Towards Explainable In-the-Wild Video Quality Assessment: A Database and a Language-Prompted Approach

Haoning Wu*
Erli Zhang*
[haoning001,ezhang005]@e.ntu.edu.sg
Nanyang Technological University

Liang Liao
Chaofeng Chen
[liang.liao,chaofeng.chen]@ntu.edu.sg
Nanyang Technological University

Jingwen Hou
Annan Wang
[jingwen003,c190190]@e.ntu.edu.sg
Nanyang Technological University

Wenxiu Sun
sunwx@tetras.ai
SenseTime Group Limited

Qiong Yan
yanqiong@tetras.ai
Sensetime Research

Weisi Lin
wslin@e.ntu.edu.sg
Nanyang Technological University

ABSTRACT

The proliferation of in-the-wild videos has greatly expanded the Video Quality Assessment (VQA) problem. Unlike early definitions that usually focus on limited distortion types, VQA on in-the-wild videos is especially challenging as it could be affected by complicated factors, including various distortions and diverse contents. Though subjective studies have collected overall quality scores for these videos, how the abstract quality scores relate with specific factors is still obscure, hindering VQA methods from more concrete quality evaluations (*e.g. sharpness of a video*). To solve this problem, we collect over two million opinions on 4,543 in-the-wild videos on 13 dimensions of quality-related factors, including in-capture authentic distortions (*e.g. motion blur, noise, flicker*), errors introduced by compression and transmission, and higher-level experiences on semantic contents and aesthetic issues (e.g. *composition, camera trajectory*), to establish the multi-dimensional **Maxwell** database. Specifically, we ask the subjects to label among a positive, a negative, and a neutral choice for each dimension. These explanation-level opinions allow us to measure the relationships between specific quality factors and abstract subjective quality ratings, and to benchmark different categories of VQA algorithms on each dimension, so as to more comprehensively analyze their strengths and weaknesses. Furthermore, we propose the **MaxVQA**, a language-prompted VQA approach that modifies vision-language foundation model CLIP to better capture important quality issues as observed in our analyses. The MaxVQA can jointly evaluate various specific quality factors and final quality scores with state-of-the-art accuracy on all dimensions, and superb generalization ability on existing datasets. Code and data available at https://github.com/VQAssessment/MaxVQA.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**.

*This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). Both authors contribute equally.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3611737

KEYWORDS

Dataset, Video Quality Assessment, Explainable, Vision-Language

ACM Reference Format:
Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023. Towards Explainable In-the-Wild Video Quality Assessment: A Database and a Language-Prompted Approach. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611737

1 INTRODUCTION

Rapid advances in technology have democratized the video production process, allowing more ordinary users to create and upload videos with smartphone cameras and public online platforms. Henceforth, Video Quality Assessment (VQA) on these prevalent in-the-wild videos is increasingly important for automatically recommending high-quality videos and improving low-quality ones.

Unlike traditional VQA [40, 48] that focuses on specific types of distortions from pristine reference videos, VQA on in-the-wild videos is challenging as it is affected by authentic and commonly intermixed distortions; moreover, the contents of in-the-wild videos diverse and may also affect their quality. Although plenty of subjective studies [12, 36, 41, 64] have collected enormous quality scores on in-the-wild videos, how specific quality-related factors (*e.g. sharpness, exposure, fluency, noise*) affect or relate with the quality of a given video is still obscure. Consequently, existing objective VQA approaches [22, 28, 45, 56] learnt from these scores are not able to evaluate these explainable concrete factors, thereby limiting their ability to provide targeted suggestions for video restoration or recommendation. Furthermore, the reliance on overall quality scores as the sole benchmark metric of VQA approaches makes it difficult to comprehensively analyze their ability on identifying specific quality issues, so as to apply them to appropriate scenarios.

