

# Pioneering Perceptual Video Fluency Assessment: A Novel Task with Benchmark Dataset and Baseline *Supplementary Material*

Qizhi Xie<sup>1,2</sup>, Kun Yuan<sup>2</sup>✉, Yunpeng Qu<sup>1,2</sup>, Ming Sun<sup>2</sup>, Chao Zhou<sup>2</sup>, Jihong Zhu<sup>1</sup>✉  
<sup>1</sup> Tsinghua University, <sup>2</sup>Kuaishou Technology

xqz20@mail.tsinghua.edu.cn, yuankun03@kuaishou.com, jhzhu@tsinghua.edu.cn

## 1. More Details of *FluVid*

### 1.1. Scoring Criteria

As described in the method section of the manuscript, we design a *novel* five-point Absolute Category Rating (ACR) standard, meticulously adhering to the ITU regulations [13–16], designed to evaluate the **perceptual fluency** of 4,606 candidate videos. It emphasizes the continuity of video across the temporal dimension, such as the stability of frame rates and the consistency of motion. Here, we conclude the score and its corresponding visual descriptions and representative scenarios (double-checked and approved by 10 visual experts) in Tab. 1, which guide 20 annotators during the human study.

### 1.2. Annotation Process

In the human study section, we describe the annotation process of FluVid. Here, we provide the annotation GUI as in Fig.1. On the left side, annotators need to watch the video playback thoroughly, focusing on the fluency of foreground objects, background, and camera movement. Viewers are allowed to watch the playback repeatedly. At the same time, annotators need to give their fluency rating on the right panel, following the criteria in Tab.1, as all of them are trained by scoring 100 anchor videos. We collect the final scores of 20 annotators on each video and average them for the final results.

### 1.3. Comparison to Existing Perceptual Datasets

In the related work and method section, we highlight that FluVid is the first-ever perceptual benchmark dataset focused on fluency. Here, we provide a comprehensive comparison among FluVid and other perceptual datasets (mostly in VQA task) in Tab.2. We discover that most VQA datasets only provide holistic video quality scores (*e.g.*, LIVE-VQC [17], KoNViD-1k [6] *etc.*). While some datasets provide scores on one or more quality dimensions, such as DOVER [23] and LSVQ [28]. They do not disen-

tangle fluency and pay insufficient attention to the temporal dimension. Additionally, though Maxwell [25] provides scores partly reflecting temporal quality (T-8), its scoring criteria are only a three-tier system, including a ternary choice of “bad”, “neutral”, and “good”. This three-tier system is not fine-grained enough to align with the human visual system (HVS) and has not been proven by a large-scale human study to be as robust as our proposed five-tier rating. FineVQ [2] constructs a VQA database FineVD, which includes temporal quality scores besides overall quality scores. However, it focuses on addressing the VQA task, treating fluency merely as a sub-dimension of overall quality. Conversely, we treat fluency as the sole focus of the newly proposed VFA. Moreover, we conduct an in-depth analysis of fluency and dissect its factors into three video components, as in the method section. Last but not least, the research focus of FineVD is on the language instructions for large multimodal models (LMMs).

## 2. More Details of *FluNet*

### 2.1. Limitations of Synthesis and Mitigation

In the method section, we emphasize that fluency is very complicated, sometimes dominated by foreground, background, or camera. Still, we only drop and duplicate the whole frame when synthesizing stuttered videos, as in Fig.5 and Alg.1 of our paper. A potential limitation for this “drop-and-duplicate” strategy is that the synthesized stutter cannot cover the complex real-world scenarios. To mitigate this, we further fine-tune FluNet on 606 videos with accurate fluency scores to enhance its generalization ability. In future work, we will 1) introduce more sophisticated synthesis, such as applying the video segmentation techniques to change the fluency level of foreground, background, and camera, respectively. 2) Scale up FluVid to enable more videos for supervised training.

