A Database for Multi-Modal Short Video Quality Assessment

Yukun Zhang[1,2], Chuan Wang[1,2], Sanyi Zhang[1,2], Xiaochun Cao[3]

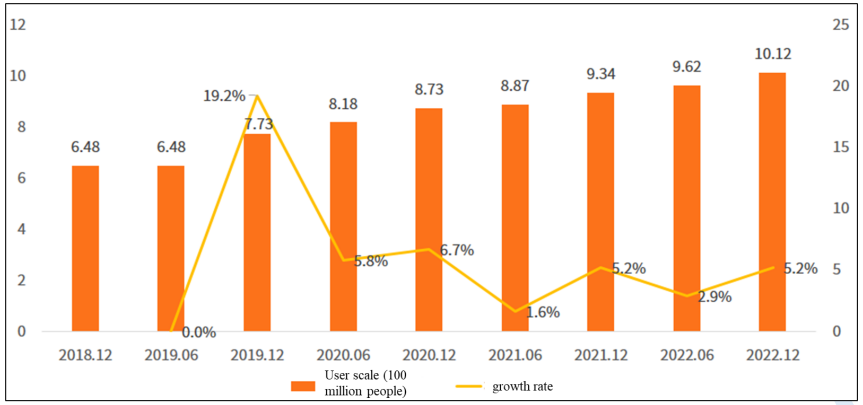
SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences
School of Cyberspace Security, Chinese Academy of Sciences
School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University

ICASSP 2023
Oral

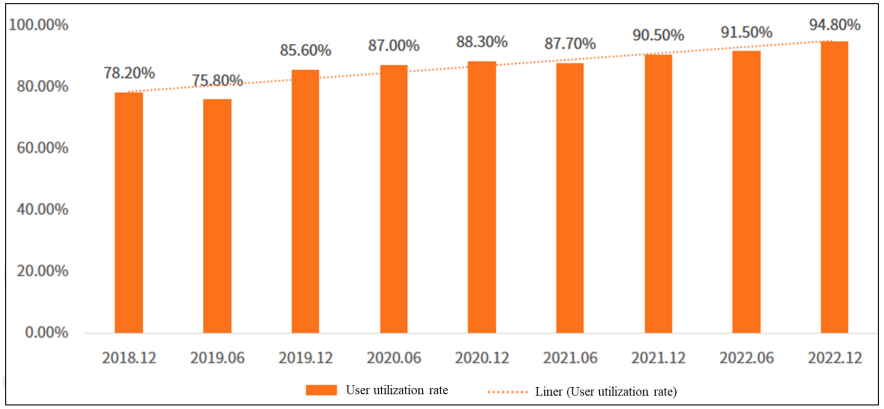
Introduction

Short videos are more attractive and influential, and it is extremely challenging to evaluate the quality of short videos.