To solve these challenges, we conduct a large-scale comprehensive subjective study to collect human opinions on specific quality factors, and explore how they affect the abstract quality scores. Specifically, we collect over two million annotations on 13 dimensions, each associated with a commonly-observed quality-related factor. These factors include in-capture authentic distortions [10, 36] (*e.g. blur, noise, poor exposure*), compression/transmission distortions [40, 48], and higher-level semantic-related (aesthetic) issues [28, 59] (*e.g. contents, color, composition*), as illustrated in Fig. 1. To better collect subjective opinions on

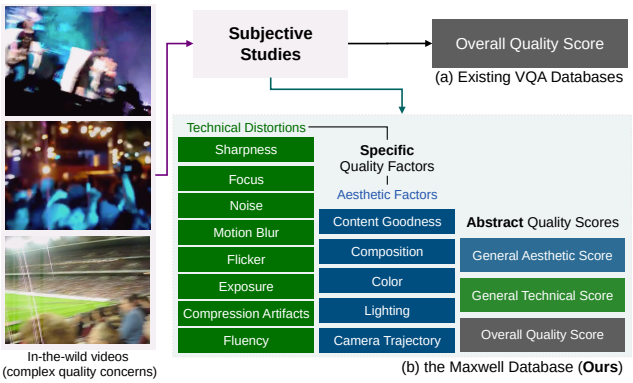

Figure 1: Quality for in-the-wild videos is influenced by complicated factors. For the first time, the proposed Maxwell database studies these *specific* quality-related factors, and explore their impacts on the *abstract* overall quality scores.

these dimensions, our study has several specific designs: *First,* we notice that different factors can simultaneously affect the quality (such as Fig. 1, which is fuzzy and with strong artifacts). To capture all factors that impact the quality of a given video, we prompt subjects to label all factors [8] for each video instead of choosing one major influencing factor [13, 69]. *Second,* in non-reference settings [31, 73], each specific factor (*take sharpness as an example*) can pose either positive (*while sharp*) or negative (*while fuzzy*) impact to video quality, while it may also have no significant effects for some videos. Thus, we define a forcefully positive-neutral-negative ternary choice for each specific factor, which ensures that subjects neither miss a factor for each video, nor face the dilemma of having to choose between good or bad when there are no significant effects. Moreover, to ensure the reliability of labels, we trained all participating subjects for each dimension before the annotation process and each axis in each video is annotated by 35 subjects following [1]. Correspondingly, we construct the Multi-Axis VQA Database with Explanation-level Labels (**Maxwell**). The multi-dimensional opinions in the Maxwell database effectively bridge the gap between quality scores and specific quality factors, allowing for automatic and targeted repairment or enhancement on in-the-wild videos.

We further conduct an extensive analysis on the subjective opinions collected in the Maxwell database. Firstly, we find that the quality of in-the-wild videos is commonly affected by multiple factors, with an average of six factors affecting the quality of each video, which justifies our decision to label all factors. Secondly, we receive slightly more positive opinions than negative opinions, with a ratio of approximately 3:2. This suggests that it is appropriate to consider positive effects besides negative impacts for each quality factor. Thirdly, each factor affects the quality of at least 37% of videos, justifying the validity of the selected specific factors. Furthermore, our study revealed that human visual system is especially sensitive to temporal quality (Flicker and Fluency), emphasizing the importance of effective temporal modeling in VQA. The analysis on Maxwell helps us to better explain subjective VQA on in-the-wild videos, and further develop more effective objective approaches.

Inspired by the analysis, we propose the Multi-Axis Video Quality Assessor (**MaxVQA**), a language-prompted VQA approach that learns from comprehensive opinions in the Maxwell to jointly predict specific quality factors and overall quality, via integrating the vision-language foundation model CLIP [38] with FAST-VQA [56], the recent state-of-the-art VQA model for in-the-wild videos. Specifically, the MaxVQA first fuses CLIP visual features with FAST-VQA features to enhance the perception on low-level textures and temporal variations. It then compares the cosine similarity between the visual features and the textual features of a pair of dimension-specific prompts, so as to predict quality score for each dimension. Unlike existing strategies [8, 13] that requires separate regressors for multi-objective quality evaluation, the MaxVQA unifies multiple dimensions with different text prompts. The proposed MaxVQA proves state-of-the-art accuracy not only on all dimensions in Maxwell, but also on overall quality assessment for existing VQA datasets. MaxVQA trained on Maxwell also shows well generalization ability on existing VQA datasets, suggesting its superb robustness.

Our contributions can be summarized as three-fold:

(1) We construct the **Maxwell** database, a comprehensive subjective study to collect over two million human opinions for 13 distinct specific quality factors on 4,543 in-the-wild videos. Our database allows for objective VQA methods to provide specific quality evaluations (*e.g.* fluency of videos).

(2) Based on Maxwell, we analyze the characteristics of different factors, and their relations with overall quality scores for in-the-wild videos. Moreover, we provide the first multi-dimensional benchmark for objective VQA approaches to analyze their ability on capturing specific quality concerns.