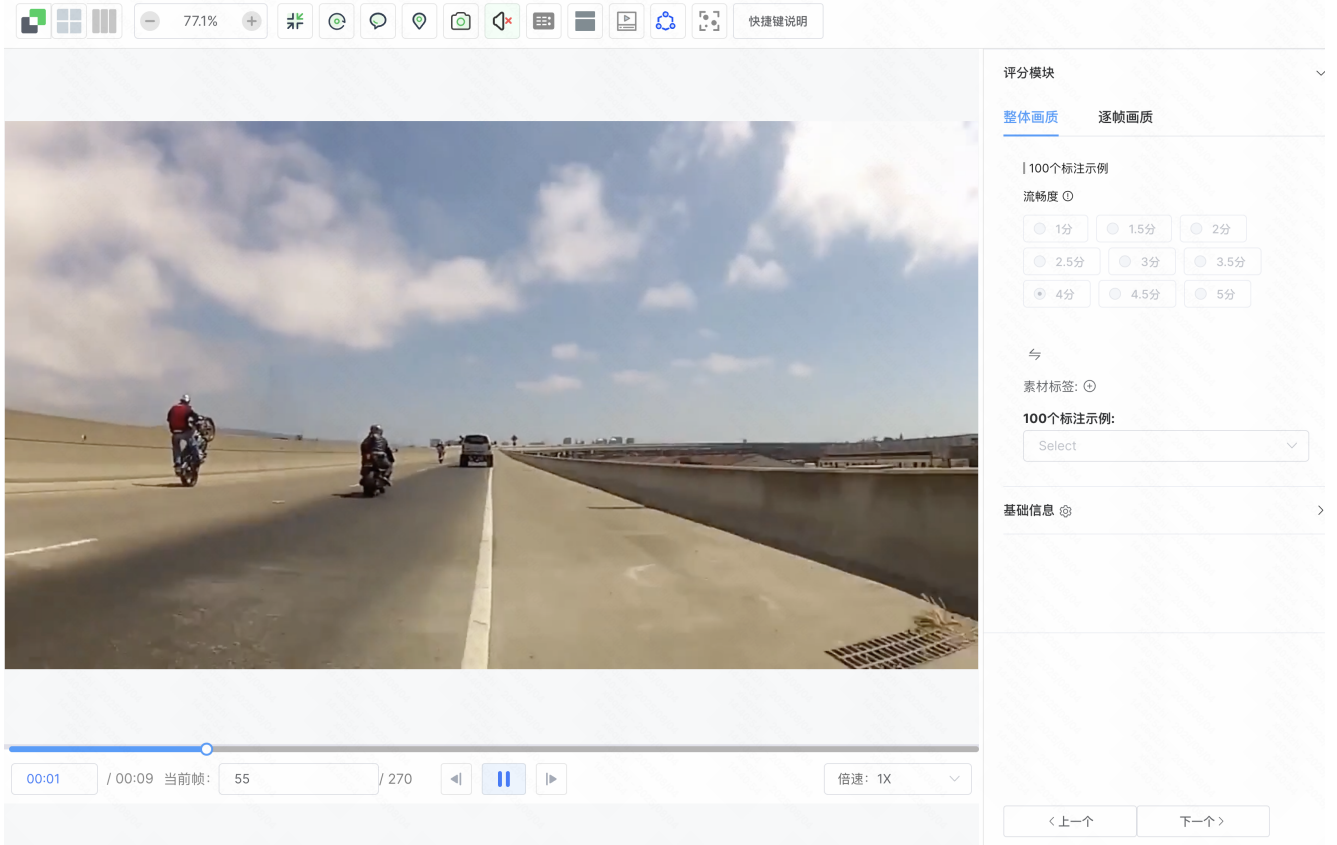


Figure 1. Annotation GUI of human study.

## 2.2. Examples of Synthesized Videos

To better illustrate our idea, we provide several exemplary synthesized videos with various drop rates in the supplementary material. By modifying the drop rate  $r_k$ , we can control the fluency level in a fine-grained manner.

## 3. More Details of Experiments

### 3.1. Rationale of Model Participant Selection

In the experimental settings section, we briefly introduce 23 model participants. Here, we explain the rationale behind our selection. First, we select VQA methods based on their performance on common VQA benchmarks, including LSVQ, LIVE-VQC, KoNViD-1k and YouTube-UGC [21]. Thus, six state-of-the-art VQA methods are selected. Second, large multimodal models (LMMs) begin to excel in multiple vision tasks, including high-level (*e.g.*, classification, detection *etc.*) and low-level tasks (*e.g.*, deblurring). Therefore, it is imperative to benchmark them on our VFA task. Among them, video LMMs are selected due to their superior ability to process videos. Specifically, we choose seven video LMMs with best performance on general video benchmarks, including VideoMME [5],

MMBench-Video [3], MVBench [9], and MLVU [30]. We also choose two video LMMs specialized in VQA, including FineVQ and VQA<sup>2</sup>-Scorer. Recent literature suggests image LMMs exhibit primary quality perception ability. Hence, we select three image LMMs with top-tier performance on general image benchmarks, including MME [27], MMBench [11], MMMU [29], MTVQA [19], and two image LMMs specialized in image quality assessment (IQA).

