LVD-2M: A Long-take Video Dataset with Temporally Dense Captions
Supplementary Material

Table of Contents

A Qualitative Results for Model Finetuning 1

B Qualitative Evaluation for Long Range Video Generation 1

C Statistics of LVD-2M and Previous Datasets 3

D Prompt Design for LLaVA and Claude3-Haiku 6

E Author Statements 7

A Qualitative Results for Model Finetuning

In this section, we present additional qualitative results to demonstrate the effectiveness of finetuning a diffusion-based image-to-video (I2V) model.

Setup. To compare the effect of LVD-2M to previous datasets on long video generation finetuning, we finetune the same pretrained diffusion-based I2V model separately on WebVid-10M [1] and LVD-2M. Both datasets are used to finetune the model for generating 65-frame videos, with the finetuning process running for 20k iterations using identical strategies.

Analysis. We identify two advantages of finetuning with LVD-2M compared to WebVid-10M. First, the camera perspective presents more variation, including translation (Fig. 1) and tracking shots around the main object (Fig. 2). In contrast, after finetuning on 65 frames on WebVid-10M, the generated videos are prone to simply repeating the first frame with small variation. Second, there are fewer significant inconsistent transitions after finetuning on LVD-2M. As shown in Fig. 3, after finetuning on WebVid-10M, the generated videos may abruptly change into white and black mask frames. This phenomenon results from the WebVid training data, where such abrupt transitions are observed for 3D art style videos. For LVD-2M, videos with such transitions are filtered out by our scene cut detection algorithm. And such cases are less observed in the videos generated by the model finetuned on LVD-2M. We also demonstrate I2V results on longer text prompts, as shown in Fig. 4.

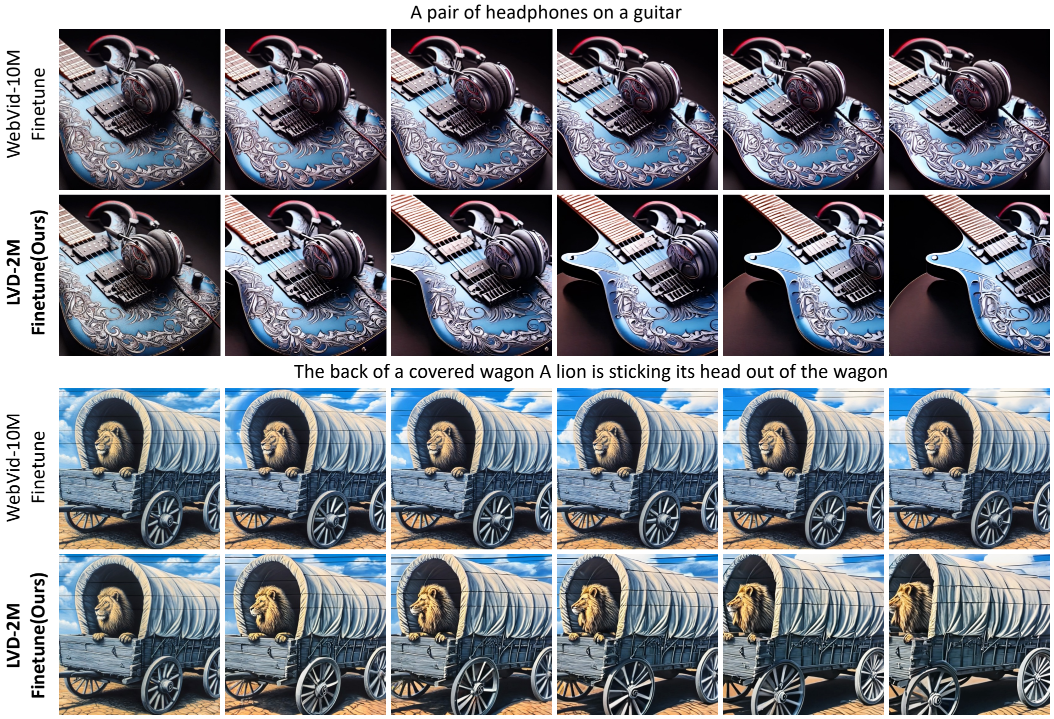
B Qualitative Evaluation for Long Range Video Generation

In this section, we present experiments about generating long videos after finetuning the LM-based T2V model on LVD-2M. We choose LM-based model because it can naturally extend the video generation to longer range by directly conditioning on previous generated frames. We also finetune the same pretrained LM-based T2V model on WebVid-10M [1] as the baseline.

