A Author Statement

We bear all responsibilities for the content, licensing, distribution, and maintenance of our datasets in StoryBench. Our datasets are released under a CC-BY-4.0 license, and our code under an Apache license. Data, code and annotation guidelines are hosted on GitHub at the following URL: `https://github.com/google/storybench`.

B Ethics Statement

The aim of StoryBench is to enable reliable measurements of progress in generative text-to-video models. While this kind of models have great potential to assist and augment human creativity [60], there are broader societal issues that need to be considered when developing these models.

First, while we annotate an evaluation set, training current, strong text-to-video models is computationally expensive. This affects not only their financial cost (*e.g.*, hardware and electricity), but also their environmental cost due to the carbon footprint of modern tensor processing hardware [61].

Second, massive amounts of data are required to train state-of-the-art generative models. Such datasets are harvested from the Web, which tend to reflect social stereotypes, oppressive viewpoints, and harmful associations to marginalized identity groups [62–64]. Other biases include those introduced by the use of examples that primarily have English texts and may reflect North American and Western European cultures [65]. We expect models trained on them to reflect these biases, and hence caution developers to assess the limitations of their models before integrating them into user-facing applications. To facilitate positive and safe integration of text-to-video models, we encourage future work to create benchmark evaluations to assess social and cultural biases of these technologies.

While multimodal models can unlock creative applications that can benefit humanity, they can also enable harmful applications. These include surveillance, especially when people are recorded and the recordings are used without their consent, or generation of harmful content, such as pornographic material. A particularly sensitive topic in this space is disinformation. When model outputs achieve realistic quality, they can be used to create convincing fake content (*i.e.*, deepfakes). These can be exploited to spread fake news, defame individuals or portray false situations. To mitigate these harms, watermarks can be applied to every generated video [66] such that it is possible to to identify whether any given video is generated by a particular model.

Due to the impacts and limitations described above, we remark that StoryBench aims to measure progress in text-to-video research. For the same reasons, we do not release our baselines to the public. By no means should our data be extended for use in sensitive domains, but rather for creative goals. We believe that generative technologies like the type of text-to-video models that can be evaluated in StoryBench can become useful tools to enhance human productivity and creativity.

The collection of our datasets has been enabled by the careful work of several participants. Due to privacy concerns, we did not include the estimated hourly wage paid to them or the total amount spent on participant compensation. We feel that individuals' hourly wage or compensation is personal information and we cannot disclose this under privacy law. However, this work was carried out by paid contractors, and we can confirm that they received their standard contracted wage, which is above the living wage in their country of employment.

C Datasheet

Motivation

Q1 **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

StoryBench was created to encourage reproducible progress in text-to-video modeling. Existing video captioning datasets consist of a single sentence describing the salient events that happen throughout the entire video. Existing dense video captioning datasets, instead, are either domain-specific (*e.g.*, instructional video) or contain captions that lack enough information to generate a video. With our annotation protocol, we describe each action separately and also map it to a precise timestamp interval, allowing us to evaluate the ability of text-to-video models to generate arbitrarily long stories. Our task of continuous story visualization is closely related to the existing one of story visualization, which was, however, limited to generate a single key-frame per caption, rather than a continuous video. With the release of StoryBench, we aim to establish a framework for reliable evaluation of forthcoming generative video technologies.

Q2 **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

StoryBench was collected by Google Research.

Q3 **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

Google Research funded the creation of StoryBench.

Q4 **Any other comments?**

No.

Composition

Q5 **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

We provide 8,900 annotations of stories across six splits. Each story contains an object serialized in JSON with the following fields: `sentence_parts`, `start_times`, `end_times`, `original_combined_sentence`, `clip_start_time`, `clip_end_time`, `story_number`, `background_description`, `dataset_name`, `video_name`, `vidln_id`, `question_info`, `num_actors_in_video`, `segment_categories`. We provide a description of each field in the README file of our code online. In addition to our annotations, an instance of the dataset requires the corresponding video file from existing datasets.

Q6 **How many instances are there in total (of each type, if appropriate)?**

The DiDeMo-CSV dev split has 744/744 videos/stories, and the test split has 655/655 videos/stories. The Oops-CSV dev split has 979/1578 videos/stories, and the test split has 979/1578 videos/stories. The UVO-CSV dev split has 1019/1665 videos/stories, and the test split has 1565/2613 videos/stories.

Q7 **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

StoryBench consists of annotations from a subset of dev and test videos from DiDeMo, Oops, and UVO. The annotated videos were selected based on a few criteria: (i) public availability as of February 22, 2023; (ii) lack of inappropriate content; (iii) annotation quality insurance; (iv) preprocessing criteria (*e.g.*, by removing videos whose first action last less than 1.5s).

Q8 **What data does each instance consist of?** *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

We provide raw annotations and corresponding video filenames (text). In addition, we also release the features used to compute our set of automatic metrics for the ground-truth videos.

Q9 **Is there a label or target associated with each instance?** *If so, please provide a description.*

The goal of the dataset is not to classify any given instance. However, we enrich the annotation of each action to easily analyze failure modes by collecting 35 labels across 6 categories (camera movements, foreground entities, foreground actions, background actions, foreground interactions, foreground transitions). We provide the full list of labels in the main body of the paper.

Q10 **Is any information missing from individual instances?** *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

No.

Q11 **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** *If so, please describe how these relationships are made explicit.*

No.

Q12 Are there recommended data splits (e.g., training, development/validation, testing)? *If so, please provide a description of these splits, explaining the rationale behind them.*

Yes. We collect annotations for the existing dev and test splits. We thus recommend using the original training/dev/test splits to avoid any leakage.

Q13 Are there any errors, sources of noise, or redundancies in the dataset? *If so, please provide a description.*

Some videos have multiple stories, which correspond to different instances in our datasets. For data collected in VidLN [20], these correspond to descriptions centered around different actors.

Q14 Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

Our benchmark and the datasets we collected rely on existing video datasets (DiDeMo, Oops, and UVO). We do not provide archival versions of the complete datasets, but the corresponding video resources are publicly available for download from their official websites.

Q15 Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? *If so, please provide a description.*

No.

Q16 Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? *If so, please describe why.*

Our collected annotations have been verified by humans not to contain inappropriate content. Moreover, our annotators flagged videos that contained sensitive data, which were then all discarded. While we did make an attempt to remove inappropriate content, we cannot exclude that a small number of inappropriate samples might have gone unnoticed.

Q17 Does the dataset relate to people? *If not, you may skip the remaining questions in this section.*

Several of our descriptions and corresponding videos are about people. All of the datasets have been verified for sensitive content, and several instances do not include people.

Q18 Does the dataset identify any subpopulations (e.g., by age, gender)?

We do not explicitly collect annotations for any subpopulation. However, it may still be possible to deduce this information from the videos and/or the written descriptions.

Q19 Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? *If so, please describe how.*

Yes, it may be possible to identify people from the videos corresponding to our annotations.

Q20 Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? *If so, please provide a description.*

Yes, our data might be considered sensitive. For instance, the associated videos reveal racial or ethnic origins of people shown in them. However, we note that we removed any videos that were found inappropriate by our annotators.

Q21 Any other comments?

We call for responsible usage of our datasets for research purposes *only* given the potential of text-guided video generation technologies to affect users.

Collection Process

Q22 How was the data associated with each instance acquired? *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly* *inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

We collected human annotations from existing, publicly available video datasets. During the collection campaign, our annotators directly looked at the raw videos. A random sample of the annotations were verified by other humans to ensure high-quality standards.

