FineVQ: Fine-Grained User Generated Content Video Quality Assessment

Supplementary Material

1. More Details of the Collected Videos

1.1. More Examples of the Collected Videos

We first show more sample videos to illustrate the abundant content in the FineVD database. As shown in Figure 1, for on-demand UGC videos, the content in FineVD covers *knowledge & technology & news, music & dancing, daily life, animation, fashion & entertainment, animal, sport, game, film & television* videos. Moreover, all categories contain both traditional UGC videos (landscape videos) and short-form videos (portrait videos), which indicates the wide coverage of videos. In particular, our FineVD also contains animation and game videos, which are typically ignored by previous databases [4, 10, 11].

Moreover, our FineVD also contains abundant live-streaming videos, which cover the categories of *mobile game, entertainment, single-player game, online game, wild & daily life, virtual streamer, multi-person interactive video, radio video*, as shown in Figure 2. The live-streaming videos also contain both traditional-form videos and short-

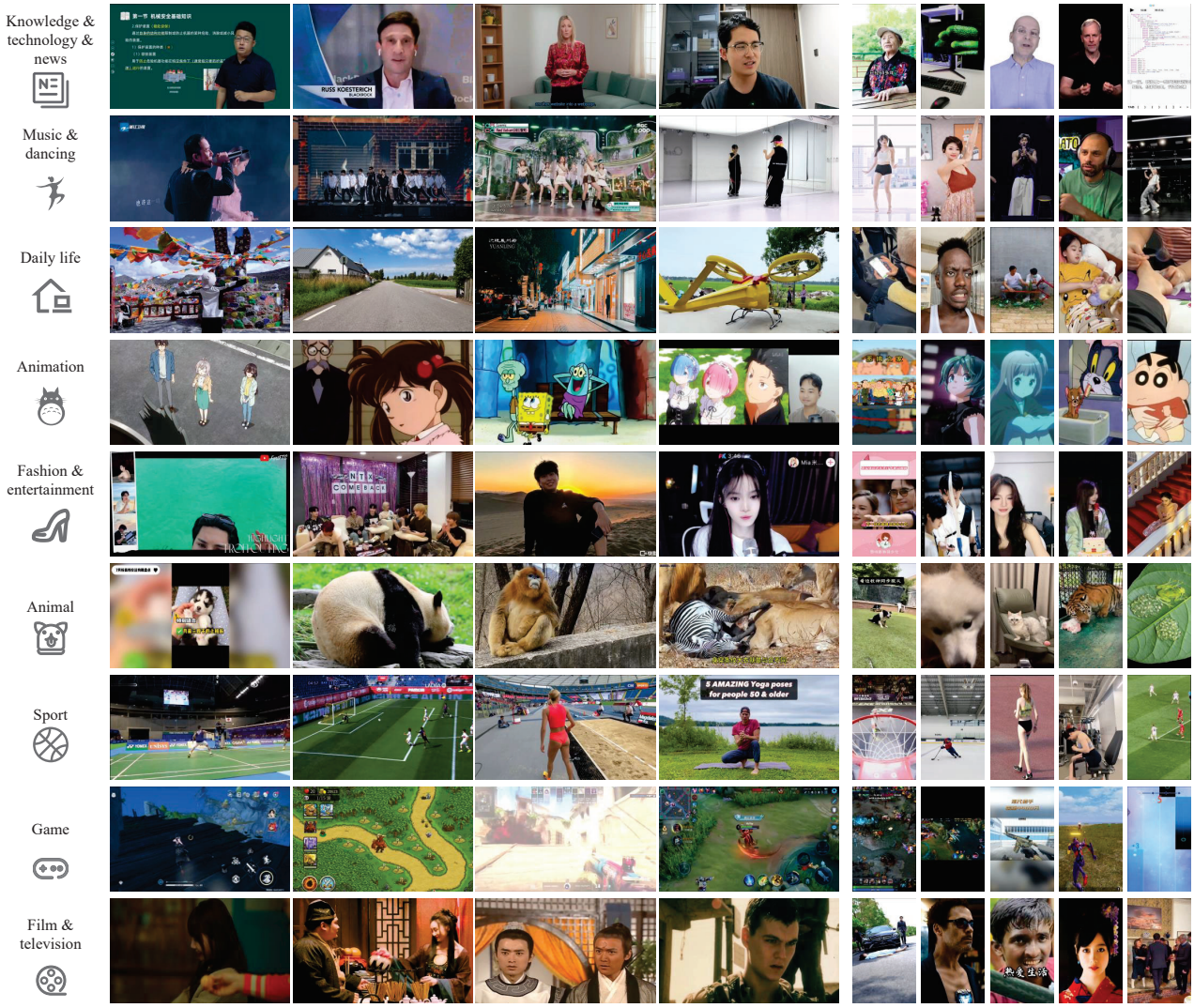

Figure 1. More examples of the on-demand UGC videos in our FineVD. Some videos are resized for better illustration.

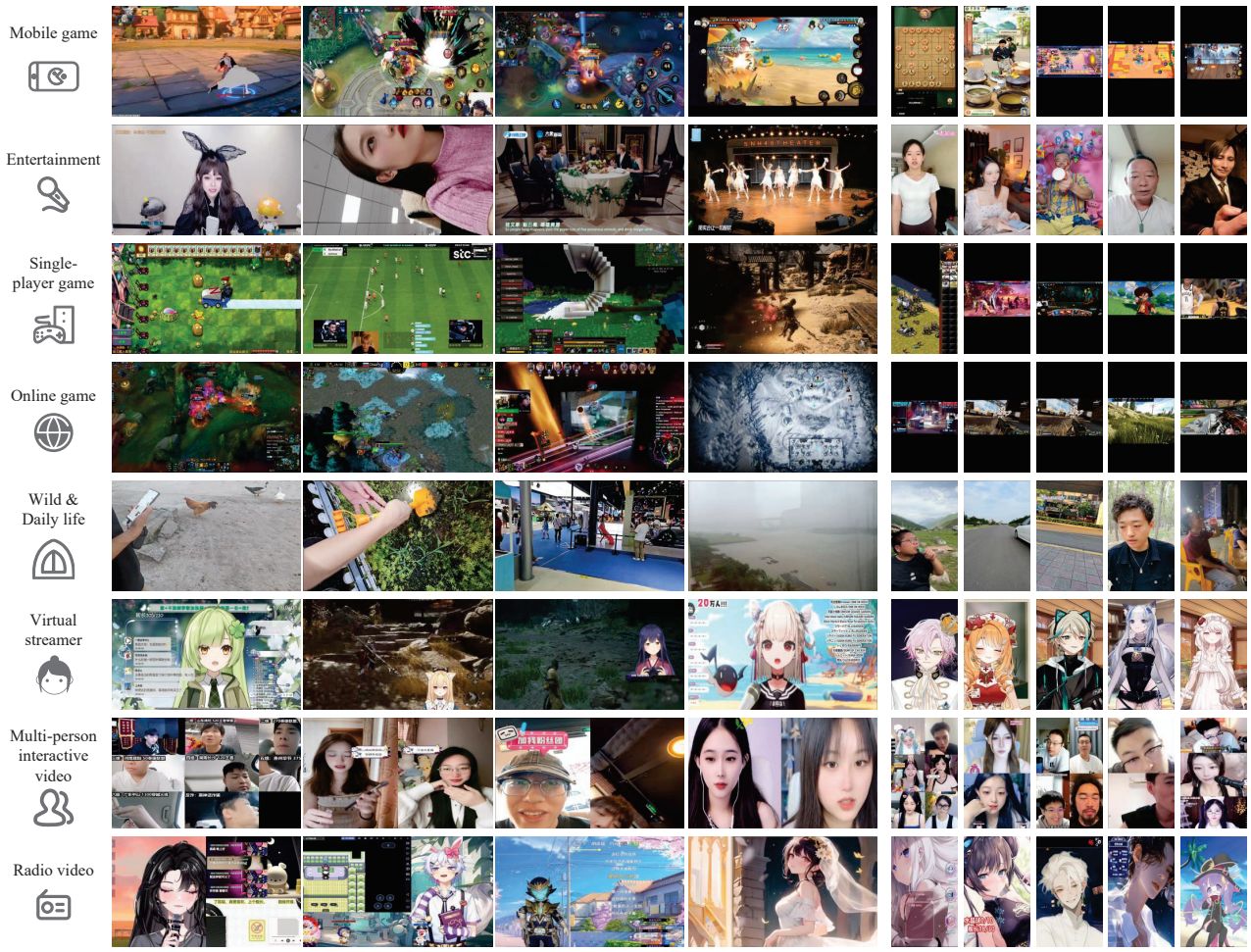

Figure 2. More examples of the live-streaming UGC videos in our FineVD. Some videos are resized for better illustration.

form videos. In particular, the virtual streamer videos and the radio videos should be distinguished, since most virtual streamer videos contain moving virtual characters, while most radio videos are static characters or wallpapers.