(3) We design the **MaxVQA**, a vision-language-based VQA model that can jointly predict specific quality factors and final quality scores, by enhancing CLIP with low-level-sensitive FAST-VQA features. MaxVQA proves state-of-the-art performance and excellent generalization ability.

2 RELATED WORK

2.1 Subjective Studies for In-the-Wild VQA

Unlike traditional settings [40, 48] that mainly focus on compression [2, 71] or transmission-related distortions, subjective studies for in-the-wild VQA [12, 41, 52, 64] directly collect human quality opinions on real-world videos. Thus, they face more complex quality-related issues [6, 10, 24, 36] and much extended content diversity [27, 53, 62]. Therefore, these quality scores collected on in-the-wild videos can be affected by complicated reasons, and it is difficult to ground the effect of a given specific factor (*e.g. sharpness or fluency of a video*). To investigate the relations between overall quality perception on in-the-wild videos and specific quality-related factors, we construct the **Maxwell** database that collects large-scale opinions for a wide range of specific quality factors together (*distortions, content-related*) on in-the-wild videos, and associated analysis that better explains subjective VQA for in-the-wild videos.

2.2 Objective VQA Methods

Recent years have witnessed rapid progresses of objective VQA methods [4, 5, 22, 26, 28, 29, 42, 45, 56, 62], which can be roughly categorized into two types: *First*, classical methods, represented by TLVQM [22] and VIDEVAL [45], which design handcraft kernels to extract the low-level patterns of various common distortions that happen on in-the-wild videos. *Second*, deep-learning-based methods. Pioneered by VSFA [28], these methods typically extract features

from pre-trained deep neural networks, and regress them with overall quality opinions, such as recent BVQA-*Li* [26]. Most recently, FAST-VQA [56], is proposed to learn quality scores end-to-end from pre-processed videos, and reaches superior accuracy with high efficiency. Although achieving higher accuracy, existing objective methods especially deep methods can hardly reason why their performance boost, nor predict the mechanism behind the quality scores. In recent years, vision-language foundation models [19, 35, 38, 54, 66], such as CLIP [38] and ALIGN [19] can measure similarity between text sentences and images, while several studies [15, 21, 51, 61, 69, 70] have attempted vision-language modeling in VQA. With the constructed **Maxwell**, we integrate CLIP with FAST-VQA, and propose the language-prompted **MaxVQA** that can jointly evaluate multiple specific quality factors and overall quality scores. The MaxVQA can analyze video quality in a more explainable way.

3 THE MAXWELL DATABASE

In this section, we elaborate on the designs and processes of the subjective study for the Maxwell database (Sec. 3.1). The subjective study is conducted **in-lab** with 35 participants [1] annotating for 4,543 videos on 16 dimensions, including 13 specific quality factors and 3 abstract quality ratings, as listed in Table 1. Afterwards, we conduct extensive analyses on the collected subjective opinions, as discussed in Sec. 3.2. The well-designed subjective study in Maxwell as well as the extensive analyses help us to better investigate the effects of different specific factors on VQA for in-the-wild videos.

3.1 Design of the Subjective Study

3.1.1 Choice of Quality-related Factors. In general, we conduct the study from two perspectives, the distortion-related technical perspective, and the semantic-related aesthetic perspective [59]. We collect the general quality opinions from the two perspectives and choose several specific factors from these perspectives, as follows.

1) **Factors in distortion-related technical perspective.** We collect opinions on common specific distortions (*e.g. flicker, compression artifacts*) that happen in real-world videos. In general, the dimensions related to these factors have clear standards, that stronger distortion relates to worse quality [13, 65]. Specifically, based on the technical origin of the distortions, they can be further grouped into *in-capture authentic distortions* [10, 36], and *post-capture distortions from compression or transmission* [2, 50, 71]. Specifically, we study six common [8, 10] in-capture authentic distortions[1]:

(T-1) **Low Sharpness**: The video does not have clear textures.
(T-2) **Out of Focus**: The salient target in video (*e.g. human in a portrait video*) is not in-focus and looks Gaussian-blurred.
(T-3) **Noise**: Random pixel-wise brightness or color variation.
(T-4) **Motion Blur**: Blurriness that happens during and is caused by motions of camera or subjects in the video.
[T-5] **Flicker**: Non-smooth variation between adjacent frames.
(T-6) **Poor Exposure**: Unrecognizable regions of frames due to extremely low (high) brightness.

and two common errors induced by compression or transmission:

(T-7) **Compression artifacts**: Block-like or moire-like artifacts introduced by compression algorithms [49, 55].