Q23 **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** *How were these mechanisms or procedures validated?*

We collected human annotations through web user interfaces that we developed. They were validated by manual inspection by us and managers from the company we hired to collect human annotations.

Q24 **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

We annotated all the videos from the evaluation sets of existing datasets that were still available online at the time of our data collection. We also discarded any videos that were found inappropriate.

Q25 **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

We hired a third-party company to collect human annotations from contractors, who received their standard contracted wage, which is above the living wage in their country of employment. The first and last author were also closely involved during the data collection to ensure that the instructions were clear and resolve any doubts raised by the crowdworkers.

Q26 **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?** *If not, please describe the timeframe in which the data associated with the instances was created.*

Our video annotations were collected between December 2022 and March 2023, but the corresponding videos were previously collected by other authors.

Q27 **Were any ethical review processes conducted (e.g., by an institutional review board)?** *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

No institutional review board conducted any ethical review process since we do not modify the original videos, and the datasets providing the videos are publicly available and have previously been published in peer-reviewed journals and conferences.

Q28 **Does the dataset relate to people?** *If not, you may skip the remaining questions in this section.*

Yes, people may appear in our annotations as well as in the corresponding videos.

Q29 **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

We collected annotations from crowdworkers, not from the individuals shown in the original videos.

Q30 **Were the individuals in question notified about the data collection?** *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

Individuals were not notified about our data collection, which involved describing their actions in publicly released videos.

Q31 **Did the individuals in question consent to the collection and use of their data?** *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

We collect annotations from existing, publicly available video datasets. We do not, however, annotate videos that were no longer available online at the time our annotation campaign was conducted, to adhere with the users' intent to remove their content online.

Q32 **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

Users can check whether any of their videos is used in our datasets from the corresponding URLs. If users wish to remove their videos after finding them sensitive, they can contact the hosting party and request to delete the content from the underlying website. Users can also contact us to request removal of the instances in our datasets corresponding to their videos.

Q33 **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

The goal of our datasets is to encourage research towards generative models that can assist and boost artists in generating novel content. However, the resulting technologies could be used to create misinformation online, such as through deepfakes. Yet, we believe that our datasets are the first of their kind to study the problem of generating videos from captions that vary over time. Hence, considering both limitations and opportunities offered by our data, we authorize the dataset for purely academic endeavors.

Q34 **Any other comments?**

No.

Preprocessing, Cleaning, and/or Labeling

Q35 **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** *If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

Yes, we ask human annotators to select which of 34 labels are related to any video segment and captions. We remove any instances (i) whose first action lasts less than 1.5s, or (ii) have a timestamp gap longer than 0.5s between any two consecutive actions.

Q36 **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *If so, please provide a link or other access point to the "raw" data.*

No, we do not save the raw data due to data retention policies in our organization.

Q37 **Is the software used to preprocess/clean/label the instances available?** *If so, please provide a link or other access point.*

Yes, we release our preprocessing scripts on GitHub.

Q38 **Any other comments?**

No.

Uses

Q39 **Has the dataset been used for any tasks already?** *If so, please provide a description.*

Our datasets have not been used for other tasks yet. However, the underlying videos have been used for their original tasks, such as temporal localization with DiDeMo, studying unintentional human action with Oops, and dense, open-world segmentation with UVO. Moreover, VidLN annotations have been used for the tasks of video narrative grounding and video question answering.

Q40 **Is there a repository that links to any or all papers or systems that use the dataset?** *If so, please provide a link or other access point.*

We encourage the community to measure progress in our benchmark and datasets at the URL `https://paperswithcode.com/dataset/storybench`.

Q41 **What (other) tasks could the dataset be used for?**

Our data can be used for the dual task of describing videos over time. In addition, our data could be used to develop automatic evaluation metrics that better align with human preferences. We also believe that the richness of our data will encourage future work to create new, exciting tasks.

Q42 **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

Our annotations describe existing video datasets that might not contain a fair distribution of individuals or groups. For our task, this means that models might be able to generate videos that are biased towards the populations represented in the training data. We encourage future work to extend our efforts towards creating training and evaluation datasets that specifically aim to increase fairness and reduce biases (*e.g.*, correlation between gender, race and jobs) of generative text-to-video models.

Q43 **Are there tasks for which the dataset should not be used?** *If so, please provide a description.*

Under no circumstances should any models developed for our benchmark be used to create deepfakes or any other form of disinformation or harm, including military and surveillance tasks. As it stands, our datasets should solely be used for research purposes.

Q44 **Any other comments?**

No.

Distribution

Q45 **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** *If so, please provide a description.*

Yes, the data will be publicly released.

Q46 **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** *Does the dataset have a digital object identifier (DOI)?*

The data will be available on GitHub.

Q47 **When will the dataset be distributed?**

From September 2023 and onward.

Q48 **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

CC-BY-4.0

Q49 **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

No, our collected annotations are released under a CC-BY-4.0 license. Third-party data are also released publicly.

Q50 **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

No.

Q51 **Any other comments?**

No.

Maintenance

Q52 **Who will be supporting/hosting/maintaining the dataset?**

Google Research will support and maintain the StoryBench annotations on GitHub. The original videos are supported by the corresponding dataset creators or services.

Q53 **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

We can be contacted either via email or through 'pull requests' on the StoryBench GitHub page.

Q54 **Is there an erratum?** *If so, please provide a link or other access point.*

There is no erratum for our first release. Errata will be documented as future releases on GitHub.

Q55 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

No, we do not plan on updating the data. However, we will update the data should there be any errors or requests for deleting specific instances. The updated data will be shared as a 'release' on GitHub.

Q56 **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** *If so, please describe these limits and explain how they will be enforced.*

We do not collect any metadata related to people in creating STORYBENCH. However, we note that our datasets consist of text annotations of existing video datasets. Should people request for their videos to be deleted from the original datasets, we invite them and users to contact us to ensure that the corresponding annotations are removed from STORYBENCH.

Q57 **Will older versions of the dataset continue to be supported/hosted/maintained?** *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

Yes, we will distribute all versions of STORYBENCH as 'releases' on GitHub.

Q58 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*

There is no plan to support and verify third-party contributions that aim at extending the datasets in STORYBENCH as our annotations correspond to standard evaluation splits of existing video datasets. However, we will update the data should there be any errors or requests for deleting specific instances. Dataset versions will be maintained as GitHub releases.

Q59 **Any other comments?**

No.

D Data Preparation Details

In this section, we provide further details on the preparation pipeline used for the evaluation (dev and test) data in STORYBENCH. It consists of the following steps: collection, preprocessing, and rating.

D.1 Data Collection

First, we use an online interface (see Figure 5) to collect stories for each video. Stories consist of multiple sentences, each describing an action, and the corresponding timestamps in the video.

Oops-CSV and UVO-CSV. For VidLN data (UVO and Oops), we provide the original VidLN caption, as well as reference split captions provided from our automatic pipeline (*c.f*. Section 4). In this stage, annotators were instructed to 'split the long sentence into shorter sentences, each describing actions that happen one after the other' and to 'add time stamps for when (the action of) each sentence starts and ends.' Moreover, our annotators are asked to click two checkboxes, whenever applicable: 'Multiple stories,' used to indicate whether the video shown in the user interface actually consists of multiple shorter clips (this is common in Oops, as the data consists of fail video *compilations*); and 'Unimportant actor,' used to indicate whether the original caption describes the events in the video from the perspective of an entity that does not play a salient role in the video (*e.g*., a person in the background). Finally, we perform a second stage of annotations where we provide annotators with the stories from the first stage, and ask them to 'continue the sentence that describes the actor's action in a natural manner by adding a concise context description of relevant actions of other actors.' This second round of annotations was required as VidLN describes the actions of a given entity (actor) throughout a video, which does not often capture the dynamics of the corresponding video segments. Our annotators narrate 2,446 and 2,779 videos from the dev sets of Oops and UVO, respectively.

