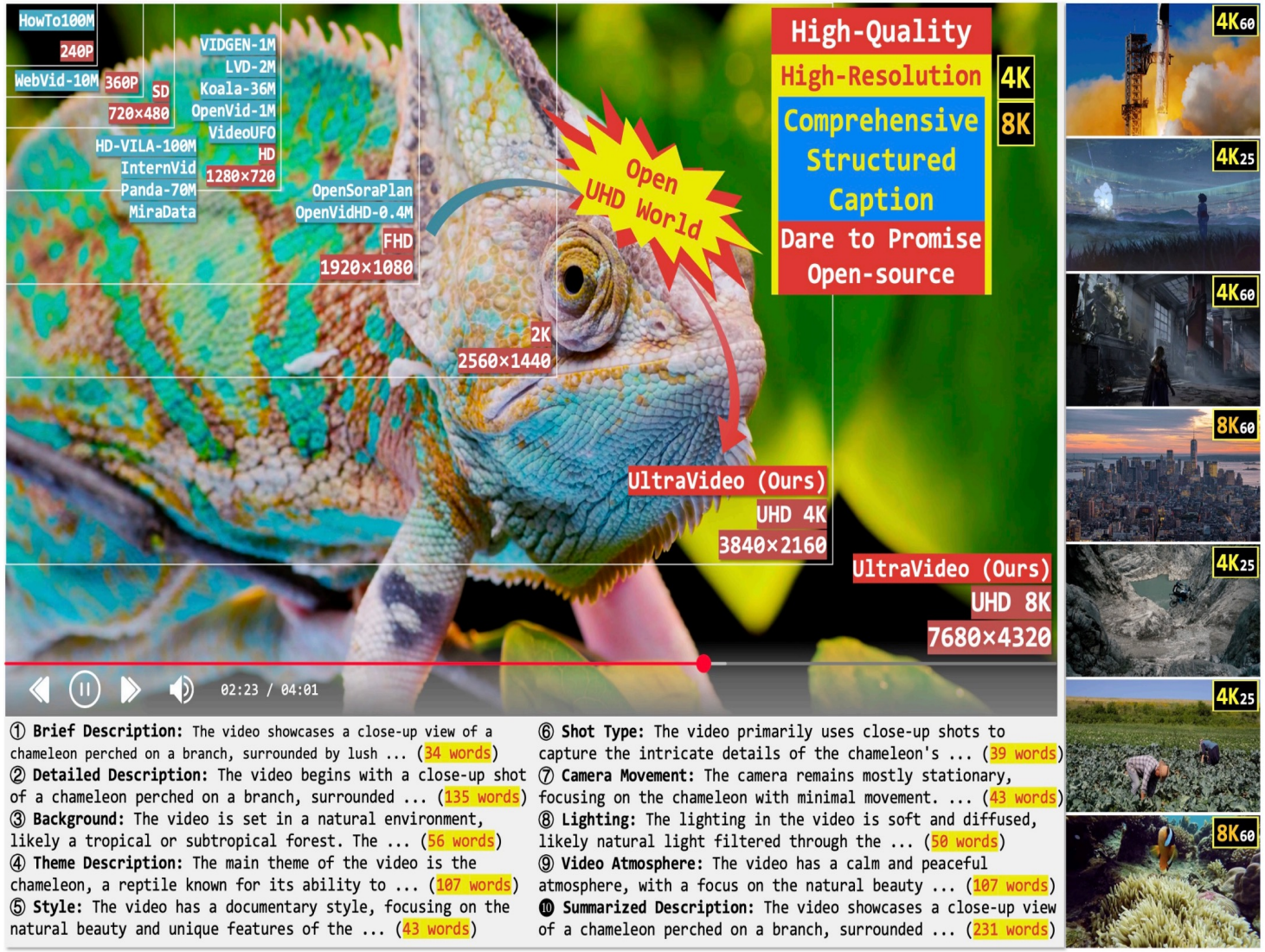
UltraVideo: High-Quality UHD Video Dataset with Comprehensive Captions

Zhucun Xue[1], **Jiangning Zhang**[†1], Teng Hu[2], Haoyang He[1], Yinan Chen[1], Yuxuan Cai[3], Yabiao Wang[1], Chengjie Wang[2], Yong Liu[1], Xiangtai Li[4], Dacheng Tao[4],

[1]Zhejiang University [2]Shanghai Jiao Tong University [3]Huazhong University of Science Technology [4]Nanyang Technological University

Project：https://xzc-zju.github.io/projects/UltraVideo/

Introduction



Motivation

- Growing demand for UHD/8K video generation.
- Existing datasets lack scale, quality, and rich captions.
- Provides high-quality data for text-to-video (T2V) research.

Contribution

- UltraVideo: UHD-4K dataset (22.4% 8K).
- Over 100 topics, each clip with 9 structured + 1 summary caption.
- Four-stage pipeline: collection, filtering, purification, captioning.
- Enables UltraWan-1K/4K models for high-quality video generation.

