Score2Instruct: Scaling Up Video Quality-Centric Instructions via Automated Dimension Scoring
Supplementary Material

Qizhi Xie[1,2], Kun Yuan[2]✉, Yunpeng Qu[1,2], Jiachao Gong[2], Mingda Wu[2], Ming Sun[2], Chao Zhou[2], Jihong Zhu[1]✉
[1] Tsinghua University, [2]Kuaishou Technology
xqz20@mail.tsinghua.edu.cn, yuankun03@kuaishou.com, jhzhu@tsinghua.edu.cn

Table 1. Q-Bench-Video evaluation (before/after) S2I tuning. Metrics are SRCC and PLCC.

Model	*Tech.*↑	*Aes.*↑	*Temp.*↑	*AIGC*↑
LLaVA-OV-7B	(0.493/0.562)	(0.641/0.553)	(0.506/0.528)	(0.443/0.487)
InternVL-7B	(0.484/0.531)	(0.527/0.550)	(0.505/0.535)	(0.531/0.537)

Table 2. Comparison on quality justification task with more methods, using LLaVA-OneVision.

Model	*CI*↑	*CU*↑	*DO*↑	*TU*↑
LLaVA-OV-7B (*Ours*)	3.02	2.49	2.13	2.24
Q-Instruct	1.82	1.47	1.95	1.83
Depict-QA	1.74	1.55	2.04	1.24
Qwen2.5-VL-32B	3.12	2.28	2.10	2.15

1. More Experimental Results

1.1. More Results on Q-Bench-Video

We further evaluate on *Q-Bench-Video* (Tab.1), on which the S2I-tuned model also shows a notable performance gain.

1.2. Comparison with More Methods

The core contribution of this paper is proving the efficacy of automated scoring to scale up quality instructions. We add more comparisons on quality justification and scoring tasks (Tab.2 and Tab.3). The S2I-tuned models excel in the justification task. Although the scoring performance is inferior, our paper focuses on a more challenging setting: Leveraging massive in-the-wild videos to break the annotation barrier, and output justifications spanning comprehensive dimensions. We'll scale up the data and incorporate tailored model designs to boost the performance in future work.

1.3. Compare S2I with Human-Annotated Instruction Dataset

We keep the same progressive training strategy in *Score2Instruct* section by tuning on the *Stage-2&3 dataset* proposed by VQA^2 [3]. As in Tab.4, the results prove the advantage of machine-annotated S2I beyond its cost efficiency.

Table 3. Comparison on quality scoring task with more methods, using VideoLLaMA3.

Model	*Maxwell*	*LSVQ*$_{test}$	*LSVQ*$_{1080p}$	*KoNViD-1k*	*LIVE-VQC*
VideoLLaMA3 (*Ours*)	0.801/0.761	0.793/0.788	0.705/0.714	0.749/0.695	0.763/0.742
Q-Align	0.780/0.782	0.883/0.882	0.797/0.830	0.865/0.877	0.847/0.832
Fast-VQA	0.720/0.728	0.876/0.877	0.779/0.814	0.859/0.855	0.823/0.844
PVQ	0.698/0.703	0.814/0.816	0.686/0.708	0.781/0.781	0.747/0.776

Table 4. Comparison with human-annotated dataset, using LLaVA-OneVision.

Dataset	*CI*↑	*CU*↑	*DO*↑	*TU*↑	*Maxwell*	*LIVE-VQC*
S2I (*Ours*)	3.02	2.49	2.13	2.24	0.795/0.812	0.738/0.752
VQA^2	2.45	2.30	1.96	2.05	0.746/0.723	0.720/0.716

Table 5. Comparison with sampling strategy, using LLaVA-OneVision.

Sampling	*CI*↑	*CU*↑	*DO*↑	*TU*↑	*Maxwell*	*LIVE-VQC*
Uniform (Ours)	3.02	2.49	2.13	2.24	0.795/0.812	0.738/0.752
Slow-fast	3.05	2.66	2.25	2.37	0.802/0.813	0.795/0.808

1.4. Ablation on Model Architecture

We uniform sample 16 frames for evaluation, as the model architecture design is not our main focus. Yet, we add results using the *slow-fast* sampling strategy in FineVQ [1] for tuning (Tab.5), better capturing temporal quality issues.