DiDeMo-CSV. For DiDeMo, we do not possess any descriptions of the video. Instead, the dataset provides detailed text queries (*e.g*., containing camera movement, temporal transition indicators, and activities) that are used to localize events in the video. We provide those queries as a reference to our annotators, and ask them to 'add a description of the background,' 'specify the number of important actors in the video,' 'refine the sentences to create a coherent story,' and 'add timestamps for when (the action of) each sentence starts and ends.' Following the original DiDeMo protocol, each video is trimmed to a maximum of 30 seconds. While the original dev and test sets contained 1,065 and 1,004 videos, respectively, only 843 and 797 videos were still publicly available as of February 22, 2023.

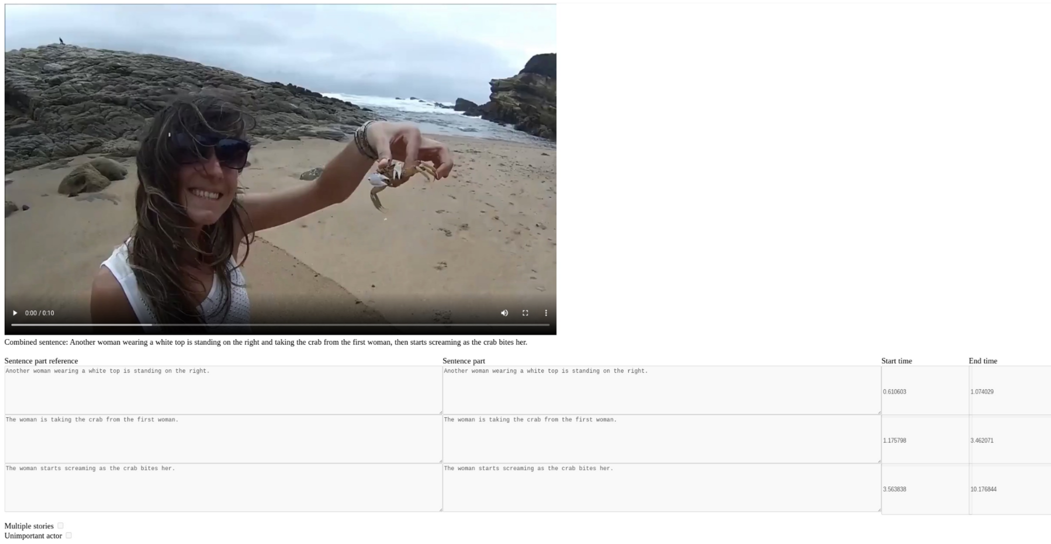

Figure 5: Example of our video annotation interface.

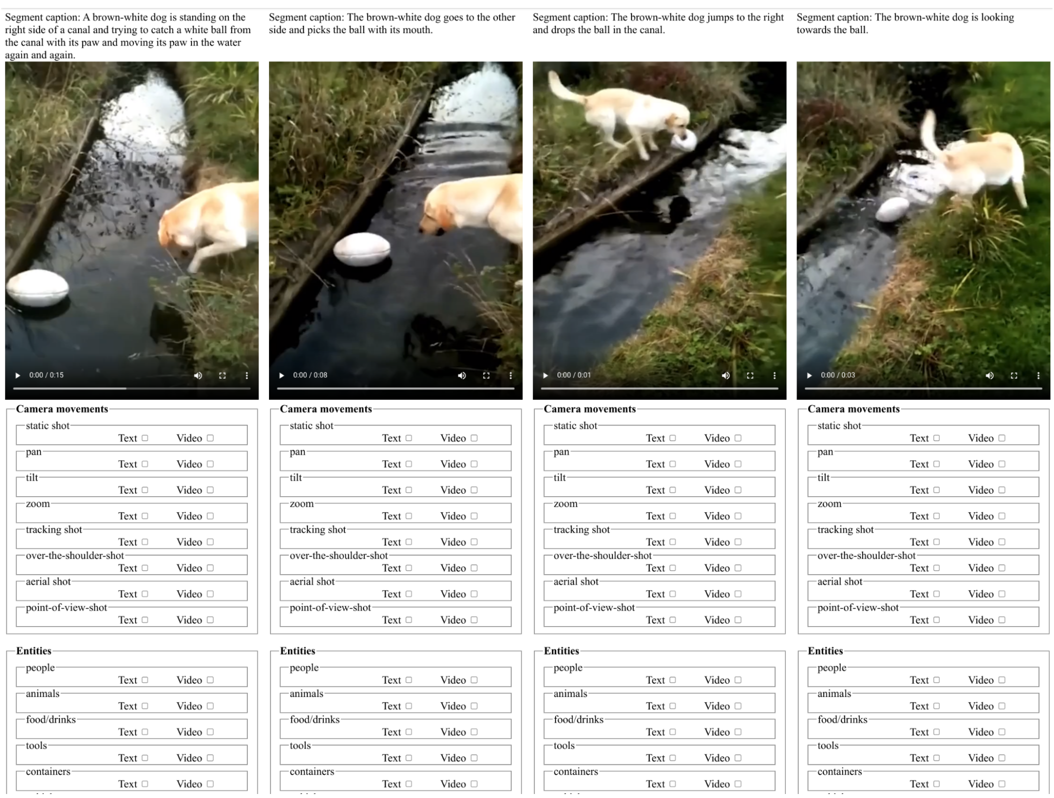

Figure 6: Example of our diagnostic labels collection interface.

In both cases, we provide additional details for both of these tasks to the annotators, as well as examples and corner cases to clearly communicate the desired annotations (available on GitHub). During this collection process, any video flagged by the annotators to contain inappropriate content was removed. Finally, a random sample of our annotations were verified by expert annotators identified by the third party company responsible for human annotations in this project.