Setup. We finetune the same LM-based model on LVD-2M and WebVid-10M separately on 65 frames (∼10s long) for 10k iterations. Due to a lack of wide accepted long-range video generation

The dataset homepage is https://github.com/SilentView/LVD-2M.

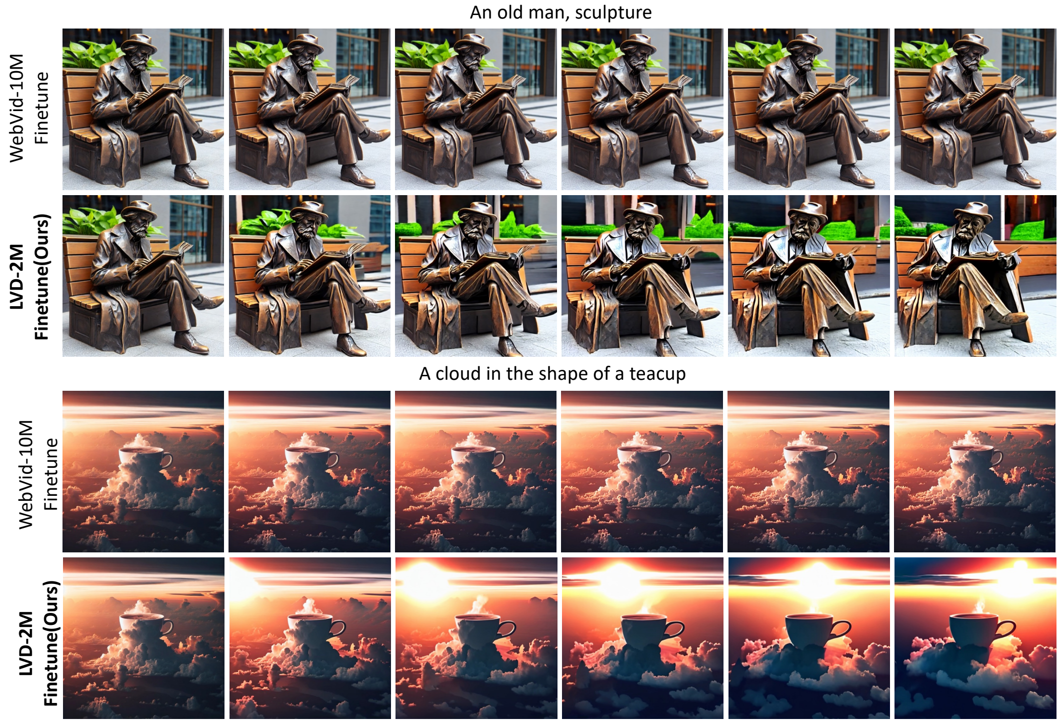
Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.



A pair of headphones on a guitar

The back of a covered wagon A lion is sticking its head out of the wagon

Figure 1: After finetuning on LVD-2M, the camera perspective will present more translation, compared to WebVid-10M.



An old man, sculpture

A cloud in the shape of a teacup

Figure 2: After finetuning on LVD-2M, the camera view rotates more often and will present more view points, compared to WebVid-10M.

