NEURAL INFORMATION PROCESSING SYSTEMS

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

Tianwei Xiong*, Yuqing Wang*, Daquan Zhou, Zhijie Lin, Jiashi Feng, Xihui Liu✉

Project page: https://silentview.github.io/LVD-2M/

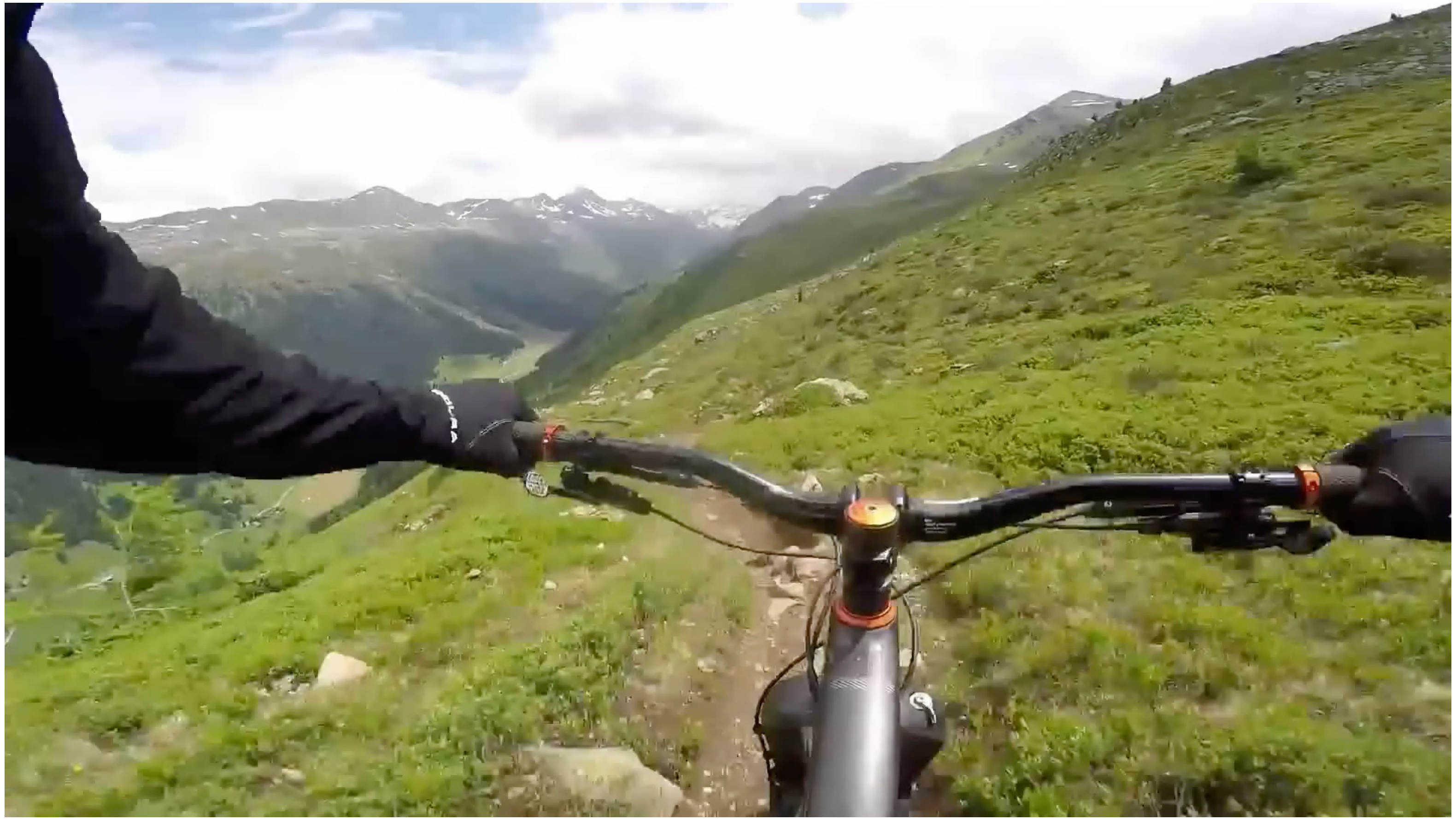
ByteDance

LVD-2M Features

- Long videos $\geq$ 10 seconds
- Long-take videos without cuts
- Large motion and diverse contents
- Temporally dense captions