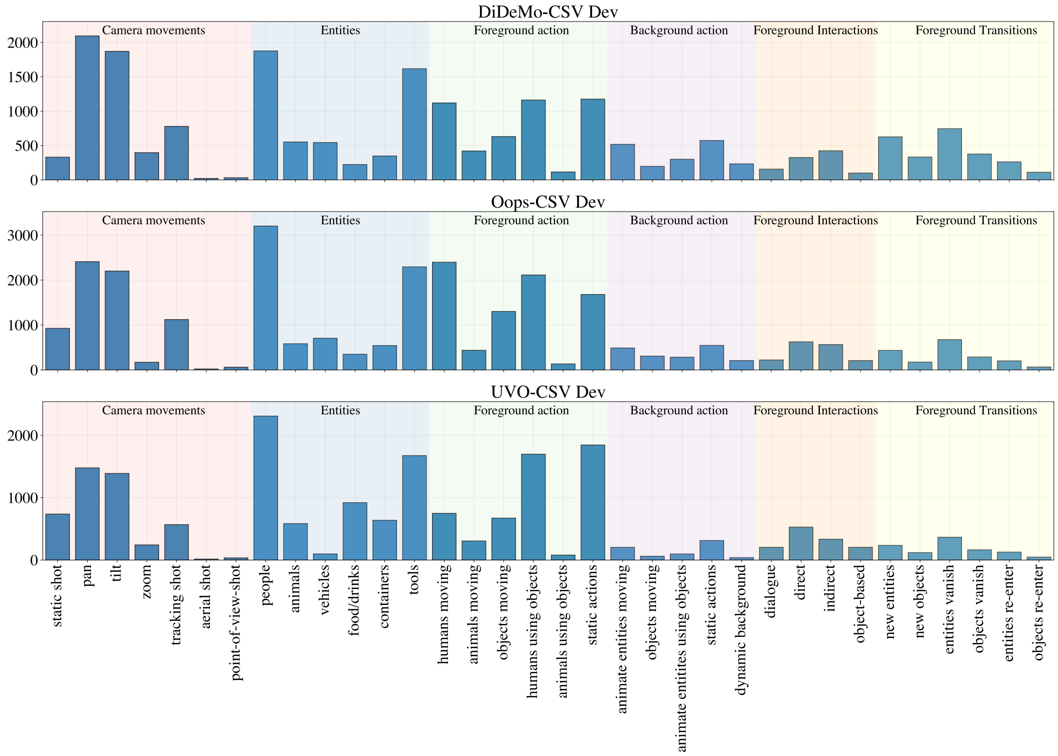

Figure 7: Distribution of collected labels per category in our dev samples.

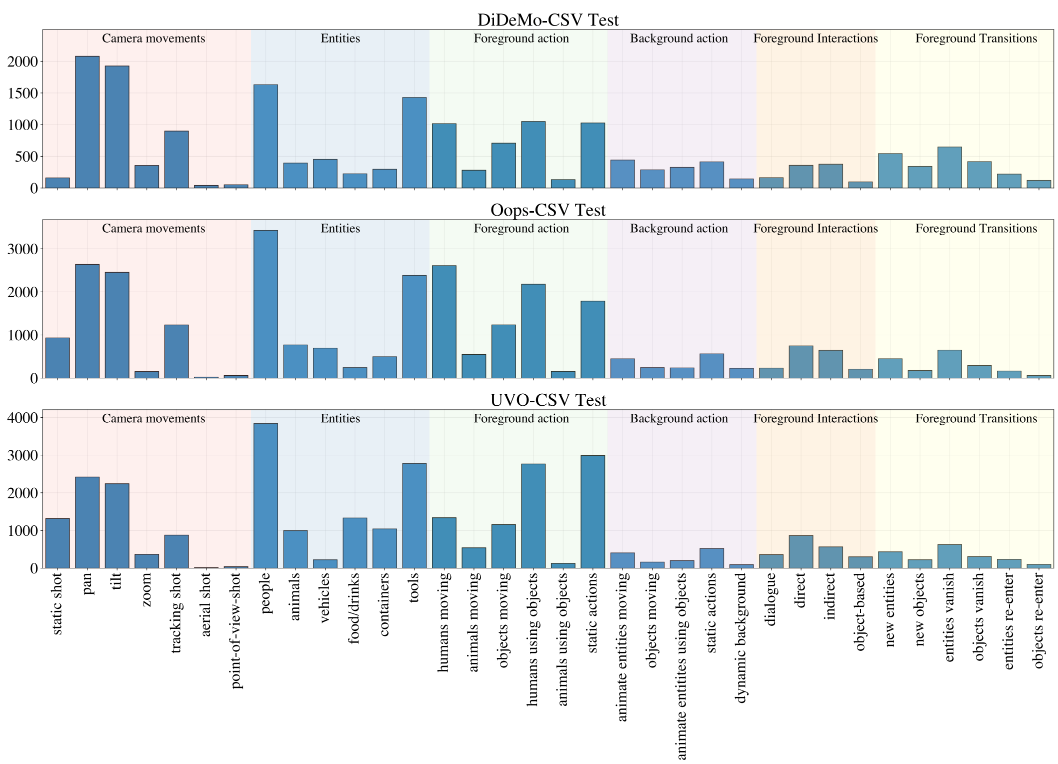

Figure 8: Distribution of collected labels per category in our test samples.

Diagnostic labels. After collecting story annotations for our videos, we enrich them with labels to help analyze the performance of forthcoming text-to-video models along different axes. With the help of artists that have been using generative AI technologies, we define 34 labels across six categories (*c.f.* Section 3). For each video segment, we then ask our annotators to tick two checkboxes per label: 'Text' if the label is mentioned in the segment caption; and 'Video' if the label is shown in the video. Figure 6 shows an example of the UI used in this process, and we release our full set of instructions online. Figures 7 and 8 show the distribution of labels in our preprocessed (*i.e.*, final) data.

Human annotation framework. We assess the high quality of our annotations as follows. First, annotators were only moved to the final data annotation process after having successfully completed a training stage. Second, we asked annotators' managers to verify the quality (*i.e.*, descriptions and timestamps match the content of the videos) of the final data by manually checking 25% of the data samples. Here, they found 97% of the samples to accurately reflect the narratives of the videos.

D.2 Data Preprocessing

Given the above collected annotations, we perform the following two preprocessing steps. First, we only keep stories whose first action is at least 1.5s long; as we use a video of 0.5s to condition text-to-video generation for the task of *story continuation*. Second, we remove any story in which two subsequent actions have more than 0.5s gap.

Figures 9 and 10 shows our final data distributions. For DiDeMo-CSV, the dev split has 744/744 videos/stories, while the test split has 655/655 videos/stories. For Oops-CSV, the dev split has 979/1578 videos/stories, while the test split has 979/1578 videos/stories. For UVO-CSV, the dev split has 1019/1665 videos/stories, while the test split has 1565/2613 videos/stories.

The preprocessed data is then adapted for each of our evaluation tasks detailed in Section 3: *action execution*, *story continuation*, and *story generation*. Finally, to evaluate our baselines, the original videos are downsampled to 8 frames per second (fps) using the 'FFmpeg' open-source software.

Figures 12 to 14 show examples of the resulting data.

D.3 Human Evaluation

Human evaluation is the preferred way to assess the capabilities of generative models. We perform side-by-side comparisons between two models, and ask human raters to choose the one (if any) that performs better according to the five criteria that we defined in Section 3. Figure 15 shows an example of the user interface developed for human evaluation.

D.4 Automatic Evaluation

Section 3.4 introduces our automatic evaluation metrics. Here, we provide our intuition of how we expect them to relate to our human evaluation metrics. **FID** would measure "visual quality" since it compares the distribution of ground-truth frames with that of generated frames. **FVD** would measure "entity consistency" and "action realism" since it compares the distribution of ground-truth videos with that of generated videos. **SIM** would measure "visual quality", "entity consistency", "background consistency", and "text adherence" as it compares ground-truth and generated frames one-to-one. **VTM** would measure "text adherence" as it compares generated videos to their prompts. **PQA** would measure "visual quality" and "action realism" as it was trained to predict the average human subjective perception of a video.

D.5 Robustness of Automatic Pipeline

In Section 4, we define an automatic pipeline to transform the original VidLN captions for Oops and UVO into multiple sentences, each approximately describing a single action, and to estimate their corresponding timestamps. Here, we compute some statistics to assess the quality of the stories generated automatically through our algorithmic pipeline by comparing them against human references for the Oops Dev set (1,578 stories).