We also compare the constructed FineVD database with other popular public UGC VQA datasets [7]. As shown in Table 1, it can be observed that our FineVD database is the first UGC VQA database that contains multi-dimensional MOS annotations and fine-grained descriptions. Moreover, compared with the recent two databases TaoLive [15] and KVQ [6], the FineVD is also established in a lab environment, but contains more diverse video content and more fine-grained annotations.

1.2. Feature Analysis

As shown in Figure 3, the constructed FineVD database exhibits broad feature characteristics across five video quality-related features, including colorfulness, brightness, contrast, SI, TI. It can be observed that the majority of features span a wide range of normalized values, indicating the feature diversity inherent in out database. Specifically, the colorfulness coverage in Figure 3 is relatively uniform, which further manifests the video content in FineVD is abundant. Moreover, the SI and TI also cover a wide range, indicating that the video content in FineVD has both spatial richness and temporal richness.

1.3. Distortion Analysis

Figure 4 illustrates the differences between different dimensions. It can be observed that the five dimensions, *i.e.,* color, noise, artifact, blur, and temporal are significantly distinct in some cases. For example, as shown in the first row, the MOSs of the color dimension for the five videos are low, but the MOSs of other dimensions are relatively high. Moreover, we also notice that the overall score is generally in the middle of the worst dimension and other dimensions,

Table 1. An overview of popular public UGC VQA datasets.

Database	Type	Year	#Cont.	#Total	Resolution	FR	Dur.	Format	Distortions	#Subj.	#Ratings	Data	Env.
CVD2014 [8]	Aut.	2014	5	234	720p, 480p	9-30	10-25	AVI	In-capture	210	30	MOS	In-lab
LIVE-Qualcomm [2]	Aut.	2016	54	208	1080p	30	15	YUV	In-capture	39	39	MOS	In-lab
UGC-VIDEO [5]	Syn.+Aut.	2019	50	550	720p	30	10	N/A	UGC+compression	30	30	DMOS	In-lab
LIVE-WC [14]	Syn.+Aut.	2020	55	275	1080p	30	10	MP4	UGC+compression	40	40	MOS	In-lab
YT-UGC+(Subset) [12]	Syn.+Aut.	2021	189	567	1080p, 720p	Diverse	20	RAW+264	UGC+compression	N/A	30	DMOS	In-lab
ICME2021 [3]	Syn.+Aut.	2021	1000	8000	N/A	N/A	N/A	N/A	UGC+compression	N/A	N/A	MOS	In-lab
TaoLive [15]	Syn.+Aut.	2023	418	3762	1080p, 720p	20	8	MP4	UGC+compression	44	44	MOS	In-lab
KVQ [6]	Syn.+Aut.	2024	600	3600	Diverse	Diverse	8	MP4	UGC+compression	15	15	MOS+Rank	In-lab
KoNViD-1k [4]	Aut.	2017	1200	1200	540p	24-30	8	MP4	In-the-wild	642	114	MOS+σ	Crowd
LIVE-VQC [10]	Aut.	2018	585	585	1080p-240p	19-30	10	MP4	In-the-wild	4776	240	MOS	Crowd
YouTube-UGC [11]	Aut.	2019	1380	1380	4k-360p	15-60	20	MKV	In-the-wild	>8k	123	MOS+σ	Crowd
LSVQ [13]	Aut.	2021	39075	39075	Diverse	Diverse	5-12	MP4	In-the-wild	6284	35	MOS	Crowd
FineVD (Ours)	Aut.	2024	6104	6104	Diverse	Diverse	8	MP4	In-the-wild	22	22	MOS×6+σ×6+Descriptions	In-lab

Note: #Cont.: The number of unique video contents. #Total: Total number of test video sequences. FR: Framerate (in fps). Dur.: Video duration/length (in seconds). #Subj.: Total number of subjects in the study. #Ratings: Average number of subjective ratings per video. Env.: Subjective experiment environment. In-lab: Experiment was conducted in a laboratory. Crowd: Experiment was conducted by crowdsourcing. Syn.: Synthetic. Aut.: Authentic.