### 3.2. Benchmark Settings

In the experimental settings section, we state the detailed benchmark setting on FluVid, including evaluation prompts and the softmax strategy. Here, we provide additional details for clarity. In terms of the number of frames, we use the default setting of each model to activate the best performance for fairness. For VQA models, the frame numbers for Fast-VQA [22], Faster-VQA [24], DOVER [23], SimpleVQA [18], and PVQ [28] are 32, 16, 32, 32 and 16, respectively. Note that, for SimpleVQA and PVQ, we pre-extract the motion features using the pre-trained SlowFast-ResNet-50 [4]. For video LMMs, the frame numbers or frame rate for Video-LLaVA [10], Chat-UniVi-v1&v1.5 [8], LLaMA-VID [20], Video-ChatGPT [12], PLLaVA [26],

Table 1. The annotation criteria for subjective labeling scores range from 1 to 5. Each scoring tier features its definition, visual description, and three representative scenarios. Please zoom in for a better view.

Score	Definition	Visual Description	Representative Scenarios
1 Bad	The temporal sequence is severely disrupted and incoherent.	The temporal sequence is exceedingly choppy, marked by frequent and noticeable stutters, frame drops, and drastic fluctuations in frame rate, rendering the actions difficult to discern coherently and causing significant discomfort during viewing.	<ol style="list-style-type: none"> <li>1. Evident frame drops result in fragmented movements, such as characters exhibiting a "jumping" motion while walking;</li> <li>2. In fast-paced scenes, such as sporting events, severe motion blur or screen tearing occurs, leading to discontinuous object trajectories;</li> <li>3. In animated or gaming videos, characters display unnatural pauses in their actions, with running movements stuttering once every second.</li> </ol>
2 Poor	The temporal sequence is noticeably disjointed, hindering the viewing experience.	The temporal sequence exhibits clear stuttering and instability, with distortions readily apparent; although one can barely discern the actions, the viewing experience is frequently interrupted, diverting attention.	<ol style="list-style-type: none"> <li>1. Occasionally, frame duplication occurs, resulting in an alternating sequence of "stuttering and jumping" actions;</li> <li>2. In slow-motion scenes, such as when a character speaks, there is a slight disjunction in head movements or gestures, manifesting as 1 to 2 stutters per second;</li> <li>3. The rhythm of animated characters' movements is uneven, with variations in stride length during walking, oscillating between large and small steps.</li> </ol>
3 Fair	The temporal sequence is moderately smooth, with occasional imperfections.	The temporal sequence is generally coherent, though it experiences occasional fluctuations in fluidity; one must concentrate to perceive the imperfections, which do not significantly hinder the overall comprehension of the content.	<ol style="list-style-type: none"> <li>1. In the long shot, there is a brief moment of stutter, manifesting as a 0.5-second buffering trace;</li> <li>2. During rapid panning shots, the edges of background objects exhibit a subtle, almost imperceptible blurring discontinuity, as seen in the slight jerking of grass textures;</li> <li>3. In scenes with multiple characters, the movements of secondary roles are slightly out of sync, such as the background actors lagging just behind the rhythm of the protagonists.</li> </ol>
4 Good	The temporal sequence flows gracefully, approaching a natural cadence.	The temporal sequence flows seamlessly and naturally, with virtually no stuttering or fluctuations in frame rate; any temporal imperfections are scarcely perceptible during viewing, and the continuity of movement approaches a professional filming standard.	<ol style="list-style-type: none"> <li>1. In high-speed motion scenes, dynamic blur appears naturally, as exemplified by the seamless clarity of tire treads while racing;</li> <li>2. The intricate movements of complex actions, such as dance or martial arts, flow together smoothly without any frame drops;</li> <li>3. During camera transitions or zooms, the shifts are fluid and flicker-free, reminiscent of cinematic-grade editing.</li> </ol>
5 Excellent	The sequence is exquisitely fluid, without any imperfections.	The temporal sequence is exquisitely coherent and smooth, with a perfectly stable frame rate; the motion portrayal aligns seamlessly with real-world perception, allowing for complete immersion in the content, free from any temporal distractions.	<ol style="list-style-type: none"> <li>1. There are no frame losses, repetitions, or misalignments; the sequence of frames in the bitstream adheres strictly to the order of capture.</li> <li>2. In ultra-high-speed motion scenes, such as the flight of a bullet or the collision of water droplets, each frame transitions with precision, and the dynamic blur aligns with the laws of physics.</li> <li>3. In virtual reality (VR) videos, the visuals are rendered in real-time without delay during head movements, effectively eliminating the temporal sources of "motion sickness."</li> </ol>

Table 2. An overview of current public perceptual datasets.