[1]To distinguish, we denote all *temporal*-related factors with brackets, *e.g. [T-5]* Flicker. Other spatial factors (*i.e. can be detected within a frame*) are in parenthesis.

Table 1: The codes and respective positive/negative descriptions during subjective study for different dimensions in the Maxwell database.

Code	Dimension Names	*Positive Description*	*Negative Description*
T-1	Sharpness	*Sharp*	*Fuzzy*
T-2	Focus	*In-Focus*	*Out-of-Focus*
T-3	Noise	*Noiseless*	*Noisy*
T-4	Motion Blur	*Clear-Motion*	*Blurry-Motion*
[T-5]	Flicker	*Stable*	*Shaky*
T-6	Exposure	*Well-exposed*	*Poorly-exposed*
T-7	Compression Artifacts	*Original*	*Compressed*
[T-8]	Fluency	*Fluent*	*Choppy*
*T-all	*Technical Perspective*	*Not Degraded	*Severely Degraded
A-1	Contents	*Good*	*Bad*
A-2	Composition	*Organized*	*Chaotic*
A-3	Color	*Vibrant*	*Faded*
A-4	Lighting	*Contrastive*	*Gloomy*
[A-5]	(Camera) Trajectory	*Consistent*	*Incoherent*
*A-all	*Aesthetic Perspective*	*Good Aesthetics	*Bad Aesthetics
*O	*Overall Quality score*	*High Quality	*Low Quality

[T-8] **Low Fluency**: Missing frames during a moving sequence.

With specific distortions identified, we can employ existing respective algorithms to restore the video. For instance, if we detect low sharpness in a video, we can apply super-resolution [3] to repair it; denoising [43], deblurring [37] and stablization [25] can restore the video when noises, motion blurs or flickers are detected; frame interpolation [17] can alleviate low fluency caused due to missing frames. These factor-level specific quality evaluations can assist automated video quality enhancement systems in the future.

2) **Factors in semantic-related aesthetic perspective.** Many studies have suggested that visual quality is affected by semantic-related factors beyond technical distortions [13, 28, 59, 68]. Unlike technical distortions, these higher-level factors are relatively under-studied in existing VQA researches. To better choose these semantic-related factors, we ask the subjects to score their feeling on videos based on four dimensions commonly concerned by existing image aesthetic assessment (IAA) studies [14, 16, 34, 63, 67], with an additional dimension for *[temporal]* aesthetics on videos:

(A-1) **Contents**: Are the contents in the video appealing?
(A-2) **Composition**: Do the video has organized and balanced composition of objects and scenes?
(A-3) **Color**: Does the video has vibrant, pleasant color?
(A-4) **Lighting**: Does the video has contrastive lighting?
[A-5] **Trajectory**: Does the camera moves in a consistent [temporal] trajectory that aligns with the scene?

3.1.2 Design of the Study. After introducing the factors to study, we discuss the concrete form for the subjective study as follows.

1) **Evaluate the Impact of Each Factor.** Many existing studies [22, 56] have suggested that different quality issues can concurrently exist and affect the quality of an in-the-wild video. Therefore, to conduct the study more comprehensively, instead of classifying one main factor that affect quality most significantly [13, 69], we ask subjects to evaluate the impact of each factor on video quality.

2) ***Good/neutral/bad*: a ternary choice.** Though each factor will affect quality of certain videos, it is unlikely for all 13 factors to jointly significantly impact the quality of any one video. Therefore, we allow for a neutral choice for each dimension denoting that the corresponding factor does not notably impact perceptual quality of the video. Moreover, considering that each specific factor can pose either positive or negative impact to video quality, we design

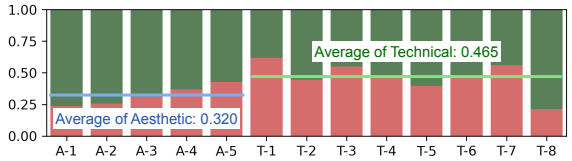

Figure 2: The tendency between negative (red) and positive (green) opinions for each dimension. The overall average of positive-to-negative ratio is 1.44:1.