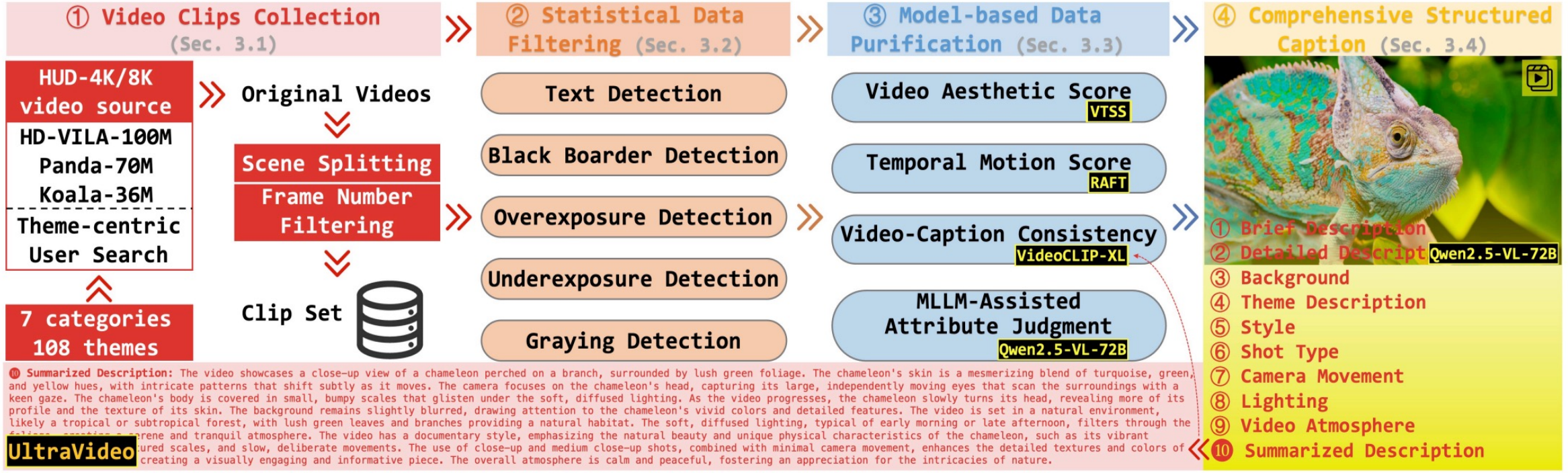
Data Construction

Stage 1: **Video Clips Collection** – Strict quality control at the source.

Stage 2: **Statistical Data Filtering** – Quantitative removal of low-quality frames.

Stage 3: **Model-based Data Purification** – AI-assisted quality evaluation.

Stage 4: **Comprehensive Structured Captioning** – Fine-grained semantic annotation.



Comparison with Popular Text-to-Video Datasets

Table 1: Comparison popular text-to-video datasets. Our UltraVideo is a *high-resolution* and *high-quality* premium T2V dataset, featuring *comprehensive structured captions* with a significantly longer average caption length. In addition to the main short version ranging from 3 seconds to 10 seconds, we also list the derived long version (*:) that exceeds 10 seconds for potential future research.

Dataset	Resolution	Structured Caption	Average Caption Length	Average Video Length	Duration	Video Clips	Year
HowTo100M [19]	240p	None	4.0 words	3.6s	134.5Khr	136M	2019
WebVid-10M [3]	360p	None	12.0 words	17.5s	52Khr	10M	2021
HD-VILA-100M [43]	720p	None	32.5 words	13.4s	371.5Khr	103M	2022
InternVid [38]	720p	None	17.6 words	11.7s	760.3Khr	234M	2023
Panda-70M [8]	720p	None	13.2 words	8.5s	166.8Khr	70.8M	2024
MiraData [14]	720p	6	318.0 words	72.1s	16Khr	330K	2024
VIDGEN-1M [31]	720p	None	89.3 words	10.6s	2.9Khr	1M	2024
LVD-2M [41]	720p	None	88.8 words	20.2s	14.6Khr	2.1M	2024
Koala-36M [36]	720p	6	202.3 words	13.6s	137Khr	36M	2024
OpenSoraPlan [16]	1080p	None	100.2 words	20.1s	2.8Khr	512K	2024
OpenVid-1M [20]	720p	None	126.5 words	7.2s	2.1Khr	1M	2025
OpenVidHD-0.4M [20]	1080p	None	104.5 words	9.6s	1.2Khr	433K	2025
VideoUFO [37]	720p	2	155.5 words	12.6s	3.5Khr	1M	2025
UltraVideo-Long (Ours)*	4K / 8K	10	850.3 words	30.9s	143hr	17K	2025
UltraVideo (Ours)	4K / 8K	10	824.2 words	5.3s	62hr	42K	2025