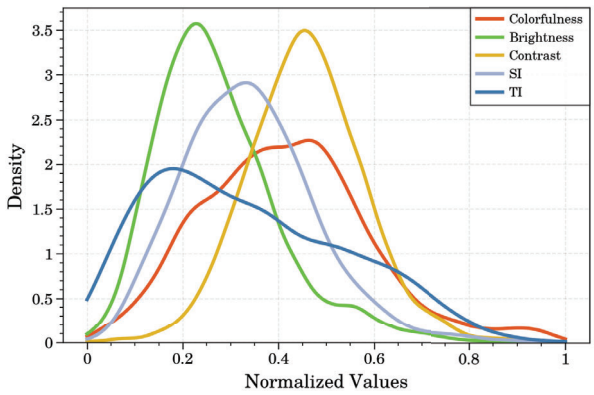

Figure 3. The feature distribution of FineVD. SI and TI indicate spatial information and temporal information, respectively.

which also manifests that the overall quality is significantly affected by the most severe distortion.

2. More Details of Subjective Study

2.1. Subjective Experiment Setup

The subjective experiment is conducted among 22 subjects. We use the 5-level rating method following the recommendation of ITU [9] to conduct the experiment, and the quality bar is labeled with five Likert adjectives, including "Bad, Poor, Fair, Good and Excellent", respectively. Given the peculiarities of the fine-grained quality assessment, we advise the subjects to give their opinion scores following the instructions:

Noise dimension: (1) Severe: There are serious image noise or particles in the video. The noise is very obvious, resulting in unclear content and seriously affecting the viewing experience. (2) Strong: The image noise in the video is obvious, affecting the viewing experience, but it does not affect the comprehensibility of the content. (3) Mild: The noise in the video exists, but it is not too significant and may affect the details. (4) Slight: In most cases, the image in the video has no obvious noise. The noise may exist for a short time. (5) Undistorted: There is almost no visible noise problem in the video.

Artifact dimension: (1) Severe: There are serious artifacts in the video, obvious block distortion, compression distortion or other obvious visual defects, which significantly interfere with the viewing experience. (2) Strong: The artifacts, flaws or distortions in the video are obvious, but they do not seriously affect the comprehensibility of the content. (3) Mild: The artifacts, flaws or distortions in the video exist, but they are not too obvious and have a slight impact on the video content. (4) Slight: The artifacts, flaws or distortions in the video are few and basically do not interfere with viewing. They may only be noticed in a few cases. (5) Undistorted: The video has almost no visible artifacts, flaws or distortions.

Blur dimension: (1) Severe: The video is extremely blurry, details are difficult to discern, and even the main objects or people cannot be identified, showing obvious pixelation or blurring effects, which seriously affects the viewing experience. (2) Strong: The video is obviously blurry, details are unclear, the main objects and outlines can be roughly identified, but lack clarity and fineness. (3) Mild: The blur in the video exists, but it is not too significant and may affect the details. (4) Slight: The video is basically clear, most objects and details can be clearly identified, with only slight or