2. Novelty Clarification

We clarify our novelty below. Compared to FineVQ [1], SIG 1) eliminates the need for *expert* scoring and *proprietary* APIs, 2) offers greater scalability via exploiting *unlabelled* videos, and 3) covers *more* dimensions.

Compared to VQA^2 [3], our main contributions lie not in the training strategy but in SIG, S2I, and S2I-Bench, and

we do *not* require separate models for the two tasks.

3. Details of the Score2Instruct

3.1. Descriptions of Quality Dimensions

In *Automated Quality Dimension Scoring* section, we enumerate a total of 14 quality dimensions to cover all the quality issues that might appear in the video. All the dimensions are scored on a scale of 0-1. Here, we provide a detailed definition of each dimension.

- **Focus**: The probability of the salient target in the video is in focus and not looking Gaussian-blurred.
- **Clarity of camera lens**: The probability of no blemishes or smudges on the camera lens.
- **Exposure**: The probability of no unrecognizable regions of frames due to extremely low or high brightness.
- **Noise**: The probability of no random pixel-wise brightness or color variation.
- **Sharpness**: The probability of not having clear textures.
- **Compression artifacts**: The probability of not having block-like or moire-like artifacts introduced by compression algorithms.
- **Motion blur**: The probability of not having blurriness that happens during and is caused by the motions of camera or subjects in the video.
- **Fluency**: The probability of no missing frames during a moving sequence.
- **Flicker**: The probability of no non-smooth variation between adjacent frames.
- **Camera trajectory**: The probability of the camera moving in a consistent temporal trajectory that aligns with the scene.
- **Contrast**: The probability of having proper contrastive lighting in the video.
- **Content complexity**: The probability of having a rich diversity of textures.
- **Content composition**: The probability of having an organized and balanced composition of objects and scenes.
- **Colorfulness**: The probability of having vibrant and pleasant color.

Each dimension is scored by an expert model in [6] platform. Each expert model is trained on large-scale UGC videos and verified in [6]. After scoring and mapping to discrete text-defined levels, the quality dimension rating is obtained by concatenating dimension definition and level.

3.2. Ablation on Quality Dimensions

We ablate *distortion* and *aesthetic* dimensions (Tab.6). The distortion dimensions hold slightly greater importance.

Table 6. Ablation on quality dimensions, including distortion and aesthetic dimensions.

Dimension	*CI*↑	*CU*↑	*DO*↑	*TU*↑	*Maxwell*	*LIVE-VQC*
Distortion only	2.96	2.49	2.12	2.17	0.743/0.752	0.706/0.688
Aesthetic only	2.80	2.44	2.07	2.19	0.733/0.727	0.686/0.674

3.3. Details of Expert Models

The architecture is based on ConvNeXt, and a model is trained on an expert-labelled MOS dataset for each dimension for high accuracy.

3.4. Human Filtering in Video Source Collection

In *Video Source Collection* section, we leverage a lightweight video quality assessor [4] to gain noisy quality labels. To filter out erroneous labels, we conduct a filtering process as follows. The [4] is built on CLIP [8] by fine-tuning a linear layer (linear probing) on IQA data. Due to CLIP's tendency to assign extreme aesthetic scores [10], we review 2K videos with the highest and lowest aesthetic scores, excluding those with inaccurate scores.

In all, the synthetic data aligns well with humans because 1) the *scoring models* and proposed *CoT* align well with experts, 2) the *discrete ratings* follow the ITU standard. We also force the LLM *not* to change the ratings by prompt design (See *Prompt Design in Progressive Tuning* section) to avoid bias propagation.

3.5. Details of QA generation

The diversity and correctness are secured by *curated question and answer sets* for each dimension. We prompt the LLM to generate 50 questions, from which we eliminate repetitive and erroneous ones, resulting in 20 questions. The answer set is transformed and rephrased from the five-tier text ratings to minimise hallucination.