Dataset	Year	Duration/s	Ref. Num.	Scale	Scope	Subjective Evaluation Format
CVD2014	2014	10-25	-	234	In-capture VQA	In-lab
Live-Qualcomm	2016	15	-	208	In-capture VQA	In-lab
KoNViD-1k	2017	8	-	1,200	In-the-wild VQA	Crowdsourced
LIVE-VQC	2018	10	-	585	In-the-wild VQA	Crowdsourced
YouTube-UGC	2019	20	-	1,500	In-the-wild VQA	Crowdsourced
LSVQ	2021	5-12	-	39,075	In-the-wild VQA	Crowdsourced
Maxwell	2023	9	-	4,543	In-the-wild VQA	Crowdsourced
FineVQ	2024	8	-	6104	In-the-wild VQA	In-lab
UGC-VIDEO	2019	>10	50	550	UGC w. compression	In-lab
LIVE-WC	2020	10	55	275	UGC w. compression	In-lab
YouTube-UGC+	2021	20	189	567	UGC w. compression	In-lab
<b>FluVid (Ours)</b>	2025	2-20	100	4,606	In-the-wild + UGC on VFA	In-lab

Qwen 2.5-VL [1], VQA<sup>2</sup>-Scorer [7], and FineVQ [2] are eight, 64, 1FPS, 100, 16, 2FPS, 1FPS, and eight frames, respectively. For all eight image LMMs, we uniformly sample 64 frames and average the per-frame scores for evaluation.

## References

- [1] Shuai Bai et al. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025. 3
- [2] Huiyu Duan et al. Finevq: Fine-grained user generated content video quality assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3206–3217, 2025. 1, 3
- [3] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024. 2
- [4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6201–6210. IEEE, 2019. 2
- [5] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 2
- [6] Vlad Hosu et al. The konstanz natural video database (konvid-1k). In *QoMEX*, pages 1–6. IEEE, 2017. 1
- [7] Ziheng Jia et al. Vqa<sup>2</sup>: Visual question answering for video quality assessment. *CoRR*, abs/2411.03795, 2024. 3
- [8] Peng Jin et al. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*, pages 13700–13710. IEEE, 2024. 2
- [9] Kunchang Li et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206. IEEE, 2024. 2
- [10] Bin Lin et al. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, pages 5971–5984. Association for Computational Linguistics, 2024. 2
- [11] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 2
- [12] Muhammad Maaz et al. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL (1)*, pages 12585–12602. Association for Computational Linguistics, 2024. 2
- [13] ITU-R Rec. T.81: Digital compression and coding of continuous-tone still images, 1992. <https://www.itu.int/rec/T-REC-T.81>. 1
- [14] ITU-R Rec. Bt.500: Methodologies for the subjective assessment of the quality of television images, 2000. <https://www.itu.int/rec/R-REC-BT.500>.
- [15] ITU-R Rec. Bt.2022: General viewing conditions for subjective assessment of quality of sdtv and hdtv television pictures on flat panel displays, 2012. <https://www.itu.int/rec/R-REC-BT.2022>.
- [16] ITU-R Rec. P.910: Subjective video quality assessment methods for multimedia applications, 2021. <https://www.itu.int/rec/t-rec-p.910>. 1
- [17] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE transactions on Image Processing*, 19(6):1427–1441, 2010. 1
- [18] Wei Sun et al. A deep learning based no-reference quality assessment model for UGC videos. In *ACM Multimedia*, pages 856–865. ACM, 2022. 2
- [19] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024. 2
- [20] Hugo Touvron et al. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. 2
- [21] Yilin Wang et al. Youtube UGC dataset for video compression research. In *MMSP*, pages 1–5. IEEE, 2019. 2
- [22] Haoning Wu et al. FAST-VQA: efficient end-to-end video quality assessment with fragment sampling. In *ECCV (6)*, pages 538–554. Springer, 2022. 2
- [23] Haoning Wu et al. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, pages 20087–20097. IEEE, 2023. 1, 2
- [24] Haoning Wu et al. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):15185–15202, 2023. 2
- [25] Haoning Wu et al. Towards explainable in-the-wild video quality assessment: A database and a language-prompted approach. In *ACM Multimedia*, pages 1045–1054. ACM, 2023. 1
- [26] Lin Xu et al. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *CoRR*, abs/2404.16994, 2024. 2
- [27] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024. 2
- [28] Zhenqiang Ying et al. Patch-vq: ‘patching up’ the video quality problem. In *CVPR*, pages 14019–14029. Computer Vision Foundation / IEEE, 2021. 1, 2
- [29] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoyi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 2
- [30] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv e-prints*, pages arXiv–2406, 2024. 2