

# SUPPLEMENTARY MATERIAL: OPENVID-1M: A LARGE-SCALE HIGH-QUALITY DATASET FOR TEXT-TO-VIDEO GENERATION

**Anonymous authors**

Paper under double-blind review

## 1 EXAMPLES OF OpenVid-1M DATASET

In Figure 1, we visualize some samples from our OpenVid-1M. We randomly select samples with  $512 \times 512$  and  $1920 \times 1080$  resolution, respectively. With well designed data processing pipeline, OpenVid-1M demonstrates superior quality and descriptive richness, particularly in aesthetics, motion, temporal consistency, caption length and clarity as well.

## 2 FILTERING RATIOS OF DATASET PROCESSING PIPELINE

We randomly sampled a subset from the collected raw data and processed it through our data processing pipeline. A panel of evaluators was then tasked with assessing these video subsets, determining whether the videos at each processing step met our requirements. Based on their preferences, we derived score thresholds and filtering ratios for each step after multiple evaluations. Figure 2 provides visualizations of the videos with varying clarity, aesthetic, motion, and temporal consistency scores computed by our pipeline.

## 3 MORE TEXT-TO-VIDEO EXAMPLES

We present more visual results of our model. As depicted in Figure 3, the first column illustrates our model’s proficiency in generating aesthetically pleasing content with a painting style. The second column showcases the superior text alignment of the videos generated by our model, accurately depicting ‘crashed down’ from the text. The third and fourth columns highlight our model’s ability to produce intricate dynamics and motions, e.g., ‘motorcycle race’ and ‘gallop across’.

## 4 VIDEO DURATIONS COMPARISON WITH OTHER DATASETS

We present video durations comparison between our OpenVid-1M and other million level text-to-video datasets in Figure 4. Specifically, OpenVid-1M consists of 1,019,957 clips, averaging 7.2 seconds each, with a total video length of 2,051 hours. Compared to previous million-level datasets, WebVid-10M contains low-quality videos with watermarks and Panda-70M contains many still, flickering, or blurry videos along with short captions. In contrast, our OpenVid-1M contains high-quality, clean videos with dense and expressive captions.



Figure 1: Examples of OpenVid-1M dsdataset.

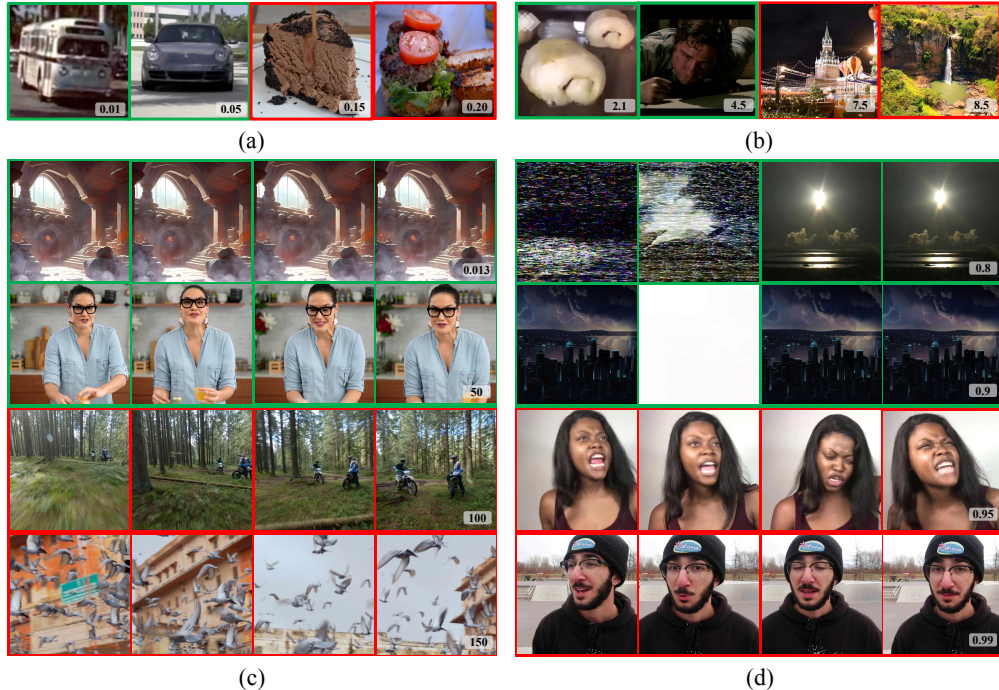


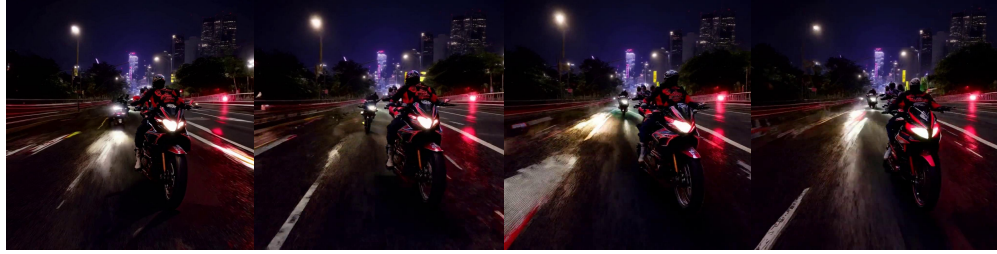
Figure 2: Visualizations of the videos with varying (a) clarity, (b) aesthetic, (c) motion, and (d) temporal consistency scores.



*“Unicorn sliding on a rainbow.”*



*“A snow avalanche crashed down mountain peak, causing destruction and mayhem.”*



*“A motorcycle race through the city streets night.”*



*“Three horses gallop across a wide open field, tails and manes flying in the wind.”*

Figure 3: More text-to-video showcases.

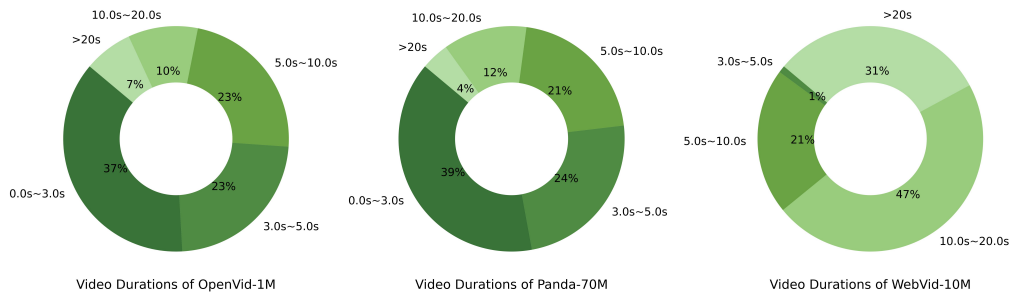


Figure 4: Comparisons on video durations with previous million level text-to-video datasets.