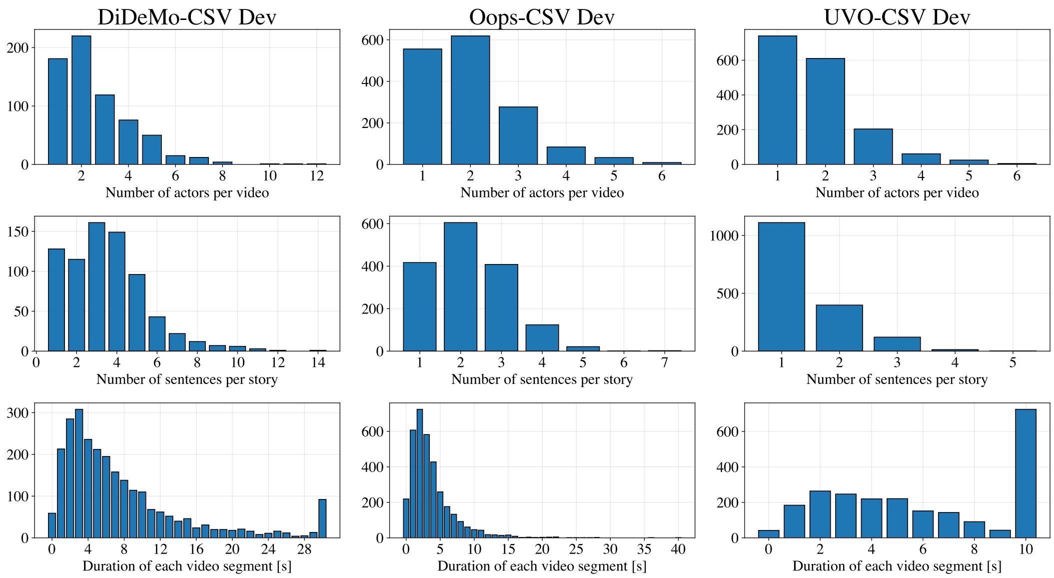

Figure 9: Statistics of our final dev sets.

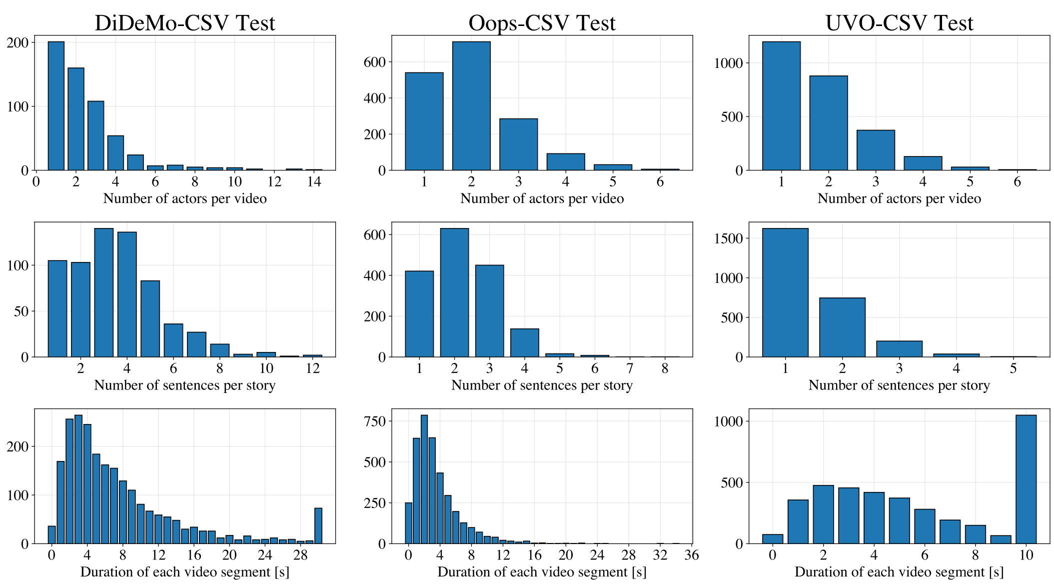

Figure 10: Statistics of our final test sets.

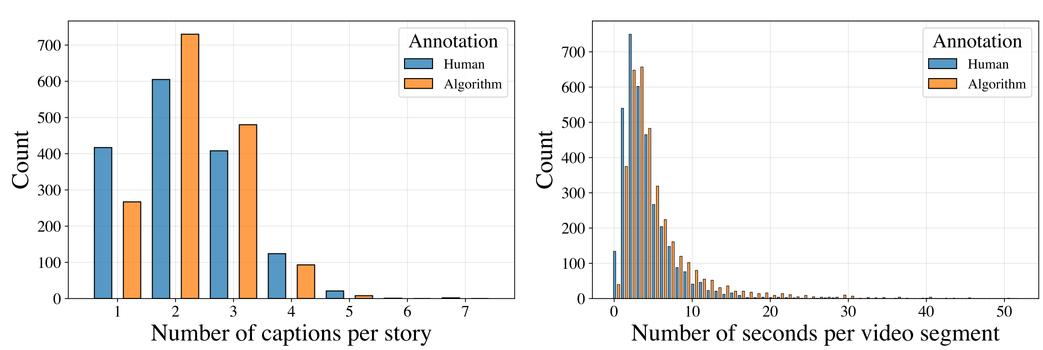

Figure 11: Robustness statistics of automated pipeline for story-like data transformation. Left: Number of captions per story. Right: Duration of video segments in seconds.

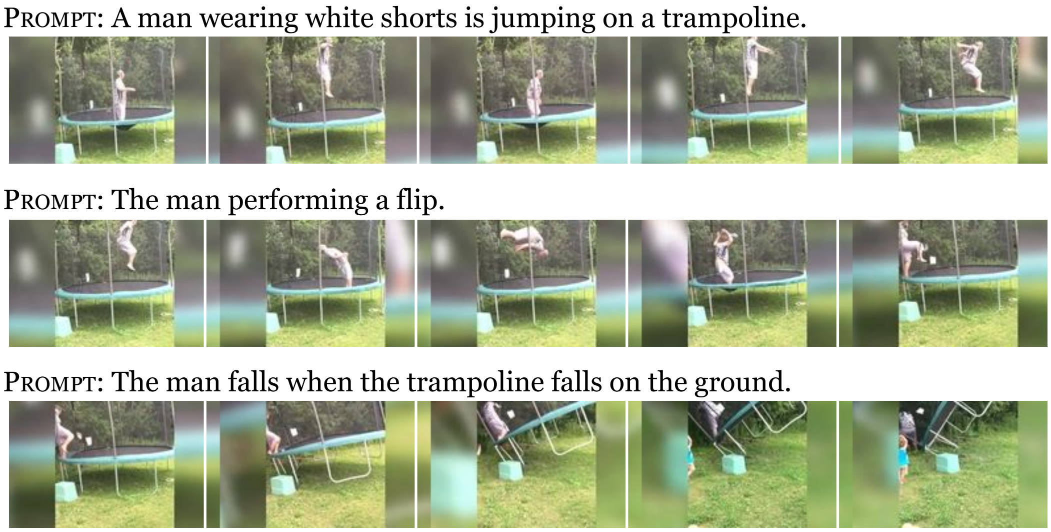

Figure 12: Example story (subsampled frames) from Oops-CSV.

Figure 13: Example story (subsampled frames) from UVO-CSV.