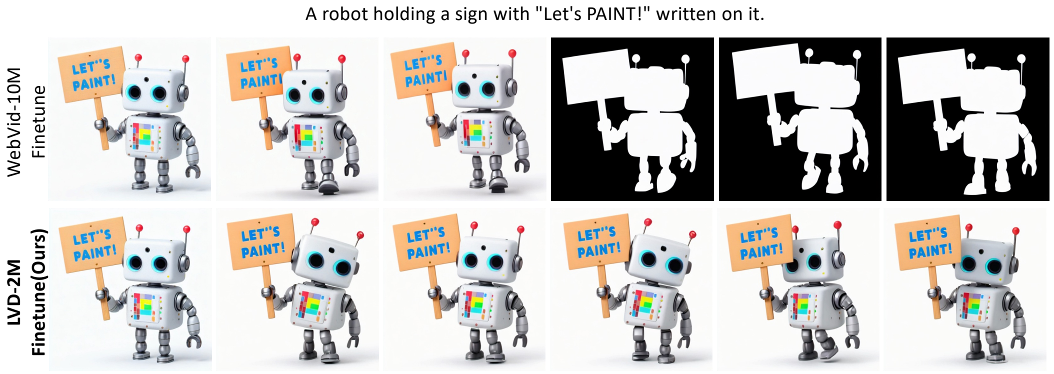


Figure 3: The problem of abrupt transition into black-white mask frames are less observed after finetuning on LVD-2M.

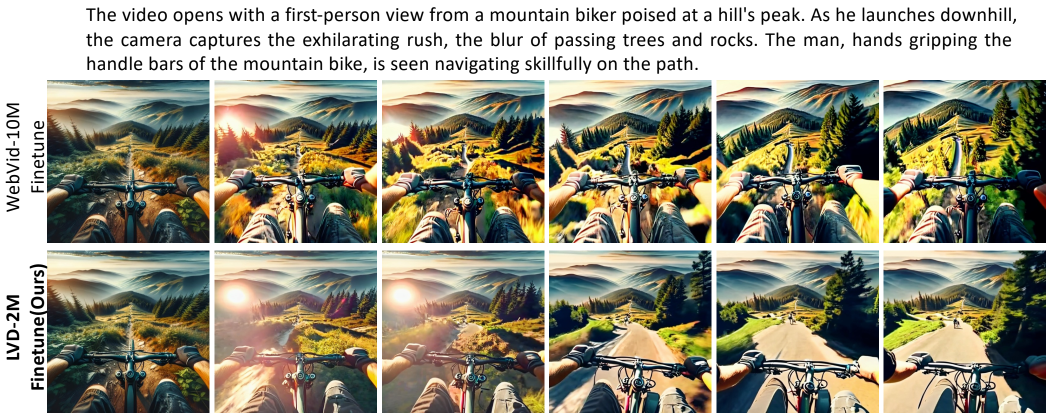

The video opens with a first-person view from a mountain biker poised at a hill's peak. As he launches downhill, the camera captures the exhilarating rush, the blur of passing trees and rocks. The man, hands gripping the handle bars of the mountain bike, is seen navigating skillfully on the path.
WebVid-10M Finetune
LVD-2M Finetune(Ours)

Figure 4: Finetuning on LVD-2M will further improve the capability of the model to generate more dynamic content, compared to WebVid-10M.

benchmark, we choose to qualitatively evaluate the finetuned models. For more details about the architecture of the LM-based model, please refer to the anonymous paper *Loong: Generating Minute-level Long Videos with Autoregressive Language Models* in our supplementary files.

Analysis. We provide a comparison of the generated videos from models finetuned on LVD-2M and WebVid-10M, as shown in Figure 5. The model finetuned on LVD-2M can generate larger motions and more diverse visual elements compared to the one finetuned on WebVid-10M. This demonstrates the effectiveness of LVD-2M in enhancing the model's capability to produce highly dynamic and engaging video content.

C Statistics of LVD-2M and Previous Datasets

In this section, we compare the dataset statistics with the source datasets of ours: WebVid-10M [1], Panda-70M [2], InternVid [3] and HD-VG [4].