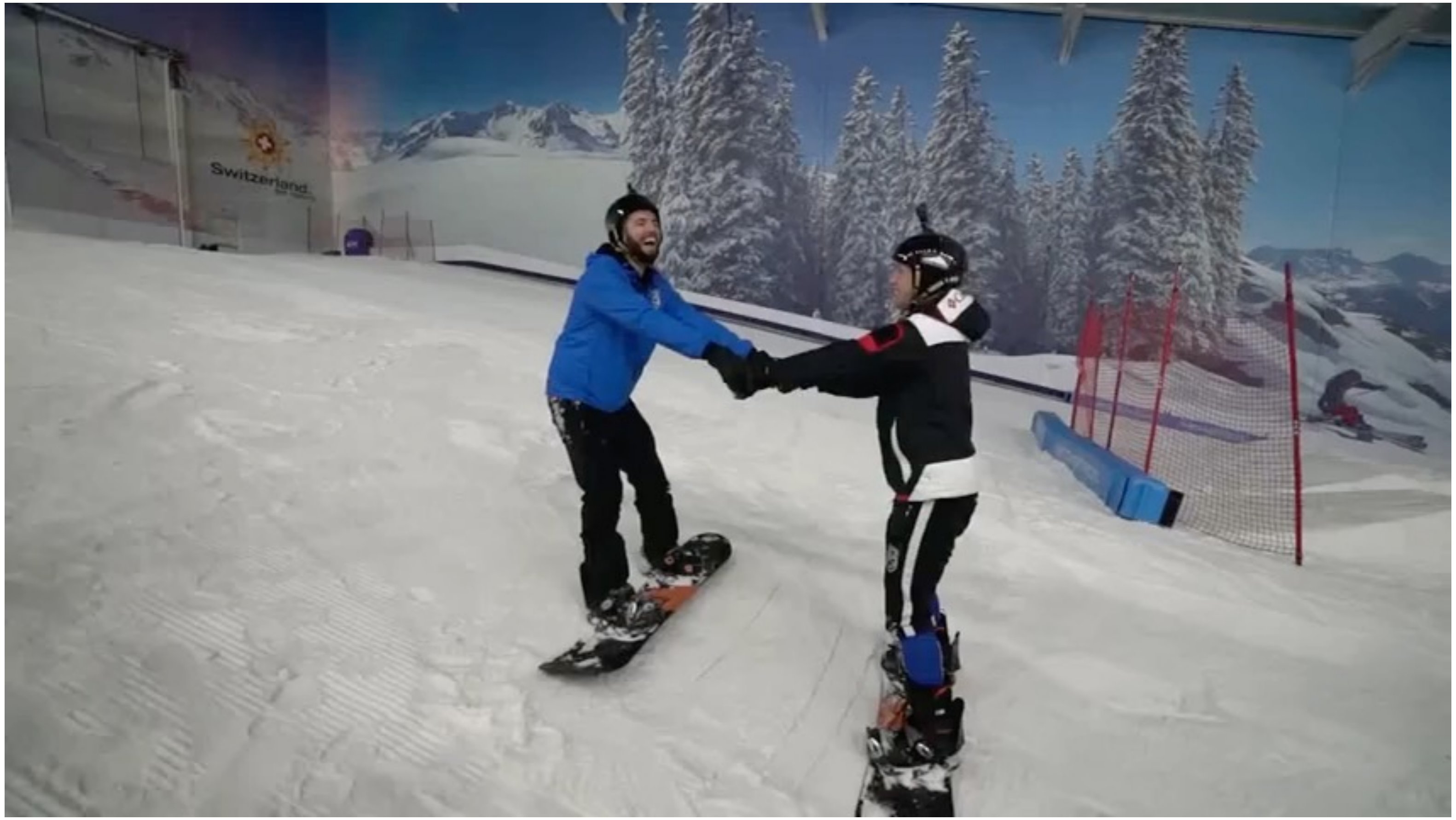
CLINCH
TEA ROOMS MRKT
Backyard Market
RINSE SHOW ROOMS
MARKET AND CAFE INSIDE
OAKLEY
ŠPORT LJUBLJANA
DVORANSKI ŠPORTI
TRIM STEZE
NOGOMET
ATLETIKA
ヒト・コミュニケーションズ
ERST