Huge number of users

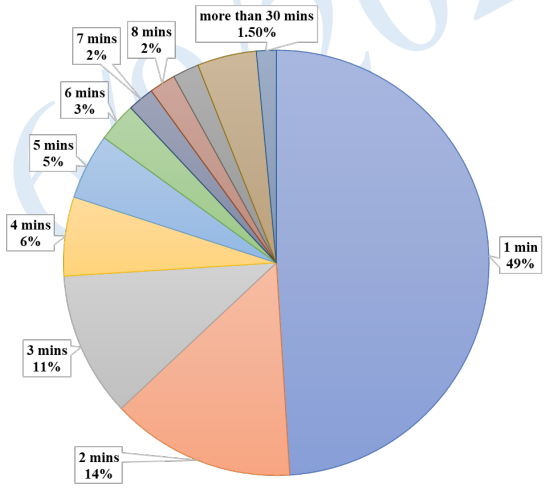


The scale of short video users continues to expand

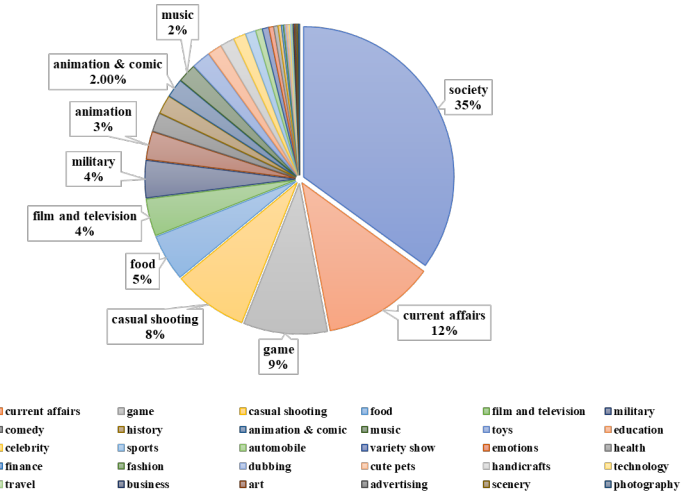


The proportion of short video users is increasing year by year.

rich in content



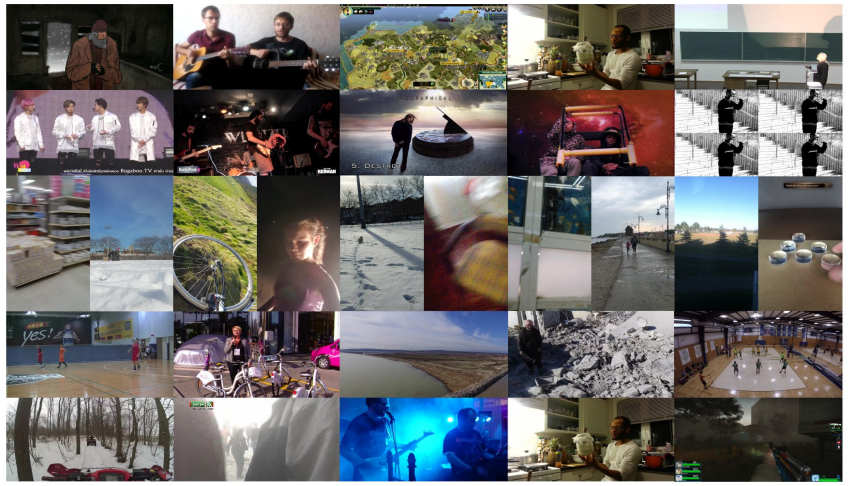
85% of popular short videos are 1-5 minutes long



There are many types of popular short videos

Introduction

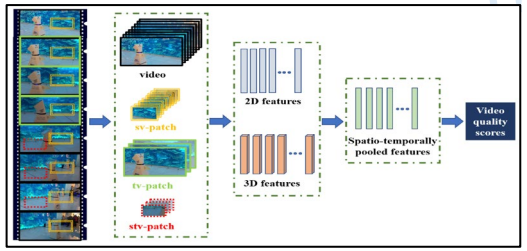
Existing video quality assessment datasets do not match short video quality assessment