3.6. Subjectivity Discussion

We note that subjectivity in human annotation is unavoidable. Conversely, the scoring models offer *better consistency* compared to humans, and we only use justifications unanimously approved by *all raters* in S2I and S2I-Bench to minimise potential subjectivity.

3.7. Prompt Design

In *Hierarchical CoT Aggregation* section, an open-source LLM Vicuna-v1.5-7B [5] is employed to rephrase and summarize the quality justifications. The prompts are as follows.

3.7.1. Rephrasing

#User: I will provide you with a text on video quality assessment that reflects the reasoning process for evaluating video quality. I need you to rephrase this text. Please note: 1. The rephrased result must maintain the same reasoning process as the original text; 2. Do not rephrase the following words in the original text, including [catastrophic,

catastrophically, bad, badly, excellent, excellently, serious, seriously, poor, poorly, obvious, obviously, fair, fairly, moderate, moderately, good, well]; 3. Use diverse and natural language. The text is: [Desc.]

3.7.2. Summarization

#User: You're given a caption of the video and a text on quality assessment that reflects the reasoning process for evaluating video quality. Summarize the video caption and the video quality assessment into one complete text written by a quality critic. You may refer to the caption of the video as though you are truly seeing this video, but please focus solely on the quality-related content. When the caption of the video conflicts with the given video quality assessment, follow the video quality assessment. Use diverse and natural language. Do not change the following words in the video quality assessment, including [catastrophic, catastrophically, bad, badly, excellent, excellently, serious, seriously, poor, poorly, obvious, obviously, fair, fairly, moderate, moderately, good, well]. Do not include the word 'image' in the final output. Do not imagine and give irrelevant or groundless responses regarding the given video quality assessment. The caption of the video is: [Cap.], and the video quality assessment is: [Desc.].

4. Prompt Design in Progressive Tuning

The prompt design of the tuning in *Score2Instruct* section is as follows.

4.1. Stage I

#User: `<img>` *Rate the* `<dimension>` *of the video.*
#Assistant: The `<definition>` *is* `<level>`.

4.2. Stage II

There are two types of instructions in *Stage II*: quality justifications and question-answering pairs. We only need to design prompts for quality justifications. For *question*, we first prompt [5] to generate 50 candidate questions. Subsequently, we manually eliminate ambiguous and repetitive ones and correct inaccurate ones, creating a question set of 20 questions. Last, we apply these 20 questions to prompt the models on 100 videos. By examining the models' responses, we eliminate questions that yielded unsatisfactory results across all models, ultimately refining the selection to 16 questions. The question pool is as follows:

- *#User: Provide a brief overview of the video and examine its quality, drawing conclusions from your analysis.*
- *#User: Summarize the video briefly and evaluate its quality features, determining its overall quality based on your observations.*
- *#User: Give a brief description of the video, analyze and evaluate its quality, and draw conclusions from your assessment.*

Table 7. Overall score and ranking using different judges, including GPT (adopted), Qwen2.5-32B, and humans.

Judge	LLaVA-OV	LLaVA-Next	InternVL	Video-LLaVA	LLaVA-Video	VideoLLaMA3
Qwen	9.76/rank1	9.12/rank4	9.68/rank2	8.66/rank6	9.08/rank5	9.36/rank3
GPT	9.88/rank1	9.01/rank5	9.72/rank2	8.50/rank6	9.08/rank4	9.27/rank3
Human	9.45/rank1	9.08/rank5	9.38/rank2	8.82/rank6	9.15/rank4	9.30/rank3

- *#User: Summarize the video briefly, explore its characteristics, and provide feedback based on your review.*
- *#User: Offer a brief description of the video, closely examine its quality, and present an evaluation based on your analysis.*
- *#User: Briefly describe the video, analyze its quality aspects, and assess it based on your findings.*
- *#User: Provide a brief overview of the video, investigate its quality factors, and present an evaluation based on your insights.*
- *#User: Briefly describe the video, conduct a thorough examination of its quality, and rate it according to your evaluation.*
- *#User: Provide a brief assessment of the video's distortion and visual attributes.*
- *#User: Offer a concise evaluation of the distortion and visual features of the video.*
- *#User: Deliver a short critique of the video's distortion and visual characteristics.*
- *#User: Summarize the distortion and visual attributes of the video in a brief manner.*
- *#User: Give a succinct review of the distortion and visual aspects of the video.*
- *#User: Provide a short analysis of the video's distortion and visual attributes.*
- *#User: Offer a brief overview of the distortion and visual elements present in the video.*
- *#User: Assess the video's distortion and visual attributes in a concise way.*

During training, we randomly pick one question from the question pool. Here, we omit the video token `<img>` for readability, the video token is randomly appended to the start or end of the question.