As shown in Figure 11 (left), the distributions of the number of sentences per story of the two approaches are very similar. In particular, we notice that our method tends to split captions into two or three segments more often than the human annotators, who, more often, prefer not to split them.

The words corresponding to each video segment are very similar between human and automatic stories. To assess this, we consider the subset of captions that have been split into the same number of sentences, so we can compute one-to-one mappings between the human and the algorithmic captions. Here, we observe a BLEU_4 score of 63.6% (BLEU_4 measures the overlap of 4-grams in the two captions), indicating a relatively high similarity of generated sentences to human references. We also note that humans were asked to enrich the original Oops and UVO captions with context information (*e.g.*, what other relevant actors are doing while a specific actor is being narrated), which our algorithmic pipeline does not explicitly tackle, leaving room for improvement in future work.

Finally, Figure 11 (right) shows that the resulting duration of the algorithmically generated video segments are slightly longer than human-annotated timestamps.

PROMPT: A white-brown dog is sitting and starts moving towards the person.

PROMPT: A person whose only hand and leg is visible is holding some food in his hand.

PROMPT: The white-brown dog is take food from the person hand and eats, while the camera focus on the dog face.

PROMPT: The person starts rubbing the dog head with his hand.

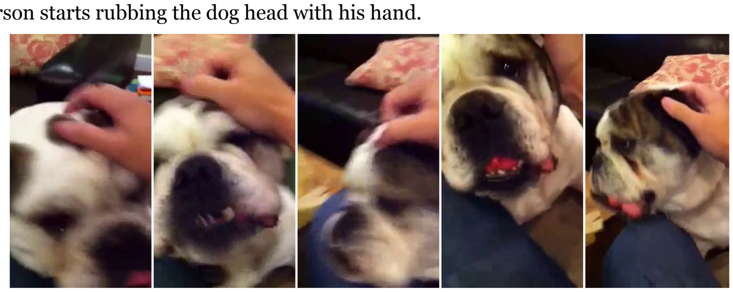

Figure 14: Example story (subsampled frames) from DiDeMo-CSV.

E Additional Results

In this section, we report our full set of results from our baselines on STORYBENCH, in terms of both human evaluations and through automatic metrics. Recall that we append -ZS for results obtained in the *zero-shot* setting, -ST for *single-task* fine-tuning, and -MT for *multi-task* fine-tuning. Each model was fine-tuned for 500K steps in less than a day on 4x4x4 TPUv4 chips. For every story, each model generates 4 output videos at 8fps using a 160×96 pixel resolution. We randomly sample one of them for human evaluation (*e.g.*, Figure 16), but report mean and standard deviation for automatic metrics.

E.1 Human Evaluation

Figure 17 shows the results of human evaluation, where each bar displays the number of wins of two given models evaluated side-by-side, as well as the number of ties (in white). For each story, we ask three human raters to compare two models and report the majority vote in Figure 17.

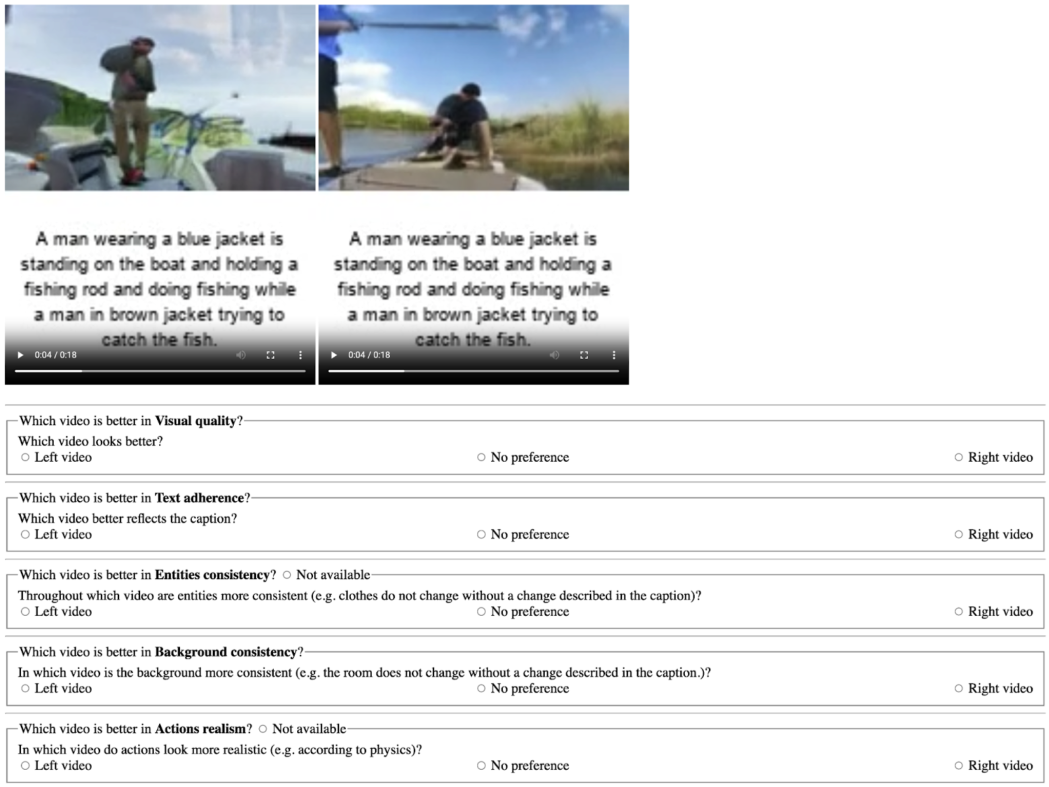

Figure 15: Example of our human rating interface.

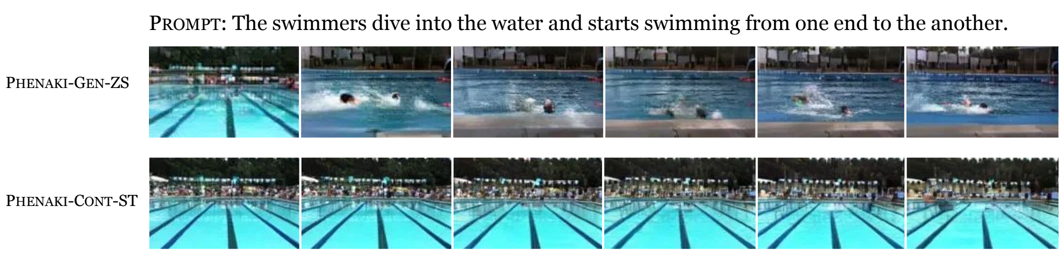

Figure 16: Example of generated actions by PHENAKI-GEN-ST and PHENAKI-CONT-ST on DiDeMo-CSV. PHENAKI-GEN-ST quickly changes the background, while PHENAKI-CONT-ST correctly synthesizes a person swimming left-to-right without distorting the background. Video subsampled by a factor 4 to be shown here.

Looking at task of *action execution* on Oops-CSV, we see that our PHENAKI-CONT-ST achieves competitive performance with our PHENAKI-GEN-ZS baseline, with better text adherence, background consistency and action realism. This result is not surprising as most of the actions in Oops are short (less than 5s). It is interesting, however, to see that our annotators find PHENAKI-CONT-ST largely better than PHENAKI-GEN-ST across all criteria. On the other hand, none of these models clearly outperforms others for the most challenging task of *story generation*.