LVD-2M Caption



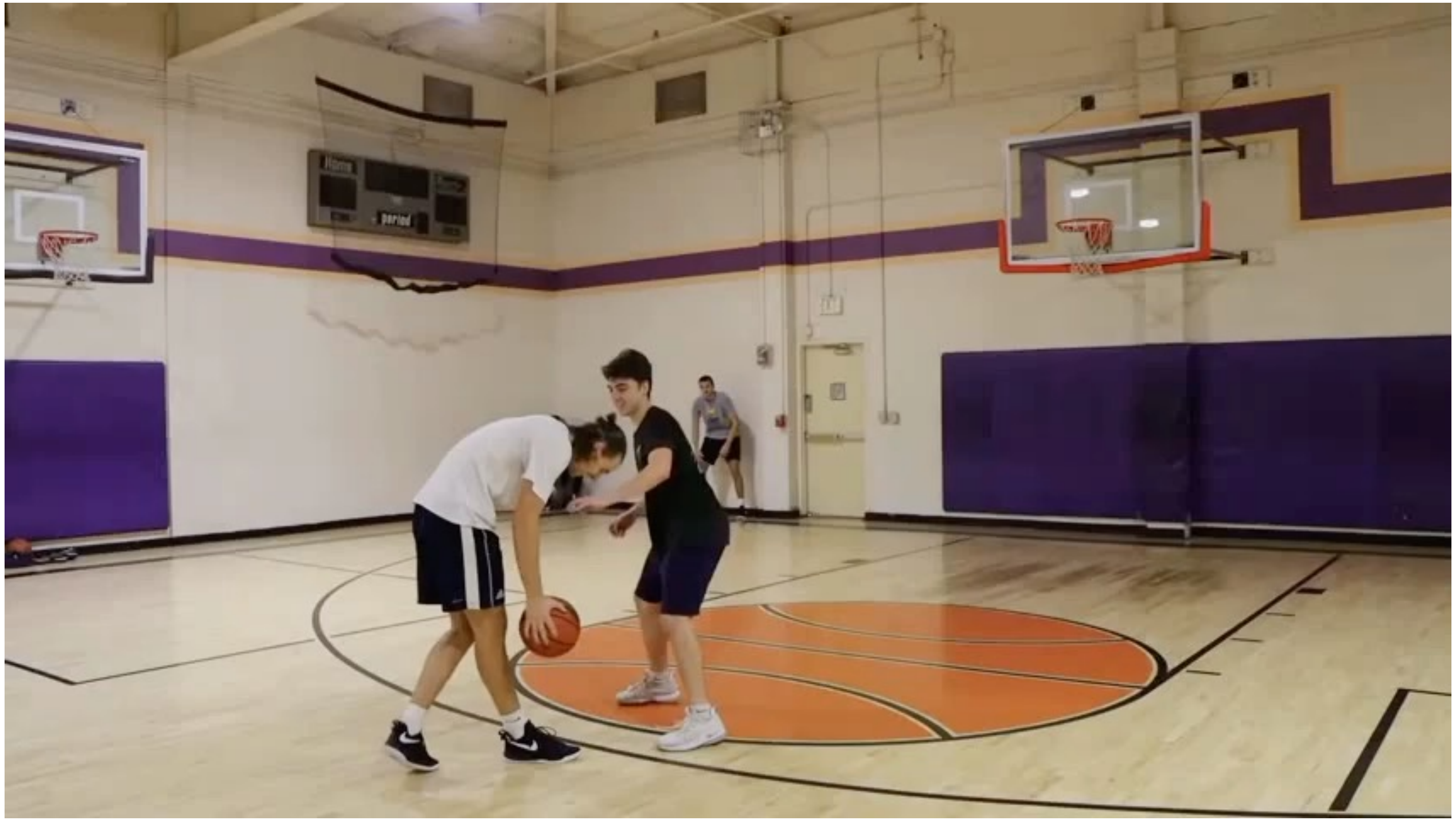
The video depicts a person riding a mountain bike through a mountainous landscape with lush greenery and rocky terrain. The rider, wearing a black outfit, navigates a dirt and rocky trail, maneuvering the bike with control and skill. The camera perspective provides a first-person view of the ride, capturing the rider's descent down the mountain. The video appears to document an outdoor adventure, showcasing the natural beauty of the surroundings.