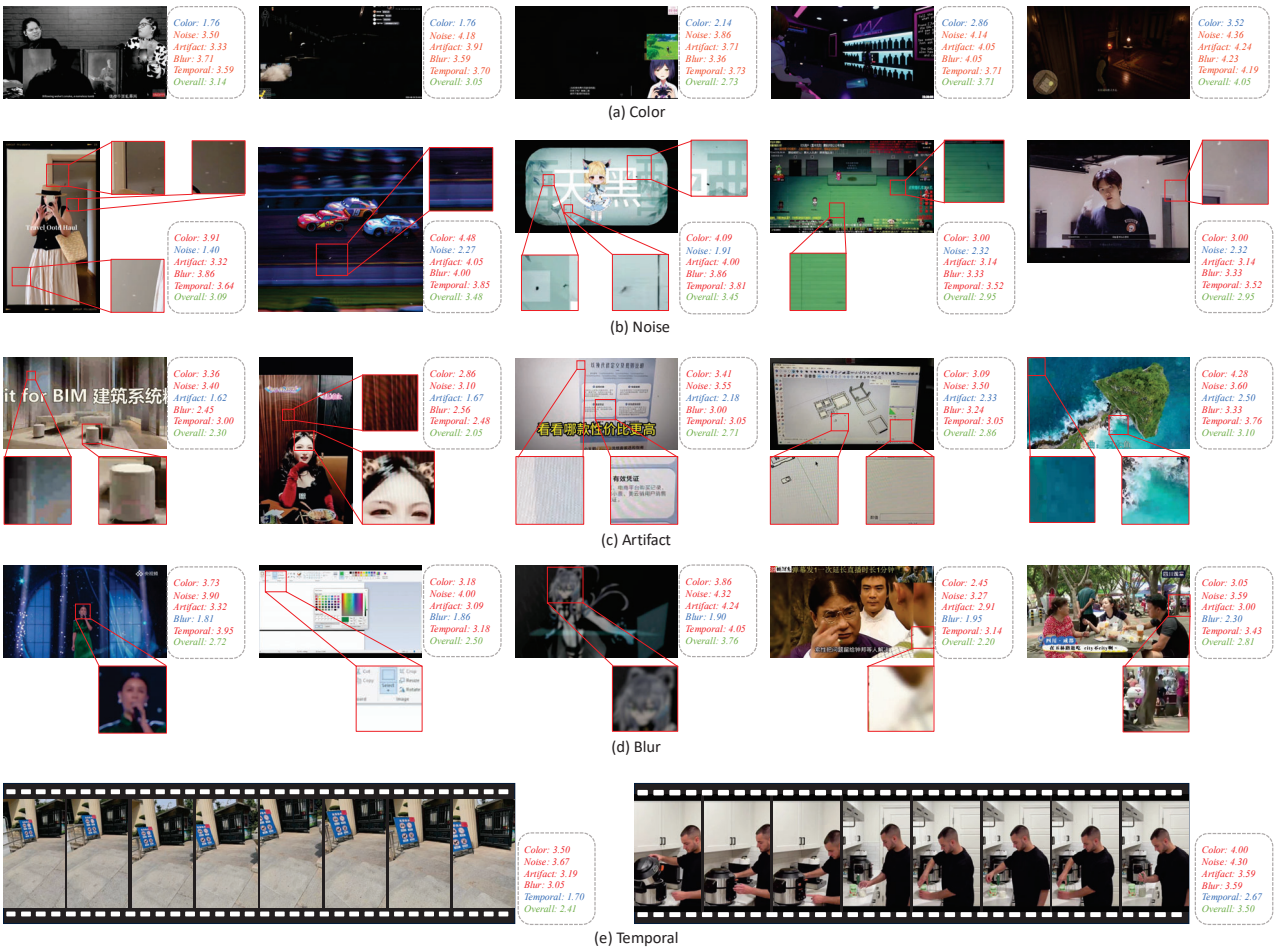

Figure 4. Illustration of the differences between different dimensions.

short-term blur. (5) Undistorted: The video is very clear, all objects and details can be clearly distinguished, and there is almost no blur phenomenon.

Color dimension: (1) Bad: There are obvious defects in color, such as unrealistic colors, hue deviation, extreme exposure or low-light. Or some obvious color errors can be observed, the colors are unnatural, and the viewing experience is seriously affected. (2) Poor: The color display is inaccurate, but not to an extreme degree. There are some problems with color contrast, saturation, and brightness, but they will not seriously affect the comprehensibility or viewing experience of the content. (3) Fair: The video color is fair, with slight problems with color contrast, saturation, and brightness. (4) Good: The color display is basically accurate, and there is no obvious distraction when watching, but the color may not be particularly attractive. (5) Excellent: The color display is completely accurate, with no visible distortion or offset. The color is attractive.

