T2V-Bench: Benchmarking Text-to-Video Models

Angel Bujalance, Amina Izbassar, Jakub Tomaszewski, Despoina Touska

UNIVERSITY OF AMSTERDAM

Problem

While text-to-video generative models are becoming more popular and accessible to the general public, the research field still lacks comprehensive evaluation benchmarks of such models. This not only poses challenges in understanding common failure modes of T2V generative models, but also far-reaching societal impacts.

Failure Examples

- Inconsistent object representation over frames.

"A brown dog is digging a hole in a green garden"

- Confusions in binding colors to corresponding objects.

"A musician playing a blue violin in a dimly lit room with a deep purple walls"

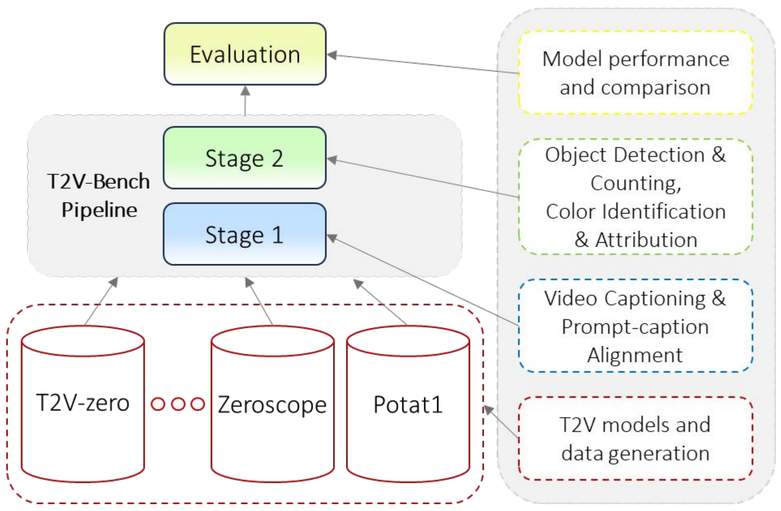
Our Work

A 2-stage method for evaluating Text-to-Video (T2V) generative models in both coarse and fine-grained, object-oriented manner.

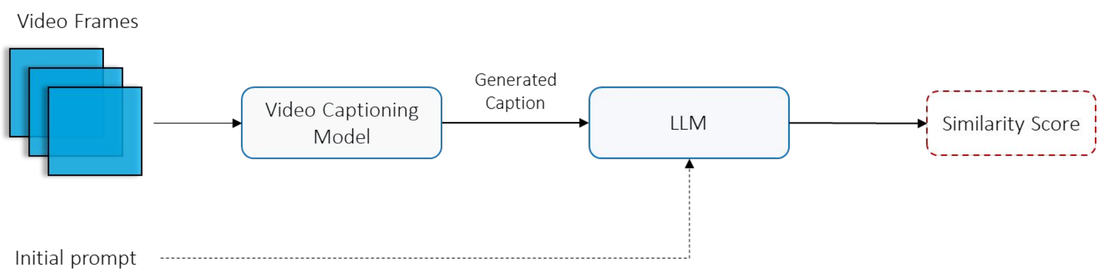
Method

Overview

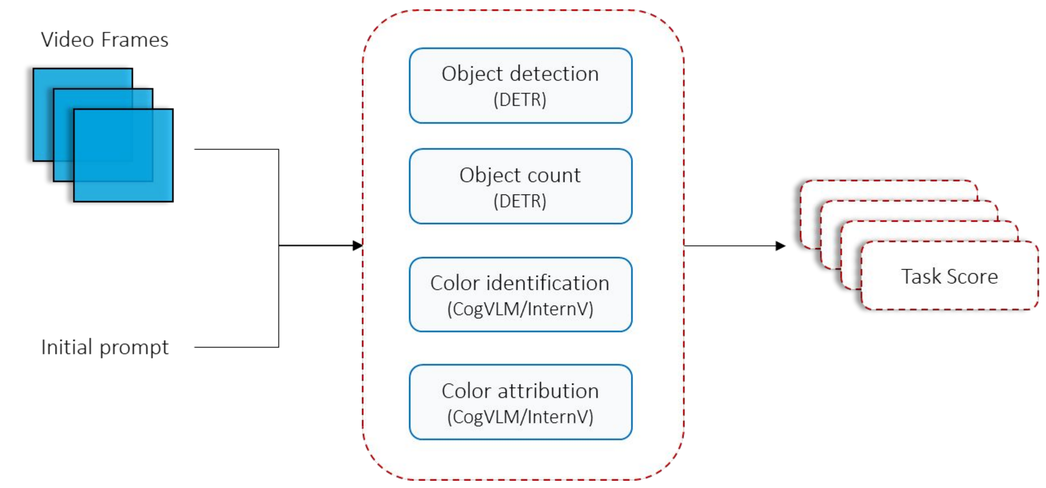
We identify five key task areas were models typically underperform: object identification, object counting, color identification, color attribution, and general video consistency. Based on this, we design an end-to-end, two-stage evaluation pipeline, that evaluates various aspects of the model's generation capabilities.