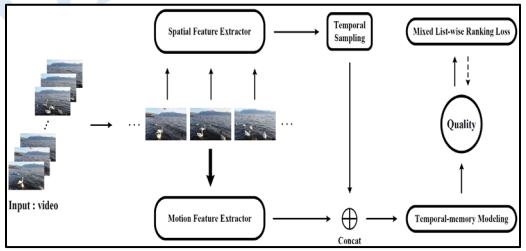
YouTube-UGC

MOS：1.706

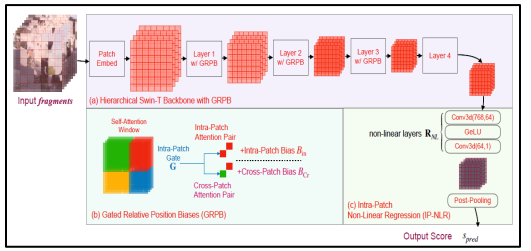
MOS：4.526



PVQ[1] (2021)
2D space + 3D space-time



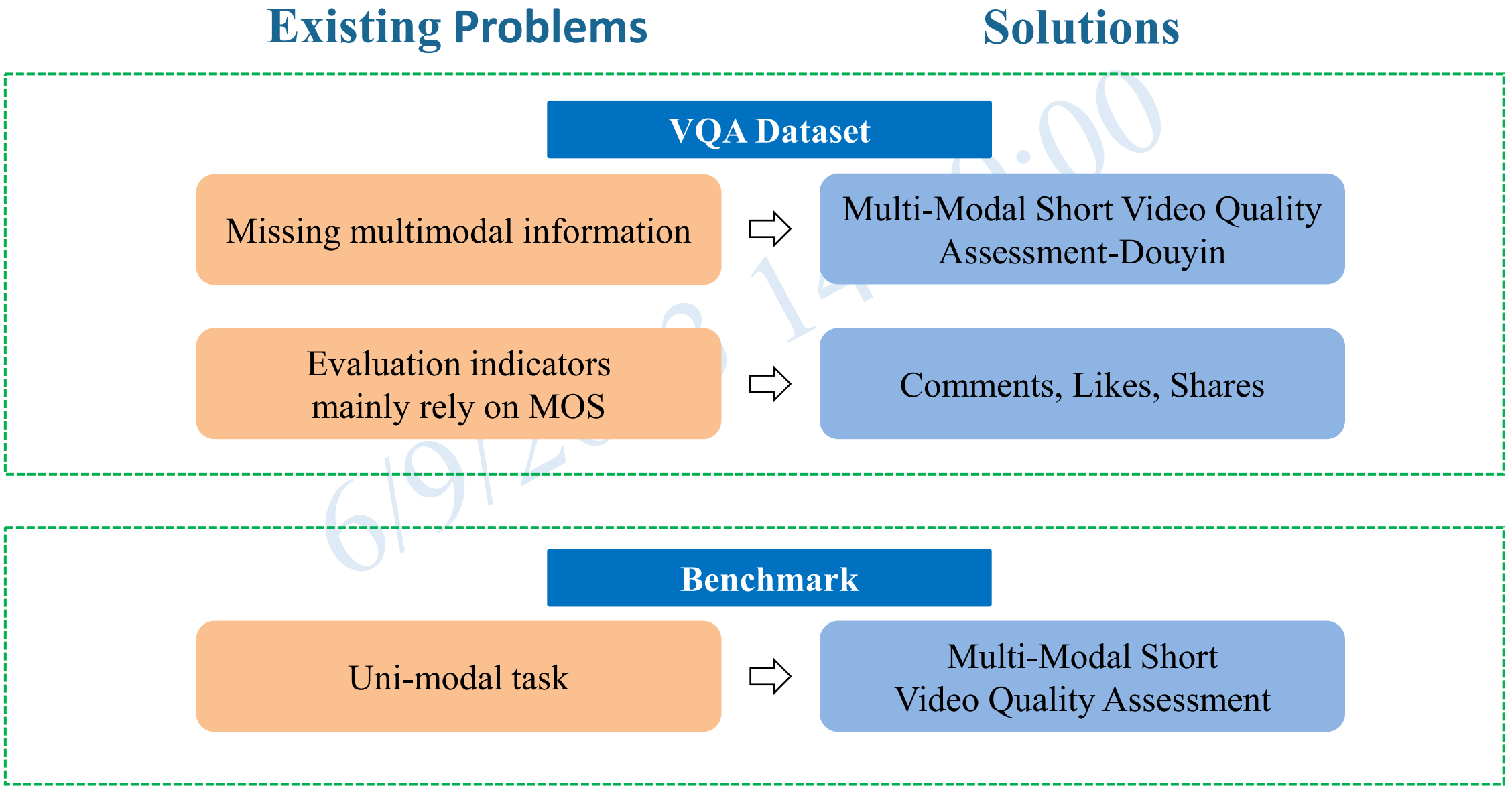
BVQA[2] (2022)
IQA+ Action recognition



FAST-VQA[3] (2022)
VQA sampling strategy

[1] Ying, et.al. , “Patch-vq:’patching up’the video quality problem,” CVRP 2021
[2] Li, et.al. , “Blindly assess quality of in-the wild videos via quality-aware pre-training and motion perception,” TCSVT 2022
[3] Wu, et.al. , “Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling,” ECCV 2022