Temporal dimension: (1) Bad: The video frequently has obvious frame rate problems, strobing, jitter or stuttering, resulting in very incoherent content. (2) Poor: The video has unstable frame rate, occasional frame skipping or slight strobing, jitter or stuttering, which obviously interferes with the coherence of the content and the viewing experience. (3) Fair: The video occasionally has slight frame rate jumps, strobing, jitter or stuttering, but it does not significantly affect the understanding of the content. (4) Good: The video maintains a stable frame rate and smooth playback in most cases. In a few cases, there is slight instability, but it does not strongly affect the overall viewing experience. (5) Excellent: The video has excellent temporal consistency, with almost no frame rate problems, strobing, jitter or stuttering, and the overall viewing experience is smooth and coherent.

Overall dimension: (1) Bad: The video quality is extremely poor, with serious issues in color, noise, artifact, blur, and temporal dimensions, which greatly affect the

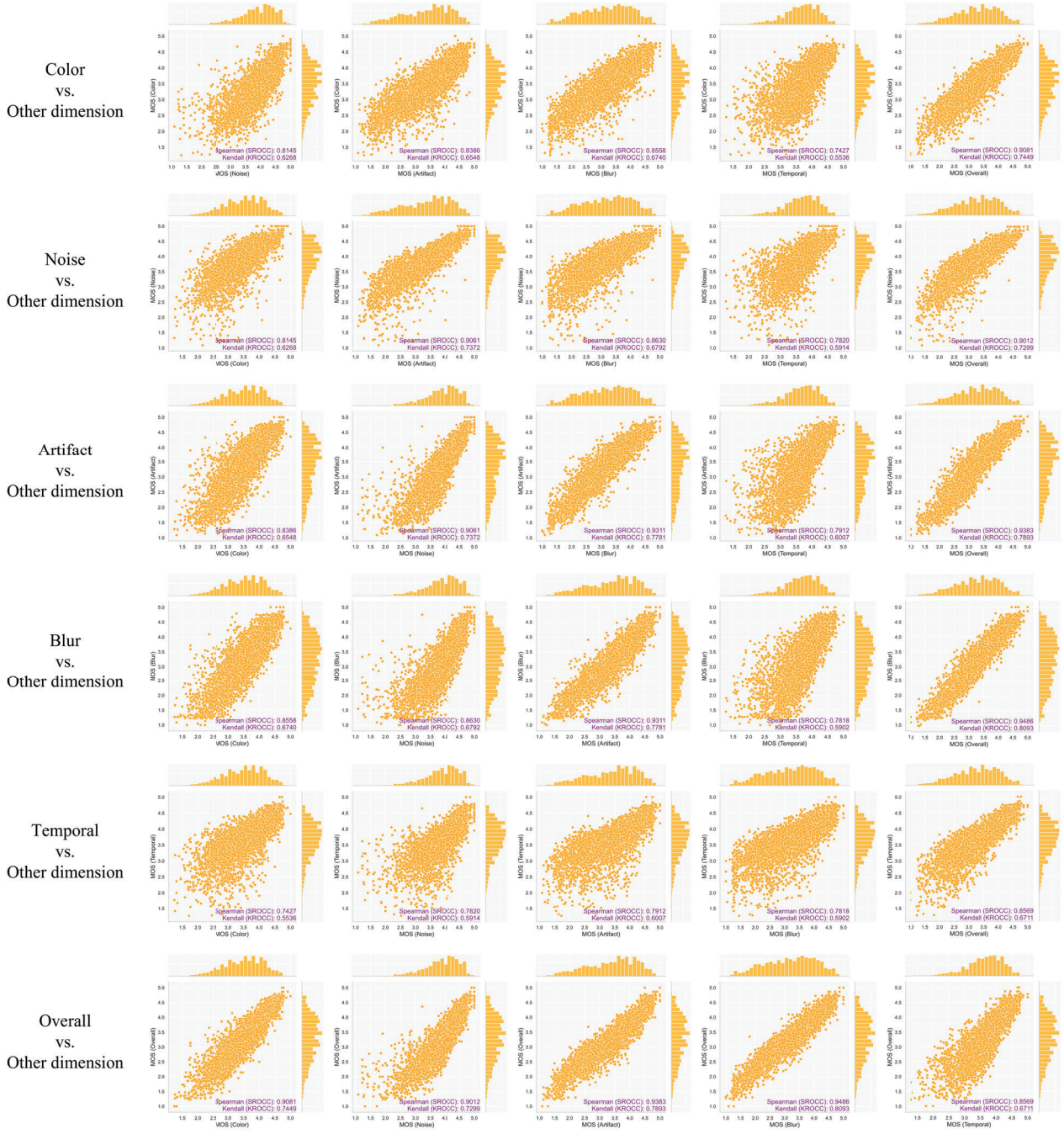

Figure 5. Illustration of the MOS correlation between any two dimensions.