Fig. 6 demonstrates the distribution of duration of the video clips. Among previous datasets, WebVid has larger portion of long videos, mainly because its videos are directly collected from stock footage providers. For other datasets whose videos are from YouTube, short video clips (<10s) almost dominate the datasets. Compared to previous datasets, LVD-2M focuses on video clips longer than 10s, resulting in the collected video clips being significantly longer. This feature of LVD-2M can be useful for learning long-range temporal modeling for video generation.

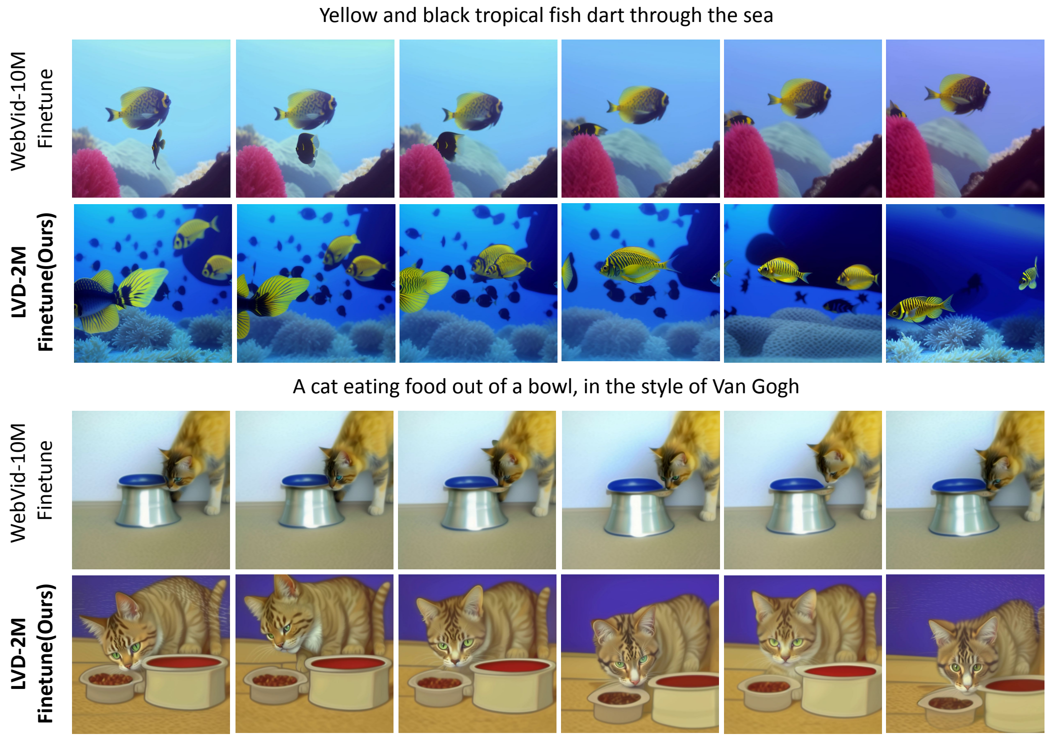

Yellow and black tropical fish dart through the sea
WebVid-10M Finetune
LVD-2M Finetune(Ours)
A cat eating food out of a bowl, in the style of Van Gogh
WebVid-10M Finetune
LVD-2M Finetune(Ours)

Figure 5: Fintuning the LM-based T2V model on LVD-2M vs. WebVid-10M. After finetuning, the model can generate richer content with larger motion. This shows that finetuning on LVD-2M can further improve the model's capability to generate more dynamic content, compared to WebVid-10M.