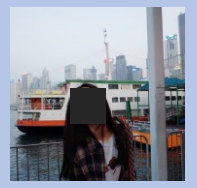
Introduction



MMSVD-Douyin Dataset

- Multimodal input:
  Text, Image, and Video

- Evaluation indicators:
  Comments, Likes, and Shares

Video Caption	#抖音美食创作人 如果你有空气炸锅，早上中午晚上都可以吃些啥，分享4个简单的小食谱～#赴重阳家宴的拿手菜	Comment: 82,581
Author's Name	**	
Personalized Signature	理想就是理想的生活。：*****	
Author Profile  Video Cover Image 		Like: 2,176,447
Video 		Share: 266,601

MMSVD-Douyin Dataset

- The statistics of MMSVD-Douyin

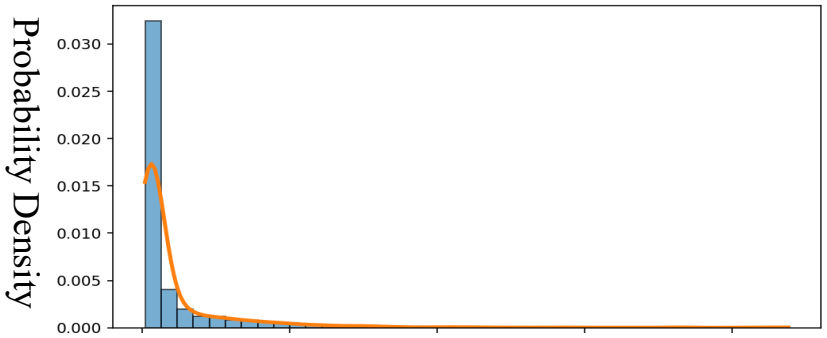
	Min	Max	Average
Video Duration (sec)	4	877	44.8
Frame Rate (frame/sec)	11	60	31
Number of Comments	0	6,096,949	114,286.8
Number of Likes	5,665	25,836,469	1,518,952.8
Number of Shares	550	4,123,251	134,917.8

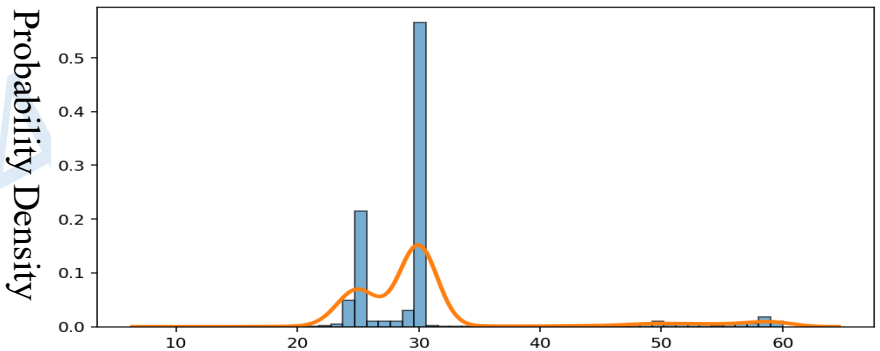
- Summary of popular video quality datasets and our dataset

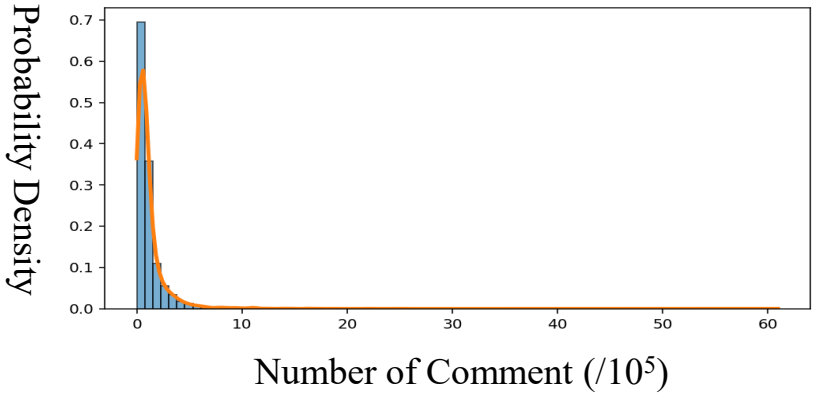
Dataset	Source	Unique Contents	Resolution	Frame Rate	Video Duration	Data Type	Evaluation Indicators
CVD2014	Captured	234	480p,720p	9-30	10-25	Video	MOS
KoNViD-1k	Flickr	1,200	540p	24-30	8	Video	MOS
LIVE-VQC	Captured	585	240p-1080p	19-30	10	Video	MOS
YouTube-UGC	YouTube	1,500	360p-4k	15-60	20	Video	MOS
MMSVD-Douyin	Douyin	4,684	480p-720p	11-60	4-877	Text, Image, Video	Comments, Likes, Shares