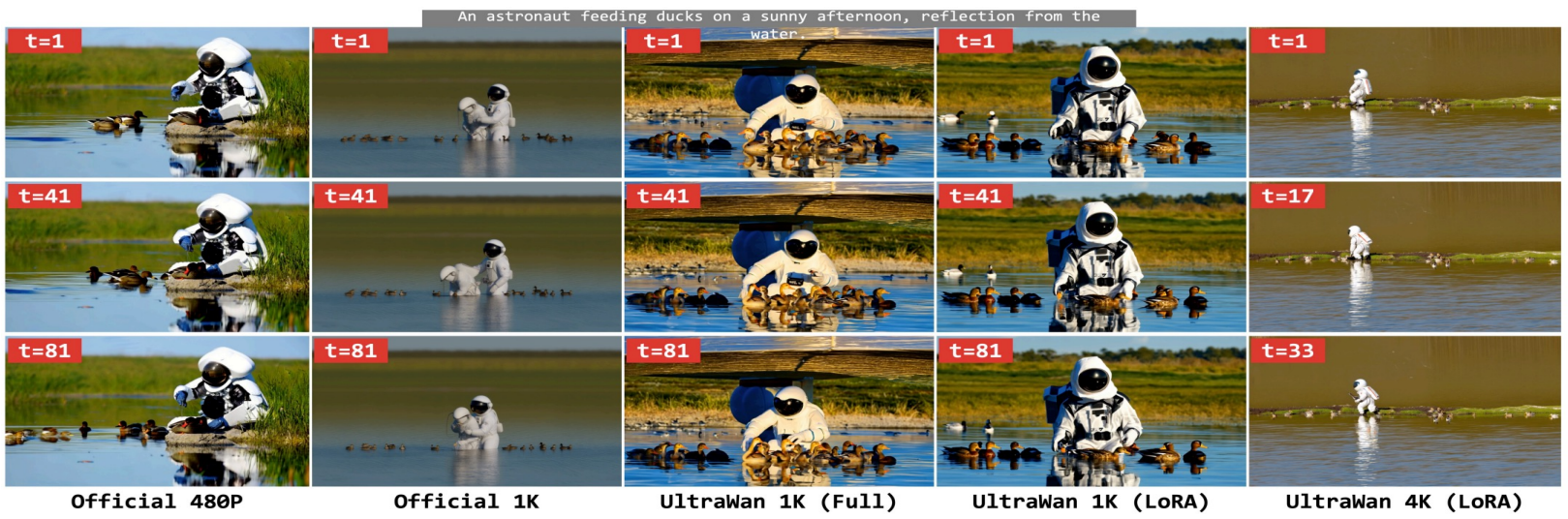
Experiments

- Based on the UltraVideo dataset, we explored the attempt of generating natively high-resolution videos, and specifically conducted fine-tuning experiments using **Wan-T2V-1.3B**

Comparison results for high-resolution video generation.

Models	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
Wan-T2V-1.3B-480p [34]	96.11%	98.06%	99.09%	98.75%	27.77%	65.83%	68.91%	66.66%
Wan-T2V-1.3B-1K [34]	95.86%	98.15%	98.07%	98.75%	66.66%	54.82%	55.12%	33.33%
UltraWan-1K (Full)	95.71%	97.94%	98.86%	99.06%	22.22%	61.52%	67.39%	66.66%
UltraWan-1K (LoRA)	97.27%	98.26%	99.33%	98.62%	66.66%	62.5%	67.74%	82.29%
UltraWan-4K (LoRA)	96.05%	98.02%	98.88%	98.47%*	66.66%*	56.81%	71.61%	50.00%

Models	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency
Wan-T2V-1.3B-480p [34]	51.04%	66.66%	100.0%	100.0%	08.33%	20.54%	24.39%	25.31%
Wan-T2V-1.3B-1K [34]	25.00%	22.22%	100.0%	36.66%	00.00%	18.75%	12.24%	20.65%
UltraWan-1K (Full)	47.91%	66.66%	100.0%	50.00%	16.66%	17.85%	19.81%	24.27%
UltraWan-1K (LoRA)	49.58%	66.66%	100.0%	75.76%	18.22%	19.57%	23.34%	23.99%
UltraWan-4K (LoRA)	42.75%	66.66%	100.0%	100.0%	00.00%	19.46%	19.31%	22.88%



Experiments

- Thanks to the structured captions in UltraVideo during training, our UltraWan exhibits stronger semantic consistency

Semantic consistency with fine-grained captions

Conclusion

- **Filled the UHD Dataset Gap:** As the first open-source UHD-4K/8K T2V dataset, UltraVideo addresses the lack of high-resolution, high-quality data in existing research, directly supporting the development of movie-level, short-video-level UHD generation applications.
- **Proven Effectiveness of the Four-Stage Pipeline**: The automated curation process (collection → statistical filtering → model purification → structured captioning) significantly improves data quality—compared to Koala-36M, the low-quality rate drops from 41.5% to 2.3%, and each video is equipped with 9 types of fine-grained structured captions (average 824 words), laying a solid foundation for semantic-controllable generation.
- **Driven Downstream Model Progress**: When used to fine-tune models like UltraWan-1K/-4K, it enables native 4K video generation with better visual quality (71.61% imaging quality in VBench) and text consistency (54.5% human preference), verifying the dataset's practical value.

Thanks

Project： https://xzc-zju.github.io/projects/UltraVideo/

Dataset： https://huggingface.co/datasets/APRIL-AIGC/UltraVideo

https://huggingface.co/datasets/APRIL-AIGC/UltraVideo-Long

Github： https://github.com/xzc-zju/UltraVideo