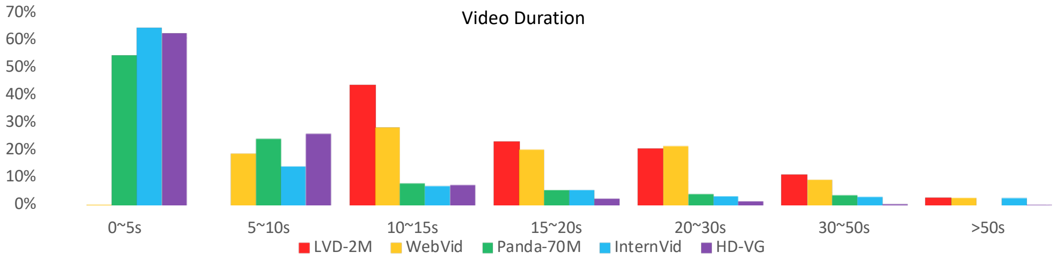


Figure 6: The distribution of video clip duration.

Fig. 7 shows the distribution of optical flow magnitude. Note that this metrics is only calculated for videos longer than 10s. Specifically for calculation, we utilize RAFT [5] with input videos scaled temporally to 2 fps and spatially to 520 × 960. The resulting score is the temporal and spatial average of the magnitudes of optical flow estimation. Videos whose average optical flow magnitude is less than 20 are filtered out from our LVD-2M.

Fig. 8 presents the distribution of caption word count. LVD-2M demonstrates a significant gap between previous datasets, with much longer captions. In our captions, we include details about the actions, characters, camera perspectives and backgrounds. And we employ Claude3-Haiku [6] for refining the captions to be more clear and concise, as we observe much redundancy in the original captions generated by LLaVA-v1.6-34B [7]. As a result, our long captions are both informative and clearly organized.

We further present a radar chart comparing LVD-2M with previous dataset, as shown in Fig. 9. We demonstrate 5 metrics, including the long-take rate measured by human raters, caption length for the

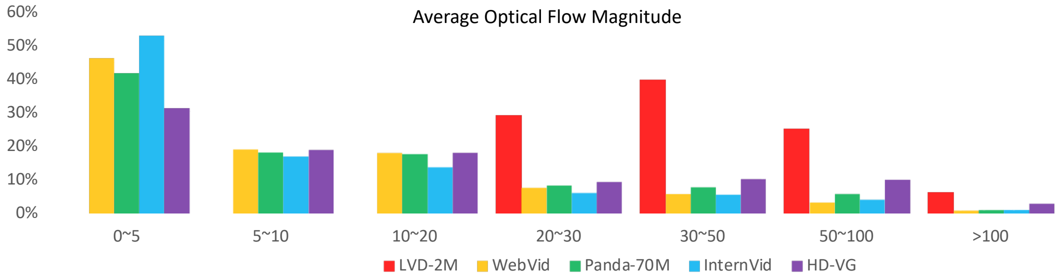


Figure 7: The distribution of average optical flow magnitude. LVD-2M demonstrate significantly larger portion of dynamic (measured by optical flow) videos.

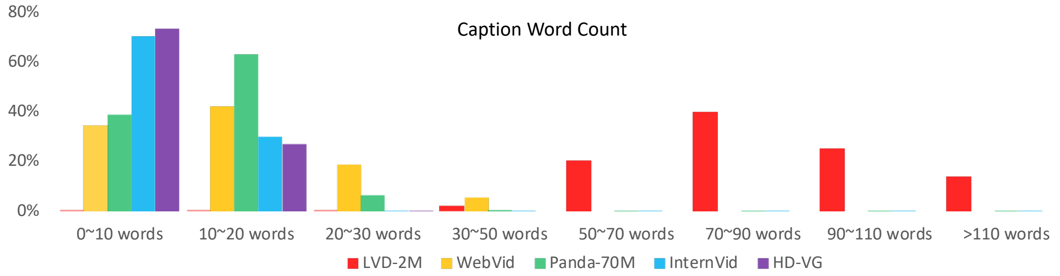


Figure 8: The distribution of caption word count.