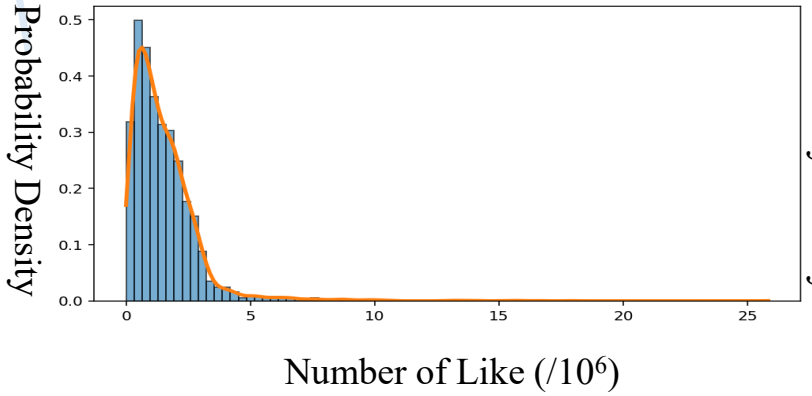
MMSVD-Douyin Dataset

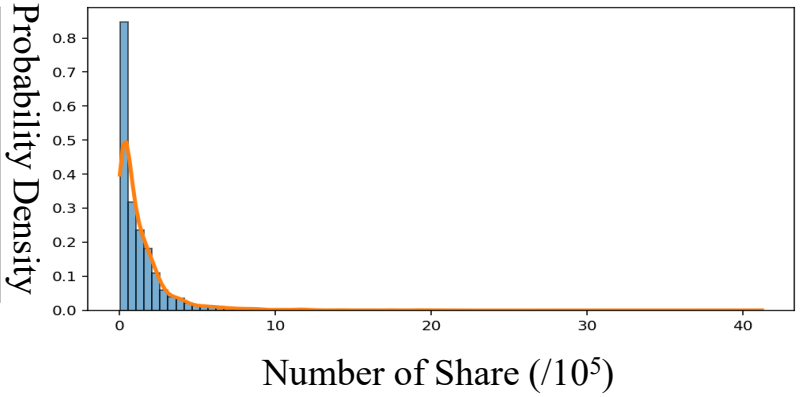
- Histograms and the fitted kernel distributions of video duration, frame rate, ”comment”, ”like”, and ”share”