LVD-2M Caption



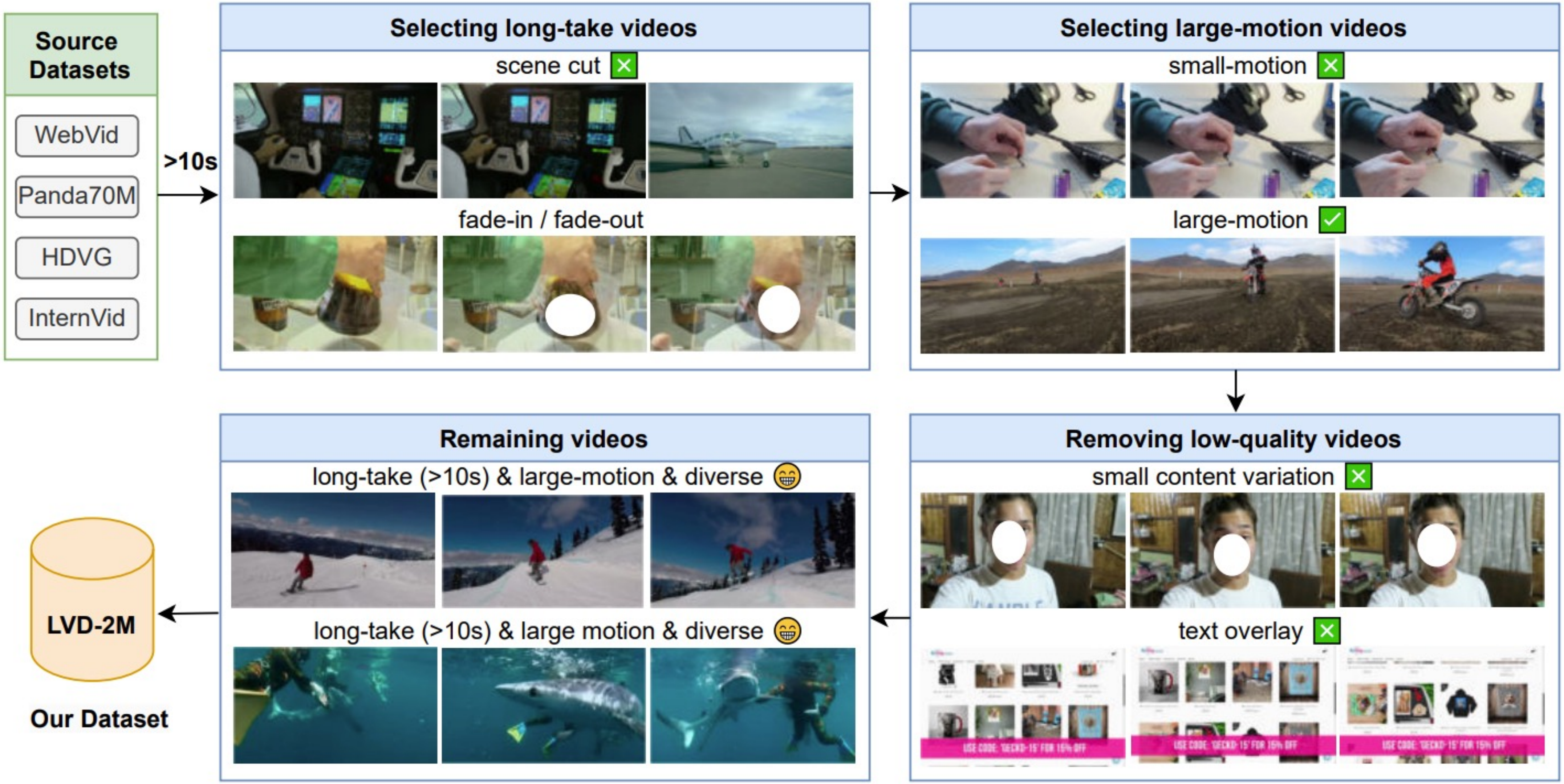
The video depicts a person learning to snowboard at an indoor snowboarding facility. The learner is wearing a blue jacket and is being assisted by another individual, both wearing snowboarding gear including helmets and goggles. The video shows the learner's progress, starting with the person holding the instructor's hand for support, then leaning forward as they begin to fall, and eventually falling onto the snow-covered ground, with the instructor remaining nearby to offer guidance.

LVD-2M Caption



The video captures an intense basketball game or practice session on an indoor court with distinctive orange, black, purple, and white color schemes. The players, dressed in athletic attire, engage in various actions such as dribbling, shooting, passing, defending, and interacting with each other. The court, equipped with multiple hoops, serves as a dynamic backdrop for the ongoing game, highlighting the competitive and active nature of the sport.