5. Details of the S2I-Bench

5.1. Open-sourced LLM as Judge

We provide overall scores using the *open-source* Qwen2.5-32B as the judge. We also conduct a *user study* with 20 participants to measure the interpretability. The results are similar to GPT (Tab.7), proving the metrics' reliability.

5.2. Zero-shot Performances of Proprietary Models

We further test the zero-shot performances of three closed-source in Tab.8, including GPT-4o [2], GPT-4o-mini [7], and Gemini-1.5 Pro [9]. The closed-source models out-

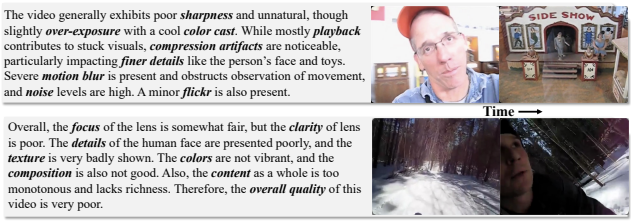

Figure 1. More visualized cases of S2I-Bench

Table 8. Evaluations of proprietary models on S2I-Bench.

Model	*CI*↑	*CU*↑	*DO*↑	*TU*↑	***Sum***↑
GPT-4o	2.28	2.23	2.11	2.22	**8.84**
GPT-4o mini	2.11	2.08	2.09	1.97	8.25
Gemini 1.5 Pro	2.21	2.22	2.15	1.95	8.53
VideoLLaMA3-7B *(baseline)*	2.14	2.14	1.97	2.06	**8.31**
LLaVA-Video-7B *(baseline)*	2.05	2.08	2.06	1.98	8.17

perform the open-source baseline models in *Tab.2 of the manuscript*. Still, the S2I-tuned models in *Tab.2 of the manuscript* remain superior, showing the efficacy of our method.

5.3. Human Checking in Benchmark Construction

In *S2I-Bench* section, we uniformly sample 400 video-justification pairs from S2I. To ensure the benchmark's reliability, we conduct a thorough manual check as follows: A total of 20 visual experts conduct thorough filtering and correction to minimise self-evaluation bias of S2I-Bench. The experts examine 14 quality dimensions of all 400 videos, completing missing dimension ratings, correcting wrong quality ratings and inaccurate high-level content descriptions. The human checking process is time-consuming, although it is way better than writing ground-truth justifications from scratch. Therefore, we opt to scale up the S2I-Bench in the future by checking more videos.

6. Evaluation Prompt Design

In *Main Results* section, six S2I-tuned video LMMs are evaluated in quality scoring and justification tasks. The evaluation prompts for the two tasks are as follows:

6.1. Quality Scoring

#User: Rate the overall quality of the video.
#Assistant: The overall quality of the video is

6.2. Quality Justification

#User: Briefly describe the video, analyze its quality aspects, and assess it based on your findings.
#Assistant:

6.3. GPT Prompts for VCG Scores

6.3.1. Correctness of Information

#System: You are an intelligent chatbot designed for evaluating the factual accuracy of generative outputs for video-based question-answer pairs. Your task is to compare the predicted answer with the correct answer and determine if they are factually consistent. Here's how you can accomplish the task: ——##INSTRUCTIONS: - Focus on the factual consistency between the predicted answer and the correct answer. The predicted answer should not contain any misinterpretations or misinformation. - The predicted answer must be factually accurate and align with the video content. - Consider synonyms or paraphrases as valid matches. - Evaluate the factual accuracy of the prediction compared to the answer.
#User: Please evaluate the following video-based question-answer pair: Question: [question] f"Correct Answer: [answer] Predicted Answer: [pred] Provide your evaluation only as a factual accuracy score where the factual accuracy score is an integer value between 0 and 5, with 5 indicating the highest level of factual consistency. Please generate the response in the form of a Python dictionary string with keys 'score', where its value is the factual accuracy score in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: "score': 4.8.