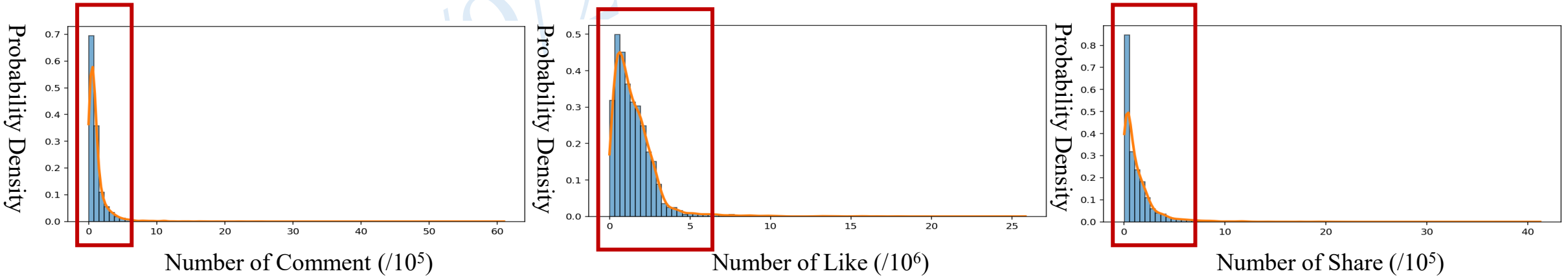




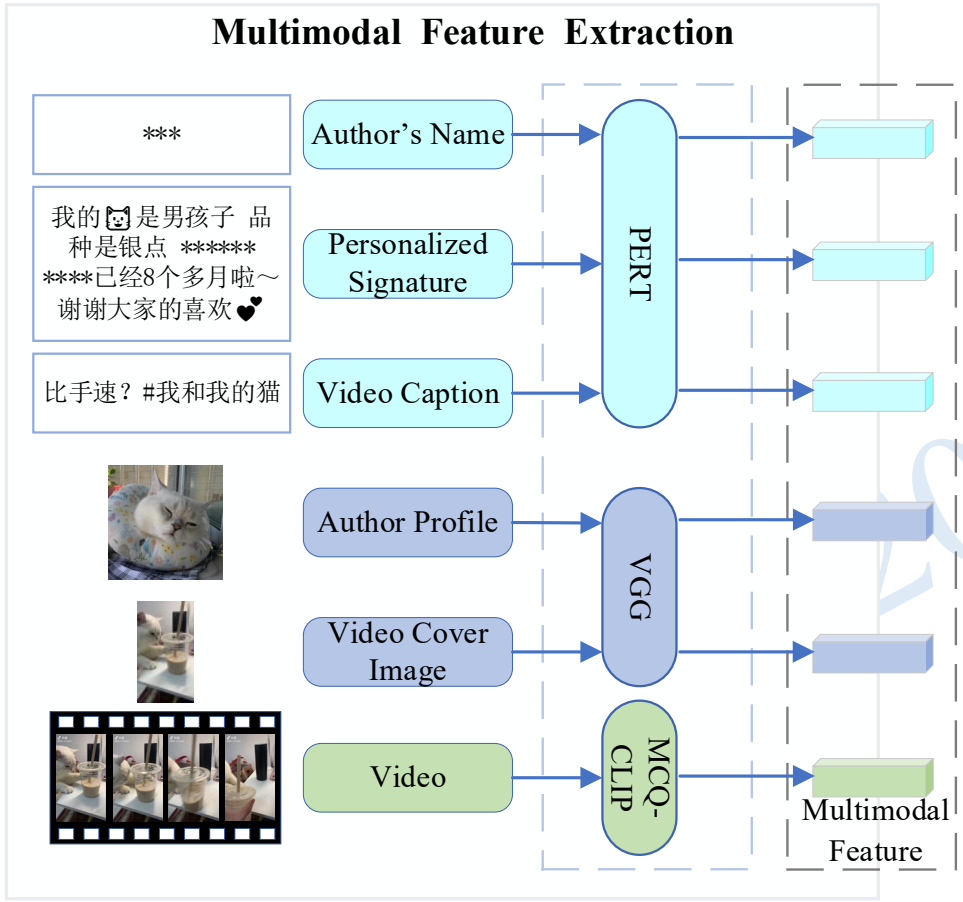
MMSVD-Douyin Dataset

- Histograms and the fitted kernel distributions of video duration, frame rate, ”comment”, ”like”, and ”share”