LVD-2M Video Data Pipeline



Source Datasets

WebVid

Panda70M

HDVG

InternVid

>10s

Selecting long-take videos

scene cut ☒

fade-in / fade-out

Selecting large-motion videos

small-motion ☒

large-motion ☑

Removing low-quality videos

small content variation ☒

text overlay ☒

Remaining videos

long-take (>10s) & large-motion & diverse 😁

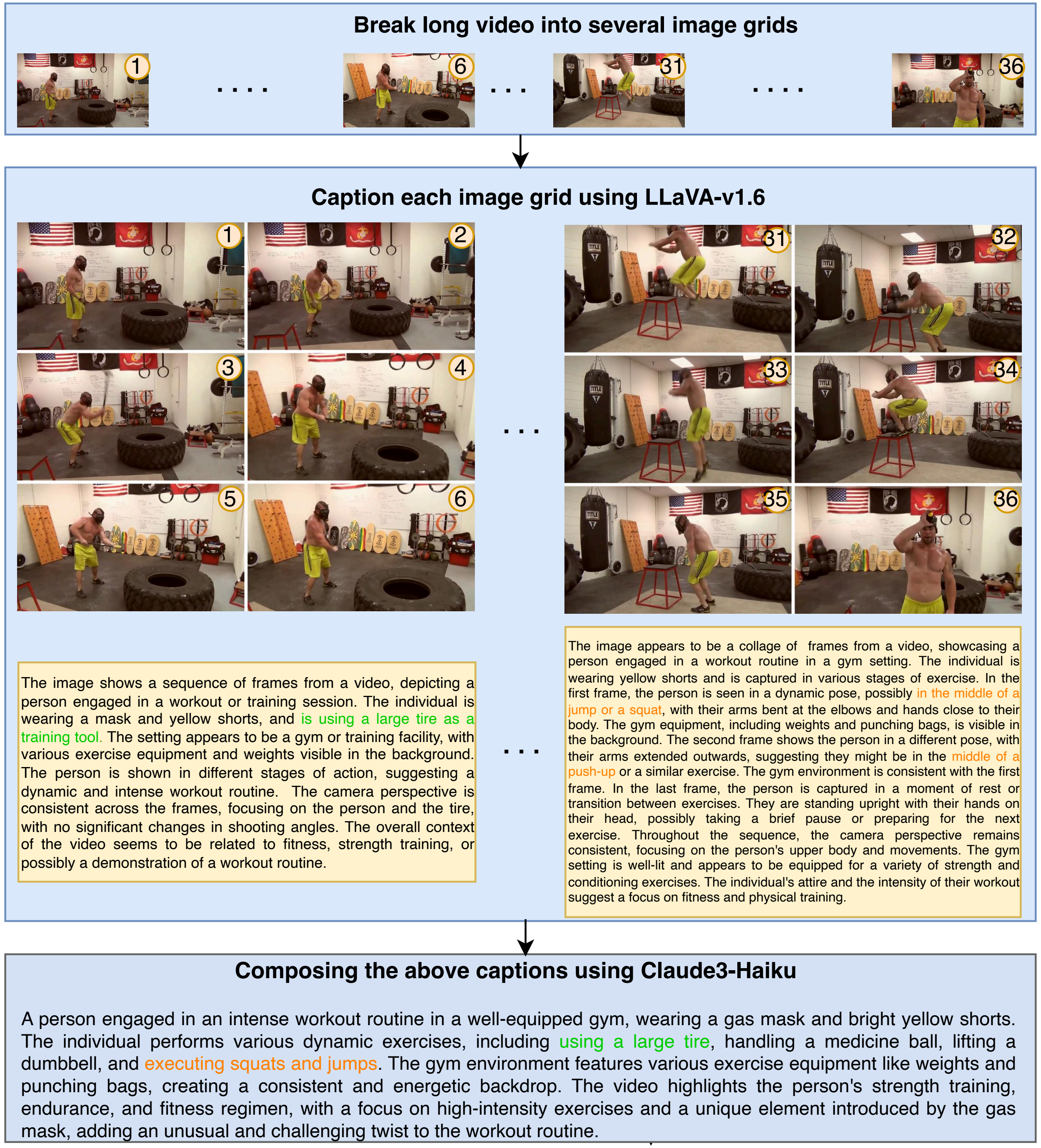
long-take (>10s) & large motion & diverse 😁

LVD-2M

Our Dataset

LVD-2M

Video Caption Pipeline



Break long video into several image grids

Caption each image grid using LLaVA-v1.6

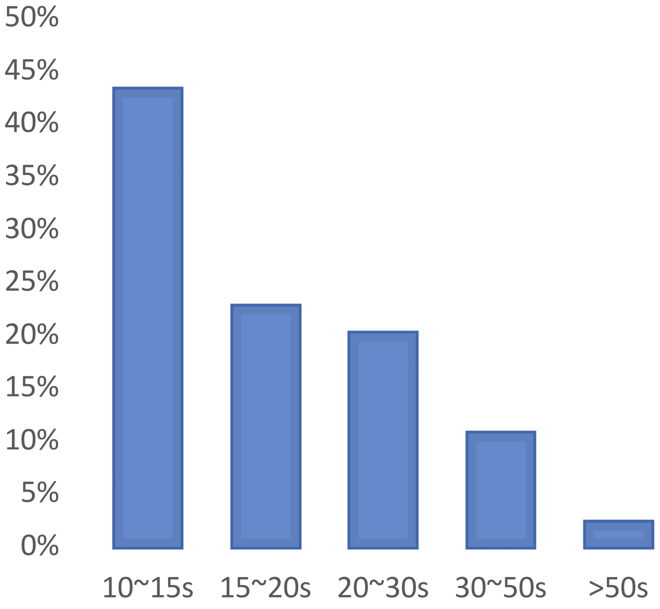
The image shows a sequence of frames from a video, depicting a person engaged in a workout or training session. The individual is wearing a mask and yellow shorts, and is using a large tire as a training tool. The setting appears to be a gym or training facility, with various exercise equipment and weights visible in the background. The person is shown in different stages of action, suggesting a dynamic and intense workout routine. The camera perspective is consistent across the frames, focusing on the person and the tire, with no significant changes in shooting angles. The overall context of the video seems to be related to fitness, strength training, or possibly a demonstration of a workout routine.