For the task of *story continuation*, PHENAKI-CONT-ST typically outperforms both PHENAKI-GEN-ZS and PHENAKI-GEN-ST, especially on Oops-CSV. On UVO-CSV and DiDeMo-CSV, PHENAKI-CONT-ST consistently outperforms PHENAKI-GEN-ST except for entity and background consistency, where human raters often have no preference between the two. Comparing multi-task models, we find that PHENAKI-CONT-MT is always preferred to PHENAKI-GEN-MT; yet PHENAKI-GEN-ZS is a strong baseline, achieving better visual quality than the fine-tuned models.

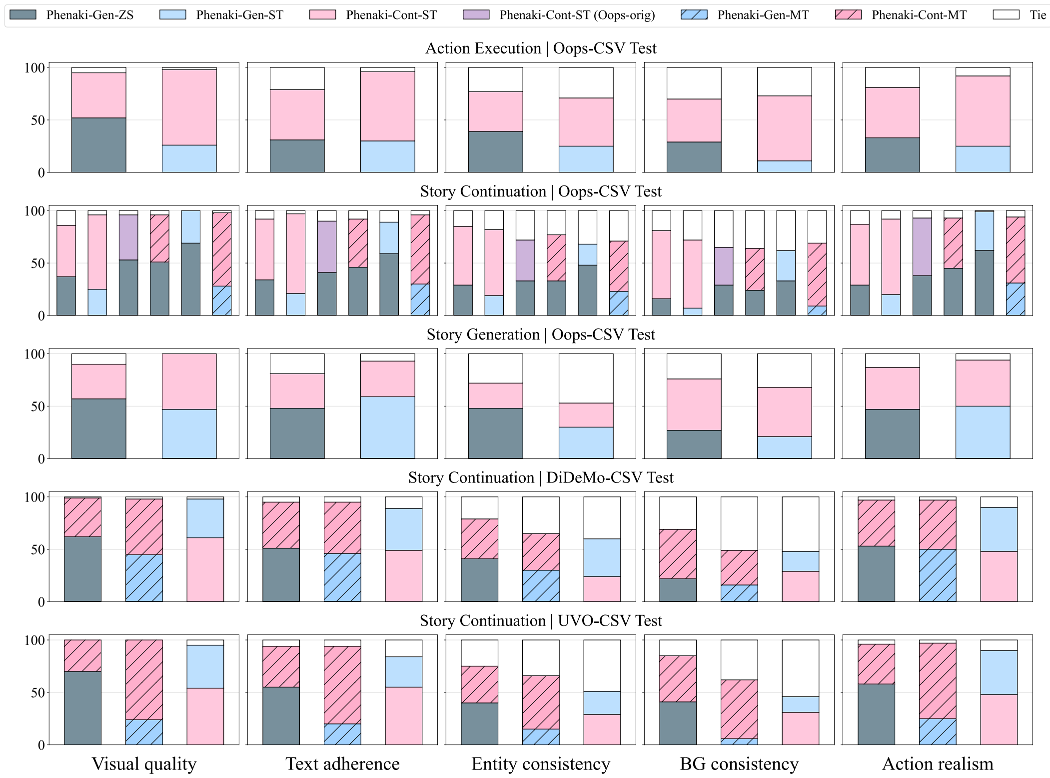

Figure 17: Results from human evaluation across datasets and tasks.

E.2 Automatic Evaluation

For completeness, Tables 8 to 10 report the performance of our baselines on all tasks and datasets when instead using CLIP to compute FID and SIM, and InternVideo to compute FVD and VTM. We find similar patterns as with other metrics (*c.f*. Section 6), but also notice that InternVideo (used for FVD and VTM) favors the videos generated by the zero-shot PHENAKI-GEN model.

Action Execution	Oops-CSV					UVO-CSV					DiDeMo-CSV				
Model (@8 fps)	FID_C↓	FVD_{IV}↓	SIM_C↑	PQA↑	VTM_{IV}↑	FID_C↓	FVD_{IV}↓	SIM_C↑	PQA↑	VTM_{IV}↑	FID_C↓	FVD_{IV}↓	SIM_C↑	PQA↑	VTM_{IV}↑
Zero-shot															
Phenaki-Gen-ZS	94.7$_{\pm 0.52}$	**126.7**$_{\pm 0.46}$	64.9$_{\pm 0.08}$	**5.8**$_{\pm 0.03}$	**22.6**$_{\pm 0.07}$	79.2$_{\pm 0.38}$	**85.3**$_{\pm 0.41}$	66.7$_{\pm 0.03}$	**8.5**$_{\pm 0.10}$	**23.0**$_{\pm 0.03}$	97.2$_{\pm 0.34}$	**78.0**$_{\pm 0.25}$	64.3$_{\pm 0.08}$	**6.7**$_{\pm 0.02}$	**22.9**$_{\pm 0.05}$
Single-Task															
Phenaki-Gen-ST	97.1$_{\pm 0.20}$	179.4$_{\pm 0.28}$	64.8$_{\pm 0.04}$	4.0$_{\pm 0.02}$	20.0$_{\pm 0.02}$	97.3$_{\pm 0.18}$	147.6$_{\pm 0.20}$	62.5$_{\pm 0.04}$	4.8$_{\pm 0.05}$	18.4$_{\pm 0.02}$	89.3$_{\pm 0.22}$	120.2$_{\pm 0.59}$	64.9$_{\pm 0.04}$	4.5$_{\pm 0.02}$	20.4$_{\pm 0.03}$
Phenaki-Cont-ST	**84.5**$_{\pm 0.02}$	171.6$_{\pm 0.58}$	**67.9**$_{\pm 0.04}$	4.8$_{\pm 0.02}$	19.9$_{\pm 0.01}$	92.4$_{\pm 0.27}$	143.2$_{\pm 0.31}$	64.0$_{\pm 0.09}$	5.6$_{\pm 0.02}$	18.7$_{\pm 0.03}$	**82.5**$_{\pm 0.22}$	107.3$_{\pm 0.46}$	**66.8**$_{\pm 0.01}$	5.6$_{\pm 0.01}$	20.1$_{\pm 0.01}$
Multi-Task															
Phenaki-Gen-MT	102.8$_{\pm 0.64}$	179.3$_{\pm 0.63}$	63.7$_{\pm 0.04}$	3.8$_{\pm 0.04}$	20.1$_{\pm 0.04}$	92.1$_{\pm 0.68}$	138.9$_{\pm 0.42}$	63.3$_{\pm 0.03}$	5.1$_{\pm 0.05}$	19.2$_{\pm 0.02}$	88.2$_{\pm 0.44}$	119.6$_{\pm 0.47}$	64.3$_{\pm 0.04}$	4.7$_{\pm 0.06}$	20.1$_{\pm 0.02}$
Phenaki-Cont-MT	86.0$_{\pm 0.52}$	171.3$_{\pm 0.56}$	67.4$_{\pm 0.11}$	4.7$_{\pm 0.01}$	20.1$_{\pm 0.03}$	**77.9**$_{\pm 0.08}$	126.8$_{\pm 0.24}$	**67.1**$_{\pm 0.06}$	6.8$_{\pm 0.01}$	19.9$_{\pm 0.02}$	85.4$_{\pm 0.14}$	106.5$_{\pm 0.29}$	66.4$_{\pm 0.07}$	5.8$_{\pm 0.03}$	19.9$_{\pm 0.03}$

Table 8: Results from automatic evaluation metrics on *action execution* tasks. Best results are in **bold**. FID and SIM use CLIP, FVD and VTM use InternVideo, and PQA uses DOVER.