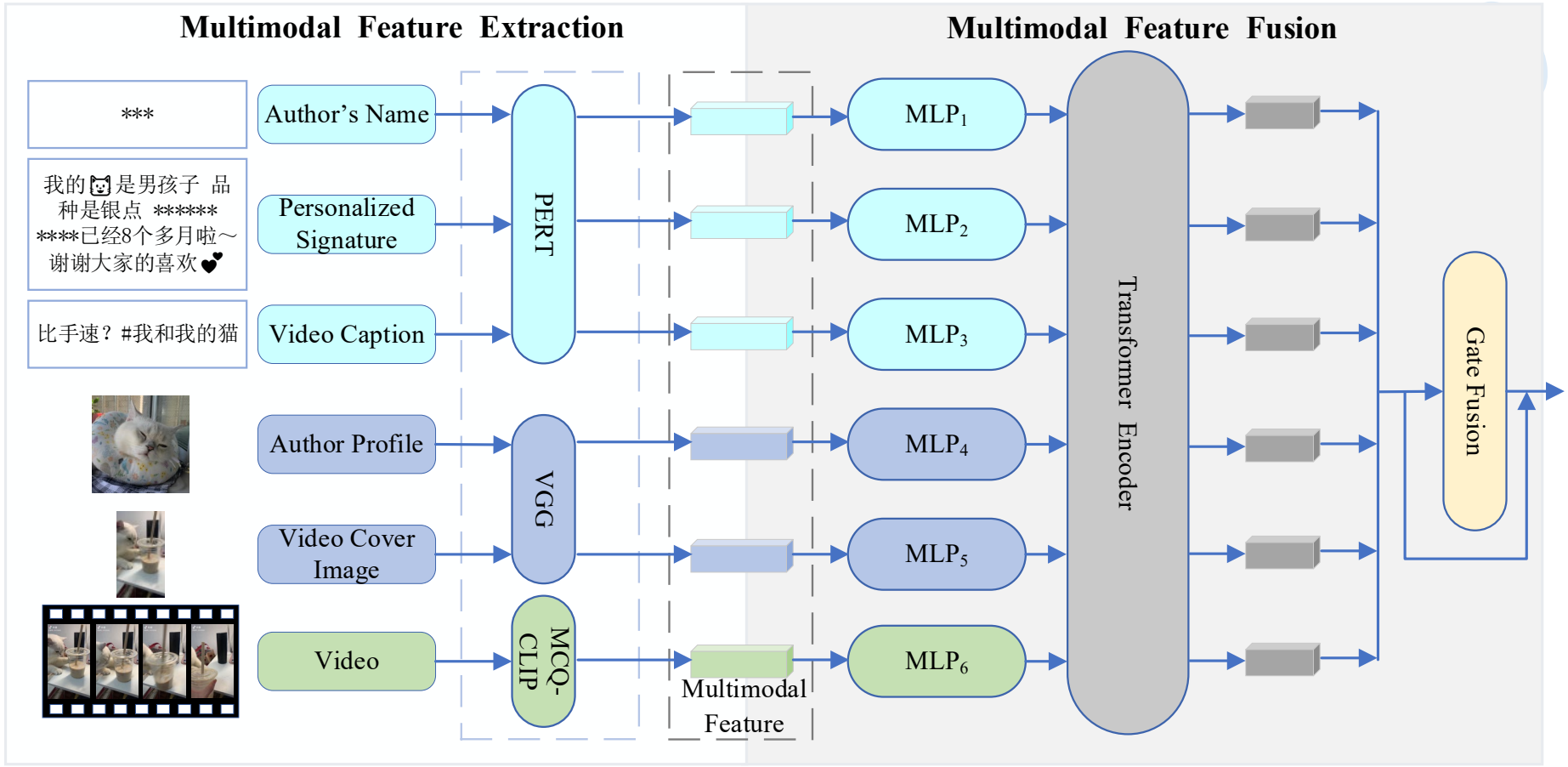
Long-tailed Distributions



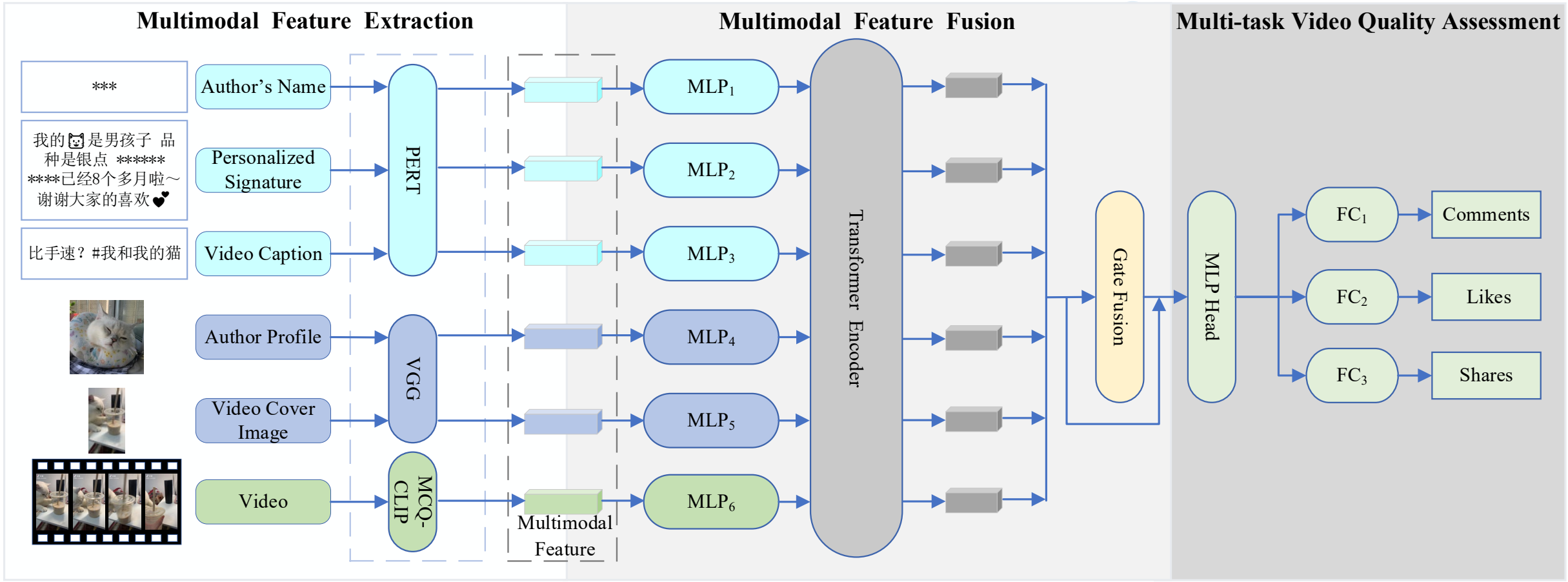
Multi-Modal Short Video Quality Assessment Benchmark



Multi-Modal Short Video Quality Assessment Benchmark



Multi-Modal Short Video Quality Assessment Benchmark



Results on MMSVD-Douyin

- Ablation study

Video	Text	Image	Criterion	MAE	Mean Error
✓			Comments	87,198	50,040
			Likes	855,611	361,993
			Shares	108,382	70,116
✓	✓		Comments	86,792	53,134
			Likes	832,529	377,522
			Shares	107,918	69,817
✓		✓	Comments	87,095	46,679
			Likes	854,247	245,329
			Shares	107,781	61,447

	Ground Truth	Prediction
Comments	58,618	69,454
Likes	1,158,660	1,212,929
Shares	58,045	72,910

	Ground Truth	Prediction
Comments	64,418	74,496
Likes	1,487,191	1,333,450
Shares	128,627	81,200

Results on MMSVD-Douyin

- Ablation study

Video	Text	Image	Criterion	MAE	Mean Error
✓			Comments	87,198	50,040
			Likes	855,611	361,993
			Shares	108,382	70,116
✓	✓		Comments	86,792	53,134
			Likes	832,529	377,522
			Shares	107,918	69,817
✓		✓	Comments	87,095	46,679
			Likes	854,247	245,329
			Shares	107,781	61,447

- Result with three modalities

Video	Text	Image	Criterion	MAE	Mean Error
✓	✓	✓	Comments	88,314	49,856
			Likes	846,417	254,633
			Shares	106,556	63,080

Conclusion

- MMSVD-Douyin: three kinds of modalities, six kinds of contents, three assessment indicators

- MulSVQA: all-around multi-modal short video quality assessment benchmark

- Future Study:
  - Add processing of **audio information**
  - Analyze the video's **emotional features**
  - Collect **comment data** of short video