The image appears to be a collage of frames from a video, showcasing a person engaged in a workout routine in a gym setting. The individual is wearing yellow shorts and is captured in various stages of exercise. In the first frame, the person is seen in a dynamic pose, possibly in the middle of a jump or a squat, with their arms bent at the elbows and hands close to their body. The gym equipment, including weights and punching bags, is visible in the background. The second frame shows the person in a different pose, with their arms extended outwards, suggesting they might be in the middle of a push-up or a similar exercise. The gym environment is consistent with the first frame. In the last frame, the person is captured in a moment of rest or transition between exercises. They are standing upright with their hands on their head, possibly taking a brief pause or preparing for the next exercise. Throughout the sequence, the camera perspective remains consistent, focusing on the person's upper body and movements. The gym setting is well-lit and appears to be equipped for a variety of strength and conditioning exercises. The individual's attire and the intensity of their workout suggest a focus on fitness and physical training.

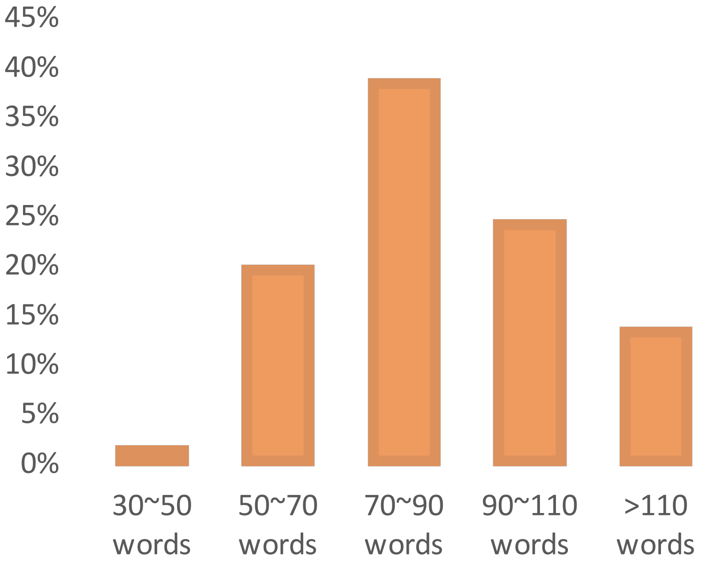
Composing the above captions using Claude3-Haiku

A person engaged in an intense workout routine in a well-equipped gym, wearing a gas mask and bright yellow shorts. The individual performs various dynamic exercises, including using a large tire, handling a medicine ball, lifting a dumbbell, and executing squats and jumps. The gym environment features various exercise equipment like weights and punching bags, creating a consistent and energetic backdrop. The video highlights the person's strength training, endurance, and fitness regimen, with a focus on high-intensity exercises and a unique element introduced by the gas mask, adding an unusual and challenging twist to the workout routine.

