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
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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Videos</th>
<th># Stories per video</th>
<th># Segments per story</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiDeMo-CV</td>
<td>1,399</td>
<td>1.00</td>
<td>3.52</td>
</tr>
<tr>
<td>Oops-CV</td>
<td>1,888</td>
<td>1.72</td>
<td>2.22</td>
</tr>
<tr>
<td>UVO-CV</td>
<td>2,917</td>
<td>1.46</td>
<td>1.46</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Videos</th>
<th># Stories per video</th>
<th># Segments per story</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiDeMo-CV</td>
<td>1,399</td>
<td>1.00</td>
<td>3.52</td>
</tr>
<tr>
<td>Oops-CV</td>
<td>1,888</td>
<td>1.72</td>
<td>2.22</td>
</tr>
<tr>
<td>UVO-CV</td>
<td>2,917</td>
<td>1.46</td>
<td>1.46</td>
</tr>
</tbody>
</table>

### Diagnostic Categories & Labels

<table>
<thead>
<tr>
<th>Category</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera Movements</td>
<td>static shot, pan, tilt, ...</td>
</tr>
<tr>
<td>Foreground Entities</td>
<td>people, animals, ...</td>
</tr>
<tr>
<td>Foreground Actions</td>
<td>humans moving, ...</td>
</tr>
<tr>
<td>Background Actions</td>
<td>objects moving, ...</td>
</tr>
<tr>
<td>Foreground Interactions</td>
<td>dialogues, direct, ...</td>
</tr>
<tr>
<td>Foreground Transitions</td>
<td>new entities, ...</td>
</tr>
</tbody>
</table>
StoryBench: Tasks

**Action Execution**

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

Story Continuation on Oops-CSV

Visual quality
Text adherence
Entity consistency
BG consistency
Action realism
Human Evaluation

Story Continuation on Oops-CSV

- Visual quality
- Text adherence
- Entity consistency
- BG consistency
- Action realism
Human Evaluation

Story Continuation on Oops-CSV

- 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

<table>
<thead>
<tr>
<th>Oops-CSV</th>
<th>Action Execution</th>
<th>Story Continuation</th>
<th>Story Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FID</td>
<td>FVD</td>
<td>SIM</td>
</tr>
<tr>
<td>Phenaki</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Gen-ZS</td>
<td>167</td>
<td>416</td>
<td>72.8</td>
</tr>
<tr>
<td>-Gen-ST</td>
<td>177</td>
<td>446</td>
<td>72.3</td>
</tr>
<tr>
<td>-Cont-ST</td>
<td>114</td>
<td>350</td>
<td>73.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Cont-MT</td>
<td>140</td>
<td>353</td>
<td>72.8</td>
</tr>
</tbody>
</table>

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