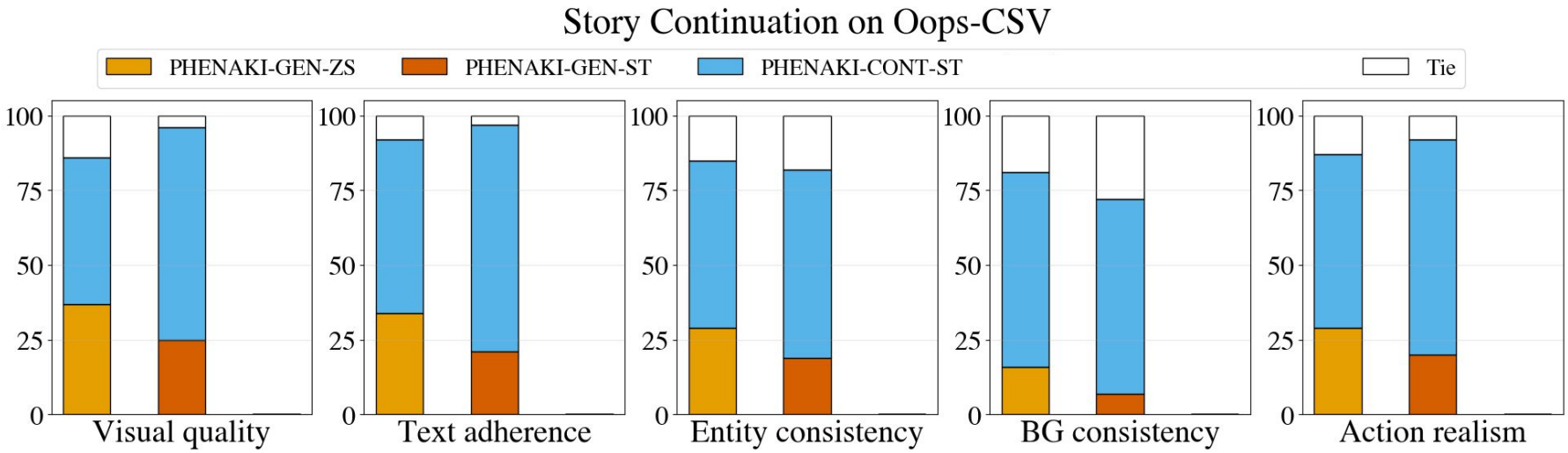
Evaluation

- -ZS: zero-shot
- -ST: single-task fine-tuning
- -MT: multi-task fine-tuning

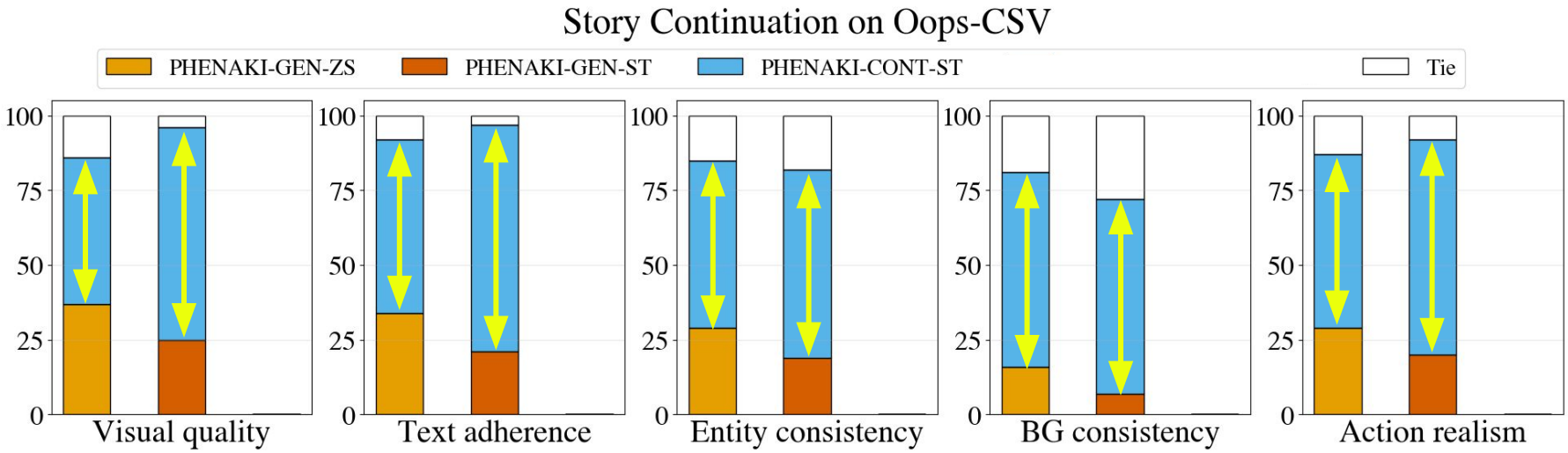
Human Evaluation



Human Evaluation

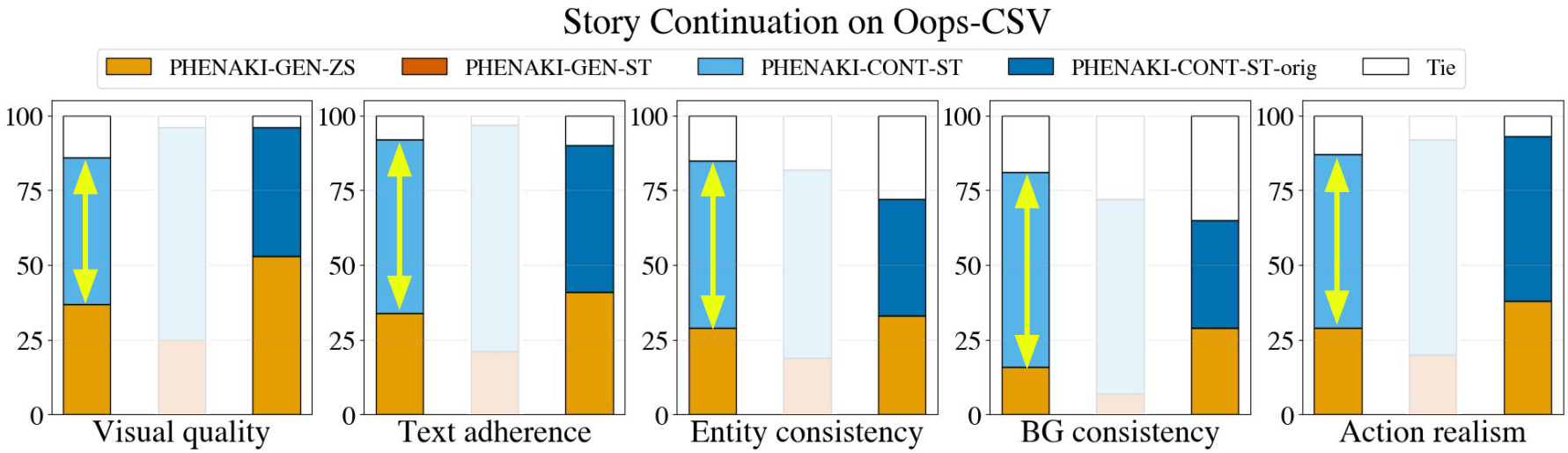


Human Evaluation



- Fine-tuning in continuation mode is effective

Human Evaluation



- Fine-tuning in continuation mode is effective
- Fine-tuning on story-like data is better

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

Overall, automatic metrics do not correlate with human ratings

Conclusion

- New annotations to generate videos from a sequence of text prompts
  - Timestamps for each text prompt
  - Diagnostic labels for each video segment
- StoryBench: a new benchmark to measure progress of text-to-video models
  - 3 different tasks, 3 datasets, and 3 evaluation setups
- Fine-tuning for continuation improves key challenges in video generation
- Our results highlight a discrepancy between human and automatic ratings

https://github.com/google/storybench