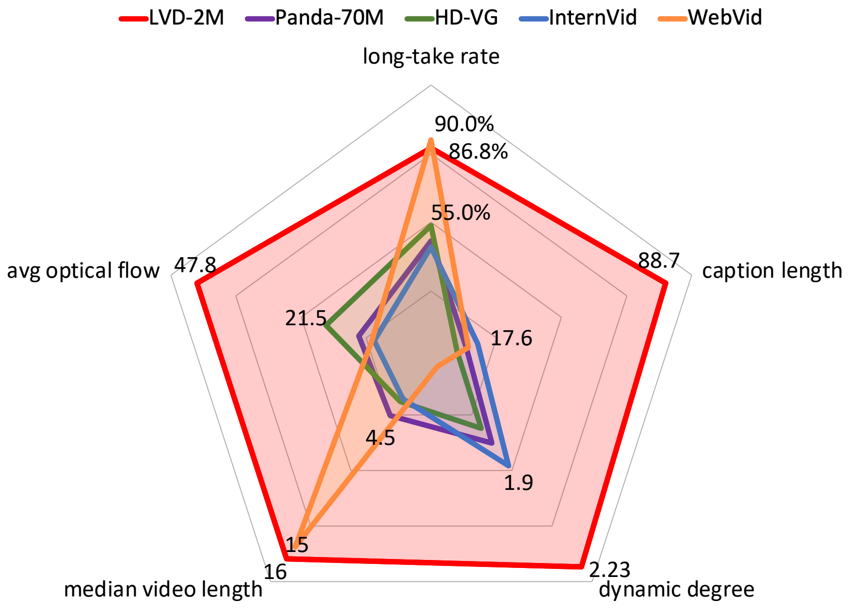
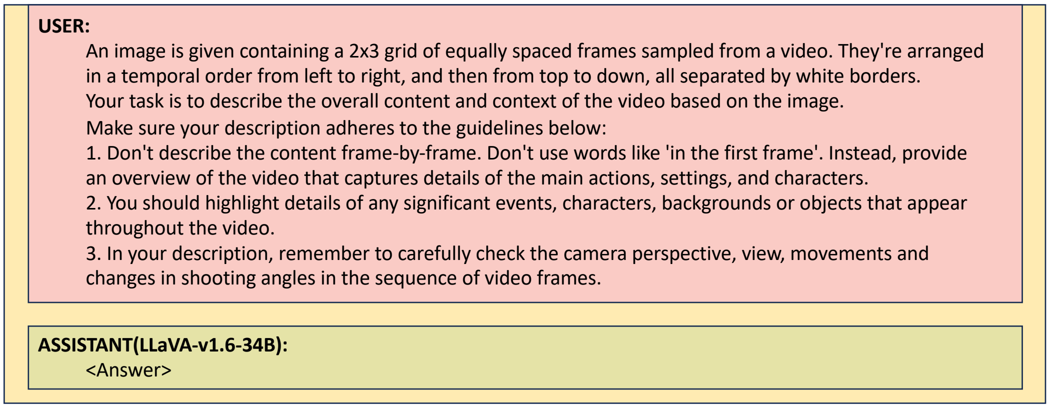


Figure 9: **LVD-2M** presents desirable quality for training of long video generation in 5 dimensions.

average caption word count, dynamic degree which is the average of human rated 1∼3 dynamic score, median video clip length and the average optical flow magnitude. For long-take rate, dynamic degree and average optical flow magnitude, the calculation is based on videos longer than 10s. Notably, for the statistics about video clip length, we choose median instead of average here because we find that the average is prone to being affected by a small portion of extremely long video clips. And median video length better reflects the portion of long videos. For the calculation of long-take rate for LVD-2M, in the main paper we exclude the data from WebVid for fair comparison, resulting 77.5%, and here we give the overall long-take rate of LVD-2M, which is 86.8%. LVD-2M presents superior quality compared to previous datasets in various dimensions.

D Prompt Design for LLaVA and Claude3-Haiku

We present the actual prompts used for our coarse-to-refined caption generation. First, 6 frames sampled from a video clip is concatenated as a 2×3 image grid as the input for LLaVA-v1.6-34B, and the VLM is instructed as in Fig. 10. If there is only one segment from the original video, the generated captions will be refined by Claude3-Haiku [6] as in Fig. 11. When there are multiple consecutive segments from the original video, we use LLaVA-v1.6-34B to generate captions for different segments independently, then we apply Claude3-Haiku for composing the chronologically ordered coarse captions to a refined caption, as shown in Fig. 12.