Stage 1: Holistic evaluation through video captioning and caption-prompt alignment assessment. Videos are captioned using an image-to-text model. and these captions are compared to the original prompts using an LLM to determine their mutual similarity.



Stage 2: Video consistency evaluation on a frame-based level. Leveraging four specific tasks: object detection and counting, color identification and attribution.



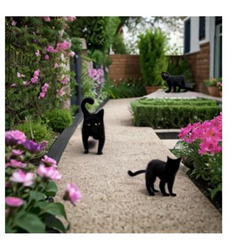
Dataset

- A collection of 700 videos.
- 50 carefully chosen prompts.
- 7 T2V models as video generators.
- Metadata essential for tasks in stage 2.

Prompt: A young boy in a bright green T-shirt is playing with a bright orange basketball on a driveway.

```
"color": {
    "T-shirt": "green",
   "basketball": "orange“
},
"object": {
    "people": 1,
     " ball": 1
}
```

Prompt: A black cat chasing a pink butterfly in a garden.



```
"color": {
      "cat": "black",
      "butterfly": "pink"
},
"object": {
      "cat": 1,
      "butterfly": 1
}
```

Prompt: A group of five people kayak on a calm turquoise lake surrounded by lush green mountains.

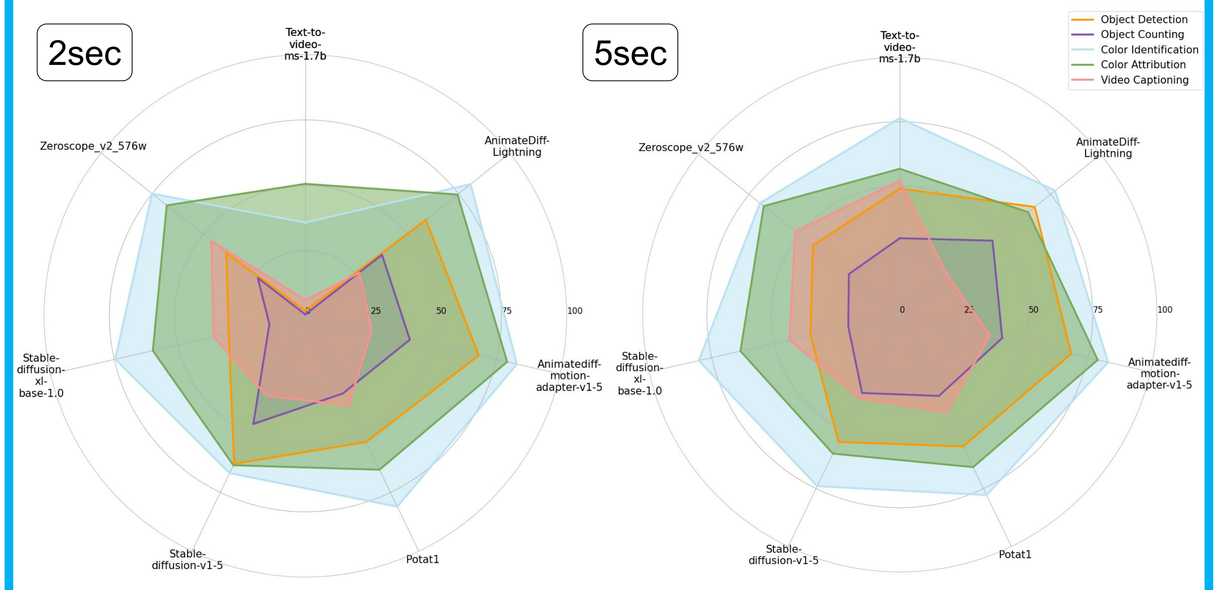
```
"color": {
     "lake": "turquoise",
     "mountain": "green"
},
"object": {
     "people": 5,
      "boat": 1
}
```

Prompt: A group of three friends are playing frisbee with a bright orange disc on a sandy beach at sunset.

```
"color": {
   "shirt": "blue",
   "sunflower": "yellow“
},
"object": {
      "people": 1
}
```

Results

- Evaluation variables:
  - 5 tasks
  - 7 T2V Models
  - 2 different video duration
- Accuracy as evaluation metric



- Text-to-Video-MS 1.7B:
  - Worst performance for **all tasks** in **2-sec** videos.
  - Best video quality for **video captioning** in **5-sec** videos.
- AnimateDiff with Motion Adapter:
  - Best results for **object detection**, **color identification**, and **color attribution** evaluation dimensions in both 2 and 5-sec videos.
- AnimateDiff Light:
  - Performs the best for consistency of the **object counting.**

Key Findings

- Object counting is the task where the models struggle the most.
- Average accuracy for detected objects remains low.
  - Some frames are generated with inconsistent object representation.
- Color identification is the task in which the models perform the best.
- Color attribution also has high accuracy scores. Reasons:
  - T2V models often attribute a color from the prompt to many other objects in the videos.
- Video captioning scores are also low, indicating that the generated videos are not semantically very close to the original prompts.
- Generated videos with longer durations tend to have higher scores.