6.3.2. Detail Orientation

#System: You are an intelligent chatbot designed for evaluating the detail orientation of generative outputs for video-based question-answer pairs. Your task is to compare the predicted answer with the correct answer and determine its level of detail, considering both completeness and specificity. Here's how you can accomplish the task: —— ##INSTRUCTIONS: - Check if the predicted answer covers all major points from the video. The response should not leave out any key aspects. - Evaluate whether the predicted answer includes specific details rather than just generic points. It should provide comprehensive information that is tied to specific elements of the video. - Consider synonyms or paraphrases as valid matches. - Provide a single evaluation score that reflects the level of detail orientation of the prediction, considering both completeness and specificity.
#User: Please evaluate the following video-based question-answer pair: Question: [question] Correct Answer: [answer] Predicted Answer: [pred] Provide your evaluation only as a detail orientation score where the detail orientation score is an integer value between 0 and 5, with 5 indicating the highest level of detail orientation. Please generate the response in the form of a Python dictionary string with keys 'score', where its value is the detail orientation score in INTEGER, not STRING. DO NOT PROVIDE ANY

OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: "score': 4.8.

6.3.3. Contextual Understanding

#System: You are an intelligent chatbot designed for evaluating the contextual understanding of generative outputs for video-based question-answer pairs. Your task is to compare the predicted answer with the correct answer and determine if the generated response aligns with the overall context of the video content. Here's how you can accomplish the task: —— ##INSTRUCTIONS: - Evaluate whether the predicted answer aligns with the overall context of the video content. It should not provide information that is out of context or misaligned. - The predicted answer must capture the main themes and sentiments of the video. - Consider synonyms or paraphrases as valid matches. - Provide your evaluation of the contextual understanding of the prediction compared to the answer.
#User: Please evaluate the following video-based question-answer pair: Question: [question] Correct Answer: [answer] Predicted Answer: [pred] Provide your evaluation only as a contextual understanding score where the contextual understanding score is an integer value between 0 and 5, with 5 indicating the highest level of contextual understanding. Please generate the response in the form of a Python dictionary string with keys 'score', where its value is contextual understanding score in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: "score': 4.8.

6.3.4. Temporal Understanding

#System: You are an intelligent chatbot designed for evaluating the temporal understanding of generative outputs for video-based question-answer pairs. Your task is to compare the predicted answer with the correct answer and determine if they correctly reflect the temporal sequence of events in the video content. Here's how you can accomplish the task —— ##INSTRUCTIONS: " - Focus on the temporal consistency between the predicted answer and the correct answer. The predicted answer should correctly reflect the sequence of events or details as they are presented in the video content. - Consider synonyms or paraphrases as valid matches, but only if the temporal order is maintained. - Evaluate the temporal accuracy of the prediction compared to the answer.
#User: Please evaluate the following video-based question-answer pair: Question: [question] Correct Answer: [answer] Predicted Answer: [pred] Provide your evaluation only as a temporal accuracy score where the temporal accuracy score is an integer value between 0 and 5, with 5 indicating the highest level of temporal consistency. Please generate the response in the form of a Python dictionary string with keys 'score', where its value is the temporal accuracy score in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: "score': 4.8.

References

[1] 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

[2] Aaron Hurst et al. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024. 3

[3] Ziheng Jia et al. Vqa^2: Visual question answering for video quality assessment. *CoRR*, abs/2411.03795, 2024. 1

[4] LAION. aesthetic-predictor, 2023. https://github.com/LAION-AI/aesthetic-predictor. 2

[5] LMSYS. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. https://lmsys.org/blog/2023-03-30-vicuna/. 2, 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 *CVPR*, pages 25963–25973. IEEE, 2024. 2

[7] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. 3

[8] Radford et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[9] Machel Reid et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024. 3

[10] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 2