Story Continuation	Oops-CSV					UVO-CSV					DiDeMo-CSV				
Model (@8 fps)	FID_C↓	FVD_{IV}↓	SIM_C↑	PQA↑	VTM_{IV}↑	FID_C↓	FVD_{IV}↓	SIM_C↑	PQA↑	VTM_{IV}↑	FID_C↓	FVD_{IV}↓	SIM_C↑	PQA↑	VTM_{IV}↑
Zero-shot															
Phenaki-Gen-ZS	103.2$_{\pm 0.87}$	**116.6**$_{\pm 0.54}$	63.1$_{\pm 0.05}$	**7.2**$_{\pm 0.06}$	**22.5**$_{\pm 0.05}$	82.2$_{\pm 0.51}$	**83.6**$_{\pm 0.44}$	65.9$_{\pm 0.04}$	**9.4**$_{\pm 0.09}$	**22.9**$_{\pm 0.03}$	108.2$_{\pm 0.43}$	**87.9**$_{\pm 0.41}$	61.7$_{\pm 0.04}$	**7.3**$_{\pm 0.07}$	**22.5**$_{\pm 0.10}$
Single-Task															
Phenaki-Gen-ST	99.6$_{\pm 0.07}$	181.5$_{\pm 0.74}$	64.0$_{\pm 0.05}$	4.3$_{\pm 0.02}$	19.7$_{\pm 0.02}$	97.8$_{\pm 0.26}$	151.1$_{\pm 0.21}$	62.5$_{\pm 0.03}$	5.0$_{\pm 0.03}$	18.2$_{\pm 0.01}$	90.9$_{\pm 0.18}$	127.5$_{\pm 0.48}$	64.2$_{\pm 0.03}$	4.0$_{\pm 0.02}$	19.9$_{\pm 0.02}$
Phenaki-Cont-ST	**89.2**$_{\pm 0.30}$	169.9$_{\pm 0.67}$	**66.3**$_{\pm 0.05}$	5.3$_{\pm 0.04}$	19.5$_{\pm 0.02}$	94.1$_{\pm 0.29}$	147.3$_{\pm 0.75}$	63.5$_{\pm 0.07}$	5.7$_{\pm 0.03}$	18.3$_{\pm 0.02}$	**89.4**$_{\pm 0.29}$	118.1$_{\pm 0.60}$	**64.5**$_{\pm 0.05}$	5.4$_{\pm 0.03}$	19.4$_{\pm 0.06}$
Multi-Task															
Phenaki-Gen-MT	105.2$_{\pm 0.26}$	182.0$_{\pm 0.17}$	63.0$_{\pm 0.04}$	4.2$_{\pm 0.02}$	19.8$_{\pm 0.02}$	92.8$_{\pm 0.58}$	141.9$_{\pm 0.22}$	63.1$_{\pm 0.03}$	5.2$_{\pm 0.03}$	18.9$_{\pm 0.02}$	90.7$_{\pm 0.24}$	126.4$_{\pm 0.82}$	63.4$_{\pm 0.02}$	4.3$_{\pm 0.04}$	19.7$_{\pm 0.04}$
Phenaki-Cont-MT	92.1$_{\pm 0.44}$	171.4$_{\pm 0.67}$	65.7$_{\pm 0.09}$	5.1$_{\pm 0.02}$	19.8$_{\pm 0.02}$	**80.4**$_{\pm 0.17}$	129.0$_{\pm 0.65}$	**66.3**$_{\pm 0.09}$	7.0$_{\pm 0.02}$	19.6$_{\pm 0.05}$	95.5$_{\pm 0.33}$	120.2$_{\pm 0.23}$	63.4$_{\pm 0.07}$	5.5$_{\pm 0.09}$	19.0$_{\pm 0.01}$

Table 9: Results from automatic evaluation metrics on *story continuation* tasks. Best results are in **bold**. FID and SIM use CLIP, FVD and VTM use InternVideo, and PQA uses DOVER.

Story Generation	Oops-CSV					UVO-CSV					DiDeMo-CSV				
Model (@8 fps)	FID_C↓	FVD_{IV}↓	SIM_C↑	PQA↑	VTM_{IV}↑	FID_C↓	FVD_{IV}↓	SIM_C↑	PQA↑	VTM_{IV}↑	FID_C↓	FVD_{IV}↓	SIM_C↑	PQA↑	VTM_{IV}↑
Zero-shot															
Phenaki-Gen-ZS	117.2$_{\pm 0.90}$	**113.0**$_{\pm 0.54}$	N/A	**8.1**$_{\pm 0.03}$	**22.9**$_{\pm 0.11}$	97.5$_{\pm 0.89}$	**88.2**$_{\pm 0.68}$	N/A	**10.0**$_{\pm 0.06}$	**22.6**$_{\pm 0.14}$	115.6$_{\pm 0.38}$	**91.1**$_{\pm 0.46}$	N/A	**7.6**$_{\pm 0.08}$	**23.0**$_{\pm 0.06}$
Single-Task															
Phenaki-Gen-ST	**103.4**$_{\pm 0.28}$	181.4$_{\pm 0.19}$	N/A	4.2$_{\pm 0.02}$	19.6$_{\pm 0.03}$	99.2$_{\pm 0.22}$	151.7$_{\pm 0.45}$	N/A	4.9$_{\pm 0.01}$	18.0$_{\pm 0.01}$	92.3$_{\pm 0.07}$	128.6$_{\pm 0.44}$	N/A	4.0$_{\pm 0.01}$	20.1$_{\pm 0.04}$
Phenaki-Cont-ST	109.1$_{\pm 0.89}$	167.9$_{\pm 0.95}$	N/A	5.4$_{\pm 0.03}$	17.9$_{\pm 0.03}$	100.6$_{\pm 0.47}$	149.6$_{\pm 0.24}$	N/A	5.4$_{\pm 0.02}$	17.4$_{\pm 0.02}$	96.1$_{\pm 0.11}$	124.6$_{\pm 0.86}$	N/A	5.4$_{\pm 0.05}$	18.9$_{\pm 0.02}$
Multi-Task															
Phenaki-Gen-MT	107.4$_{\pm 0.71}$	180.0$_{\pm 0.52}$	N/A	4.1$_{\pm 0.02}$	19.8$_{\pm 0.06}$	**94.8**$_{\pm 0.25}$	143.0$_{\pm 0.28}$	N/A	5.1$_{\pm 0.03}$	18.8$_{\pm 0.01}$	**92.0**$_{\pm 0.20}$	127.3$_{\pm 0.46}$	N/A	4.3$_{\pm 0.06}$	19.8$_{\pm 0.01}$
Phenaki-Cont-MT	114.3$_{\pm 0.12}$	171.0$_{\pm 0.53}$	N/A	5.0$_{\pm 0.11}$	18.0$_{\pm 0.06}$	99.8$_{\pm 0.28}$	132.7$_{\pm 0.52}$	N/A	6.3$_{\pm 0.08}$	17.5$_{\pm 0.05}$	105.6$_{\pm 0.26}$	125.9$_{\pm 1.03}$	N/A	5.2$_{\pm 0.03}$	18.3$_{\pm 0.06}$

Table 10: Results from automatic evaluation metrics on *story generation* tasks. Best results are in **bold**. FID uses CLIP, FVD and VTM use InternVideo, and PQA uses DOVER.