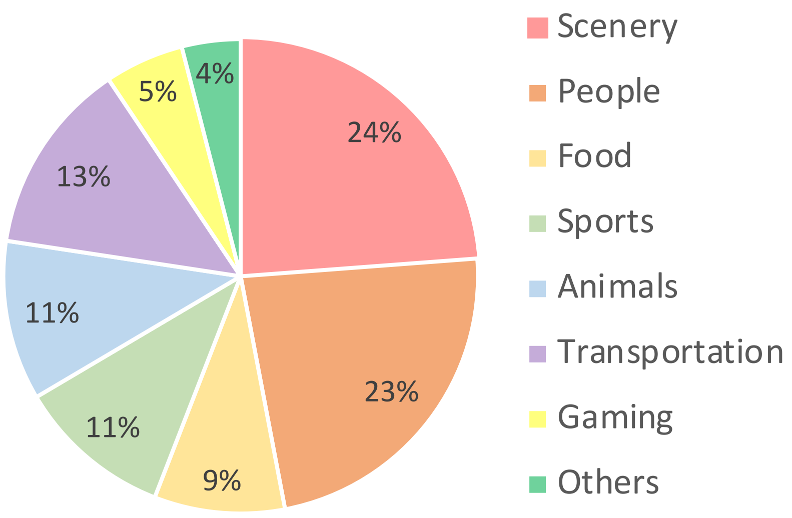
LVD-2M Statistics



Video Duration



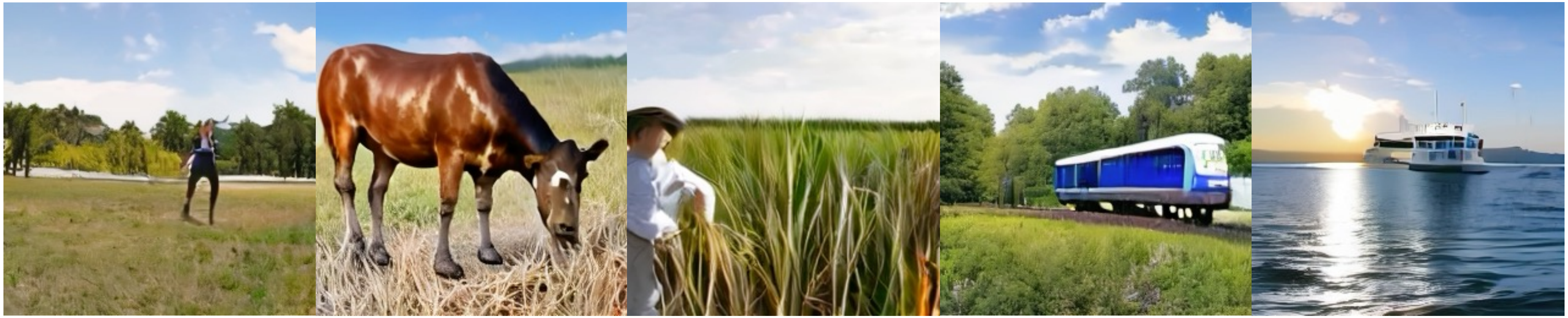
Caption Word Count



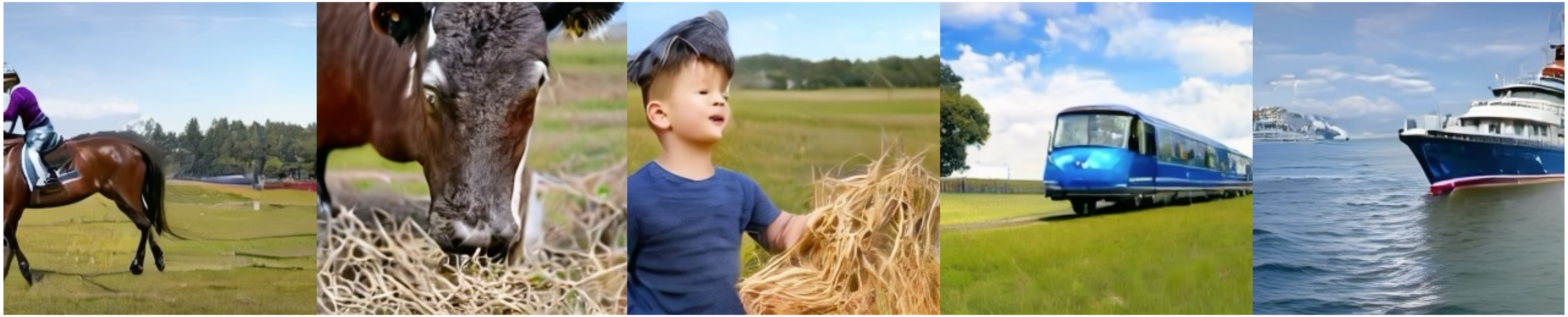
Video Categories

LVD-2M for Model Finetuning

Long-frame-length Fine-tuning on WebVid-10M[1]



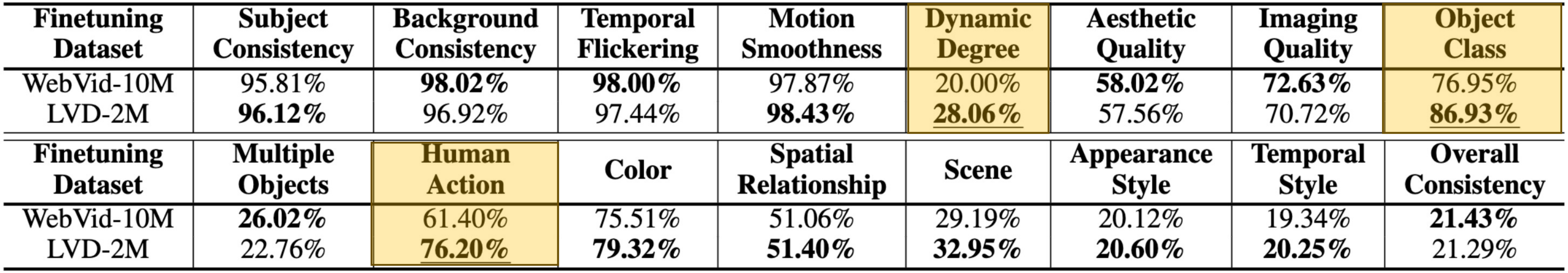
Long-frame-length Fine-tuning on LVD-2M



[1] Bain, Max, et al. "Frozen in time: A joint video and image encoder for end-to-end retrieval." ICCV. 2021.

LVD-2M for Model Finetuning

Vbench[1] evaluation for the two finetuned diffusion-based T2V models on LVD-2M and WebVid-10M [2] separately. Metrics exhibiting an absolute difference greater than 8% between the two models are underlined for emphasis.