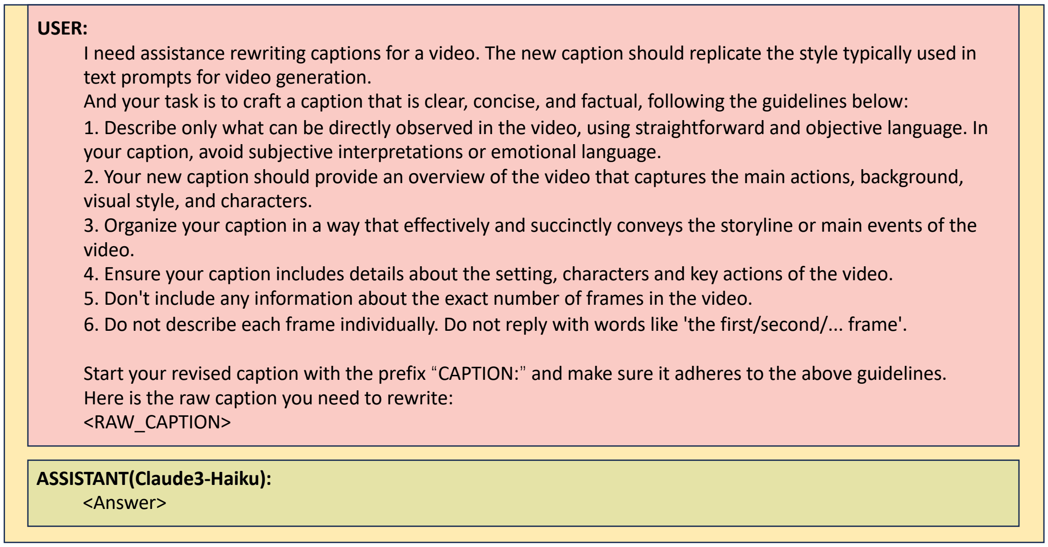
USER:

An image is given containing a 2x3 grid of equally spaced frames sampled from a video. They're arranged in a temporal order from left to right, and then from top to down, all separated by white borders.
Your task is to describe the overall content and context of the video based on the image.
Make sure your description adheres to the guidelines below:
1. Don't describe the content frame-by-frame. Don't use words like 'in the first frame'. Instead, provide an overview of the video that captures details of the main actions, settings, and characters.
2. You should highlight details of any significant events, characters, backgrounds or objects that appear throughout the video.
3. In your description, remember to carefully check the camera perspective, view, movements and changes in shooting angles in the sequence of video frames.

ASSISTANT(LLaVA-v1.6-34B):

<Answer>

Figure 10: The prompt used for instructing LLaVA-v1.6-34B [7] to generate relatively coarse captions for video clips.



USER:

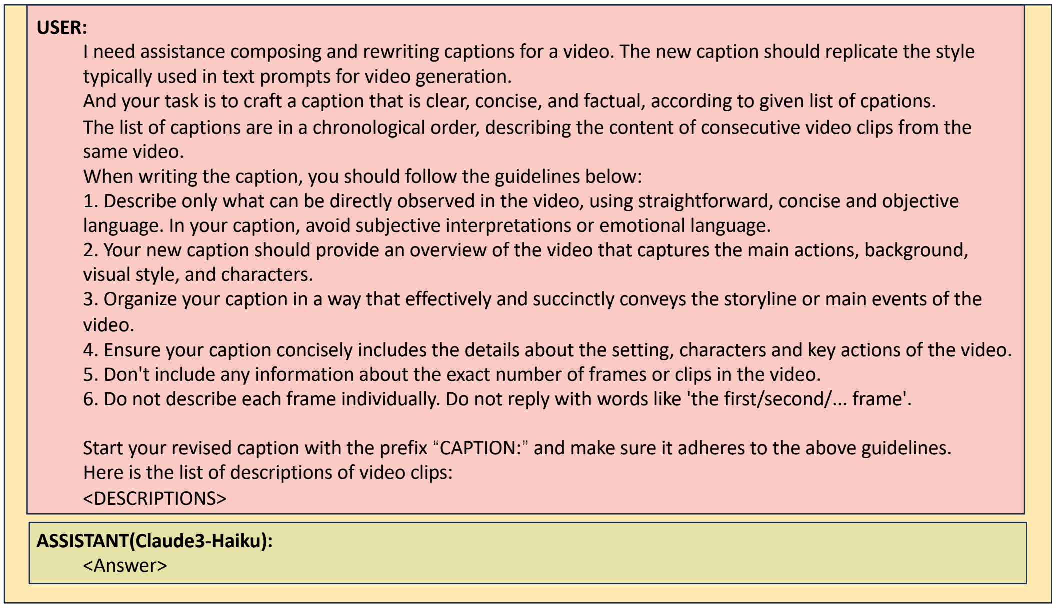
I need assistance rewriting captions for a video. The new caption should replicate the style typically used in text prompts for video generation.
And your task is to craft a caption that is clear, concise, and factual, following the guidelines below:
1. Describe only what can be directly observed in the video, using straightforward and objective language. In your caption, avoid subjective interpretations or emotional language.
2. Your new caption should provide an overview of the video that captures the main actions, background, visual style, and characters.
3. Organize your caption in a way that effectively and succinctly conveys the storyline or main events of the video.
4. Ensure your caption includes details about the setting, characters and key actions of the video.
5. Don't include any information about the exact number of frames in the video.
6. Do not describe each frame individually. Do not reply with words like 'the first/second/... frame'.

Start your revised caption with the prefix "CAPTION:" and make sure it adheres to the above guidelines.
Here is the raw caption you need to rewrite:
<RAW_CAPTION>

ASSISTANT(Claude3-Haiku):

<Answer>

Figure 11: The prompt used for instructing Claude3-Haiku [6] to refine the coarse captions from LLaVA-v1.6.



USER:

I need assistance composing and rewriting captions for a video. The new caption should replicate the style typically used in text prompts for video generation.
And your task is to craft a caption that is clear, concise, and factual, according to given list of cpations.
The list of captions are in a chronological order, describing the content of consecutive video clips from the same video.
When writing the caption, you should follow the guidelines below:
1. Describe only what can be directly observed in the video, using straightforward, concise and objective language. In your caption, avoid subjective interpretations or emotional language.
2. Your new caption should provide an overview of the video that captures the main actions, background, visual style, and characters.
3. Organize your caption in a way that effectively and succinctly conveys the storyline or main events of the video.
4. Ensure your caption concisely includes the details about the setting, characters and key actions of the video.
5. Don't include any information about the exact number of frames or clips in the video.
6. Do not describe each frame individually. Do not reply with words like 'the first/second/... frame'.

Start your revised caption with the prefix “CAPTION:” and make sure it adheres to the above guidelines.
Here is the list of descriptions of video clips:
<DESCRIPTIONS>

ASSISTANT(Claude3-Haiku):

<Answer>

Figure 12: The prompt used for instructing Claude3-Haiku [6] to refine the coarse captions from LLaVA-v1.6.

E Author Statements

The dataset is open and the data is collected from publicly available resources. For using this dataset, please check for the related license[1]. For the released data records and dataset documentation, please check our homepage at https://github.com/SilentView/LVD-2M.

[1] https://raw.githubusercontent.com/microsoft/XPretrain/main/hd-vila-100m/LICENSE

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. 2021.

[2] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*, 2024.

[3] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2023.

[4] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023.

[5] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[6] Anthropic. Claude3-Haiku. `https://www.anthropic.com/news/claude-3-family`, 2024.

[7] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.