viewing experience. (2) Poor: The video quality is poor, with obvious issues in color, noise, artifact, blur, and temporal dimensions, which affect the viewing experience. (3) Fair: The video quality is fair, with certain problems in color, noise, artifact, blur, and temporal dimensions, which affect the viewing experience to some extent. (4) Good: The video quality is good, the content is relatively clear, and there is no significant distortion. (5) Excellent: The video quality is excellent, the content is clear and rich in details, the colors are vivid and realistic, the picture is stable and smooth, and there is no distortion.

For the quality attribute annotation, we ask the subjects

to annotate obvious distortions in the video by selecting the options. If there is no corresponding distortion in the existing options, the subjects can manually enter the distortion.

2.2. Subjective Data Processing

For the quality data screening process, we first calculate the kurtosis score of the raw subjective quality ratings for each image to detect it is a Gaussian case or a non-Gaussian case. Then, for the Gaussian case, the raw score for an image is considered to be an outlier if it is outside 2 standard deviations (stds) about the mean score of that image; for the non-Gaussian case, it is regarded as an outlier if it is outside $\sqrt{20}$ stds about the mean score of that image. A subject is removed if more than 5% of his/her evaluations are outliers. For the quality attribute labeling, we choose the options selected by more than half of the subjects as the quality attribute labels. To generate question-answering (QA) pairs, we query "yes-or-no" question for all five dimensions, and additional two "which exist" and "which most affect" questions over all dimensions. An overall quality QA pair is also generated.

As a result, we finally obtain numerous fine-grained quality labels for 6104 UGC videos, including 36624 MOSs (6104×6) and 48832 (6104×(5+2+1))) QA pairs.

2.3. Correlation of Different Dimensions

To further understand human perceptual quality differences across different dimensions, we also analyze the correlation between any two dimensions in our FineVD database. As shown in Figure 5, the correlations are significantly different for different dimension pairs. First of all, we observe that the color dimension has relatively low correlations with the noise, artifact, blur, and temporal dimensions, but a relatively high correlation with the overall dimension, which manifests that color is a distinct assessment dimension and significantly influence the overall quality. The noise dimension has a relatively high correlation with the artifact dimension, indicating these two dimensions are related to some extend. Moreover, the artifact and the blur dimensions exhibit a strong correlation, likely because artifacts often cause blurring. The temporal dimension has the lowest correlations with all other dimensions, manifesting that the temporal dimension is the most special evaluation dimension. Finally, the overall dimension has high correlations with almost all dimensions, which further illustrates that the overall quality rating is influenced by all dimensions.

3. More Details of Our FineVQ Model

3.1. Loss Functions

We use both language loss and L1 loss as the loss functions to optimize the training process. Specifically, the language loss is used to restrict the FineVQ to produce specific quality attribute, while L1 loss is used to regress the quality scores. The language loss function can be formulated as:

$$\mathcal{L}_{\text{language}} = -\frac{1}{N}\sum_{i=1}^{N}\log P(y_{\text{label}}|y_{\text{pred}}), \tag{1}$$

where y_{pred} is the predicted token, y_{label} is the ground truth token, $P(y_{\text{label}}|y_{\text{pred}})$ indicates the probability, N is the number of tokens. The L1 loss can be formulated as:

$$\mathcal{L}_1 = \frac{1}{N}|q_{\text{pred}} - q_{\text{label}}|, \tag{2}$$

where q_{pred} is the predicted quality score, q_{label} is the ground truth quality score, N is the number of videos in a batch. The overall loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{language}} + \mathcal{L}_1. \tag{3}$$

3.2. Model and Training Details

We further describe the dimension of FineVQ in detail. For E_I, the output feature dimension is 4096, then two MLPs with the dimension of 4096 is followed to refine the features. For E_M, the output feature dimension is 2304, the two MLPs are followed, which map the feature dimension form 2304 to 4096. Then, the extracted features are fed into the LLM, whose feature dimension is also 4096. During training, the image encoder E_I, motion encoder E_M, text encoder and decoder, and the large language model are frozen, while the projectors and the LoRA weights are trainable.