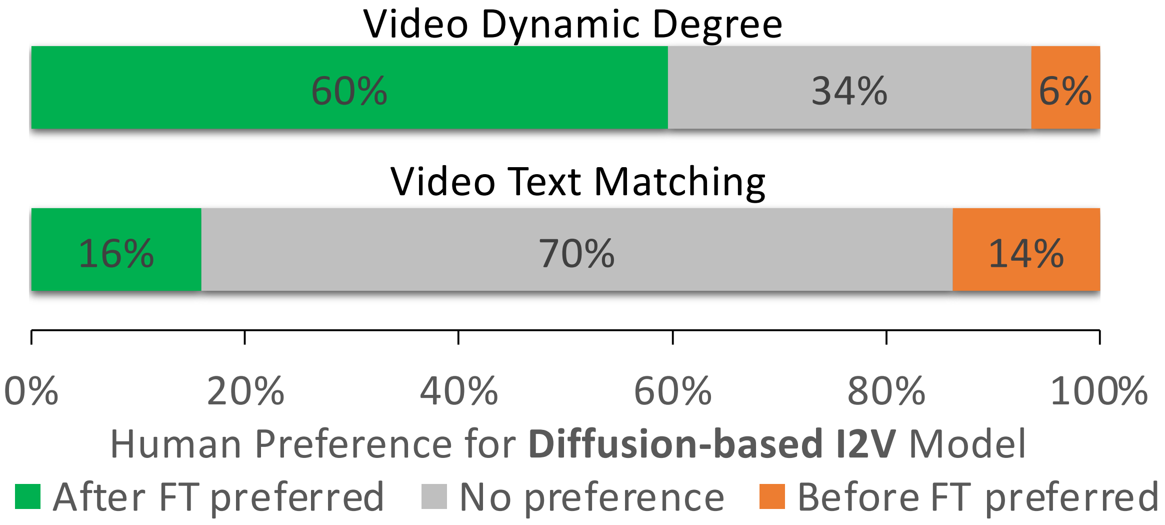


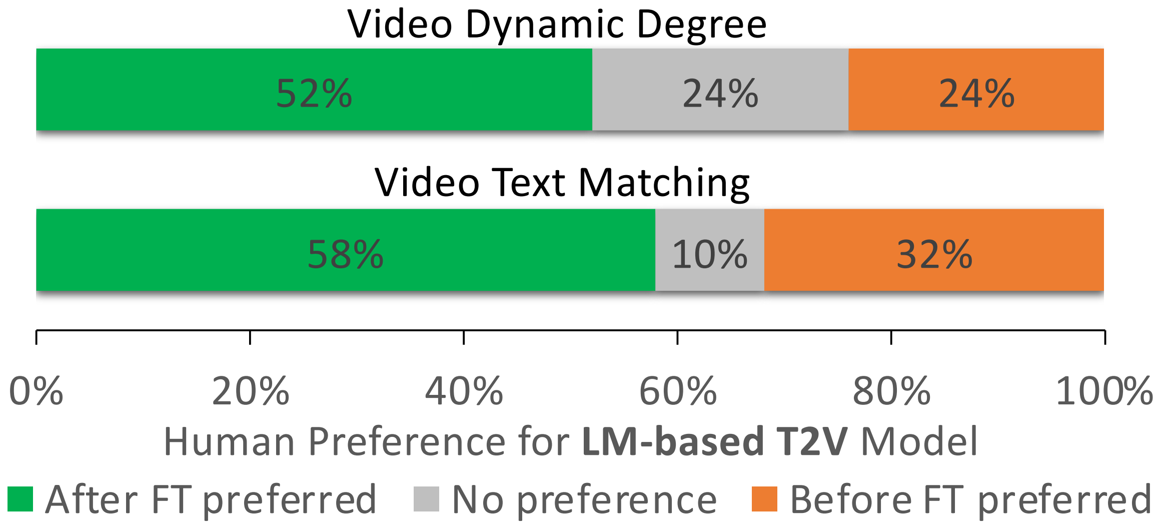
Finetuning Dataset	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
WebVid-10M	95.81%	**98.02%**	**98.00%**	97.87%	20.00%	**58.02%**	**72.63%**	76.95%
LVD-2M	**96.12%**	96.92%	97.44%	**98.43%**	**28.06%**	57.56%	70.72%	**86.93%**

Finetuning Dataset	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency
WebVid-10M	**26.02%**	61.40%	75.51%	51.06%	29.19%	20.12%	19.34%	**21.43%**
LVD-2M	22.76%	**76.20%**	**79.32%**	**51.40%**	**32.95%**	**20.60%**	**20.25%**	21.29%

[1] Huang, Ziqi, et al. "Vbench: Comprehensive benchmark suite for video generative models." CVPR. 2024.
[2] Bain, Max, et al. "Frozen in time: A joint video and image encoder for end-to-end retrieval." ICCV. 2021.

LVD-2M for Model Finetuning





LVD-2M

Project Page: https://silentview.github.io/LVD-2M/