4. More Experimental Results

4.1. Influence of Extracted Video Frames

We further conduct an ablation experiment to study the influence of the selected video frame number of image encoder E_I and motion encoder E_M, respectively. As shown in Table 2, reducing frame numbers for both image encoder E_I and motion encoder E_M can decrease the performance. Specifically, comparing the first, second and the last rows in Table 2, we can observe that setting the selected frames to 8 for E_I leads to better performance compared to 4 frames and 1 frame. Moreover, it can be observed that reducing the input frame numbers of E_M also lower the final performance. Thus, the video frame selection is important in our FineVQ model.

4.2. Improvement of Instruction Tuning on the Attribute Prediction Task

We further compare the performance of FineVQ and our base model, *i.e.,* InternVL2 (*8B*) [1] on the *quality attribute prediction* task. It can be observed that the established

Table 2. Influence of extracted video frames. F indicates the whole frames.

Strategy		LIVE VQC [10]		
E_I frames	E_M frames	SRCC	PLCC	KRCC
1	F	0.8474	0.8657	0.6743
4	F	0.8609	0.8792	0.6905
8	$F/16$	0.8354	0.8330	0.6444
8	$F/4$	0.8492	0.8525	0.6739
8	F	**0.8951**	**0.8950**	**0.7297**

Table 3. Comparison between FineVQ and the base model InternVL2 (*8B*) [1] on our established FineVD database in terms of the *quality attribute prediction task*. The *"yes-or-no"* type represents the judgment on whether the corresponding dimension is degraded. The *"which"* type indicates which distortion exists or has the most impact on the quality of the video.

Question Type	*Yes-or-no*					*Which*	
Model / Attribute	Color	Noise	Artifact	Blur	Temporal	Exist	Most
InternVL2 (*8B*) [1]	58.46%	63.58%	50.69%	54.33%	70.28%	28.25%	43.21%
FineVQ (Ours)	**73.52%**	**72.74%**	**51.87%**	**64.76%**	**86.91%**	**91.93%**	**65.06%**
Improvement	**15.06%**	**9.16%**	**1.18%**	**10.43%**	**16.63%**	**63.68%**	**21.85%**

FineVD database and FineVQ model can significantly improve the low-level quality attribute perception ability for the base model, with increasing over 10% for almost all sub-tasks.

References

[1] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2024. 6, 7

[2] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 28(9): 2061–2077, 2017. 3

[3] Wang Haiqiang, Li Gary, Liu Shan, and Kuo C.-C. Jay. Icme 2021 ugc-vqa challenge. http://ugcvqa.com/, 2021. [Online]. 3

[4] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *Proceedings of the IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2017. 1, 3

[5] Yang Li, Shengbin Meng, Xinfeng Zhang, Shiqi Wang, Yue Wang, and Siwei Ma. Ugc-video: Perceptual quality assessment of user-generated videos. In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 35–38, 2020. 3

[6] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25963–25973, 2024. 2, 3

[7] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. Perceptual video quality assessment: A survey. *Science China Information Sciences*, 67(11):211301, 2024. 2

[8] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. Cvd2014—a database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing (TIP)*, 25(7):3073–3086, 2016. 3

[9] BT Series. Methodology for the subjective assessment of the quality of television pictures. *Recommendation ITU-R BT*, pages 500–13, 2012. 3

[10] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing (TIP)*, 28(2):612–627, 2018. 1, 3, 7

[11] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5, 2019. 1, 3

[12] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. Rich features for perceptual quality assessment of ugc videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13435–13444, 2021. 3

[13] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq:'patching up'the video quality problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14019–14029, 2021. 3

[14] Xiangxu Yu, Neil Birkbeck, Yilin Wang, Christos G Bampis, Balu Adsumilli, and Alan C Bovik. Predicting the quality of compressed videos with pre-existing distortions. *IEEE Transactions on Image Processing (TIP)*, 30:7511–7526, 2021. 3

[15] Zicheng Zhang, Wei Wu, Wei Sun, Danyang Tu, Wei Lu, Xiongkuo Min, Ying Chen, and Guangtao Zhai. Md-vqa: Multi-dimensional quality assessment for ugc live videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1746–1755, 2023. 2, 3