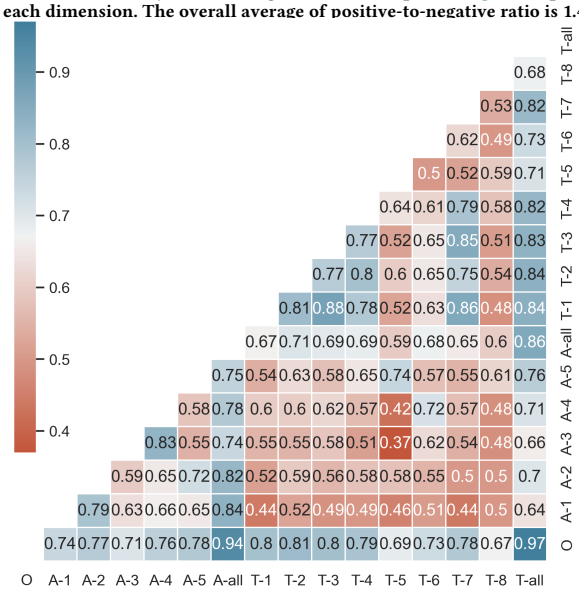

Figure 3: The correlation (*Pearson Linear, PLCC*) map among various quality factors and overall quality scores. See Table 1 for full names for codes.

a ternary choice question for each dimension, including the neutral choice and a pair of antonyms to describe the positive and negative choices. The annotation form in our study is exemplified as follows:

```
(T-1) Sharpness:      Fuzzy < ----neutral---- > Sharp
(Please choose)       [ ]          [ ]          [ ]
```

Descriptions for positive and negative choices are listed in Table 1.

3.1.3 Collection of Videos. To ensure the diversity of annotated videos, we collect them from large video databases [20, 44] and finally sub-sample 4,543 videos [47] that aligns with the distribution of the candidate set for annotation. The resolution of these videos ranges from 240P to 1080P, with an average duration 9s.

3.1.4 Training, Testing, and Annotation. To ensure that the subjects have clear and common understanding on all factor dimensions, we conduct the subjective studies **in-lab**. Moreover, before annotation, we collect three examples each for the positive (*e.g. stable for Flicker axis*), negative (*e.g. shaky for Flicker axis*), and neutral cases for all dimensions to **train** the participants. Moreover, similar as existing efforts [12, 64], we also derive a testing process and reject the subjects that do not pass the testing. After rejection, every video has annotations from at least 31 accepted subjects. Denote the negative, neutral and positive opinions as $[-1, 0, 1]$, the mean factor opinion score ($\mathrm{MOS}_{a,i}$) for factor a of video i are obtained by averaging the raw opinions $\mathrm{OS}^k_{a,i}|^K_{k=0}$ from K accepted subjects.

3.2 Analyses on the Subjective Opinions

3.2.1 Proportions of Different Opinions. First, we examine the proportion of non-neutral opinions in each dimension, which ranges from [0.37,0.56] with an average of 0.45. This means that the quality of an average video is affected by on-average **5.83** factors among the 13 factors, where each factor impacts the quality of **at least 37%** of videos, proving the rationality of selecting these factors and labeling impact of each one. Moreover, we measure the tendency between positive and negative opinions through relative proportions after removing the neutral opinions, as shown in Fig. 2. The tendency for is balanced for technical distortions but significantly biased towards positive opinions for aesthetic factors. This result supports our design of the positive choices and suggests that human visual system may not only perceive *quality* in a negative way.

3.2.2 Correlation among Dimensions. Next, we visualize the correlation map of opinions among all dimensions (factors and abstract quality scores) in Fig. 3, from which we notice several interesting observations. ***First***, all factors are highly relevant to, but also notably different from overall quality (**O**), with PLCC ranging in [0.67,0.81]. ***Second***, some distortions tend to happen together in real-world videos. For example, Sharpness (**T-1**) most strongly associates with Noise (**T-3**), which might be because low-definition video capturing devices may also be worse on dealing with noises. ***Third***, we observe a general low relevance between effects of aesthetic and technical factors (*average PLCC about 0.5*). For example, the correlation between Color (**A-3**) and Flicker (**T-5**)) is very low (0.37). This observation suggests that they usually have different influence to overall quality assessment on in-the-wild videos, supporting our design to divide these factors into two different perspectives during the subjective study. ***Fourth***, the inter-relation between spatial and temporal distortions is also low, suggesting the importance of specific temporal modeling in VQA. These observations provide valuable guidance on improving future objective VQA models.

3.2.3 Absolute Responses of Factors. In addition to correlation maps, we would also like to discover human sensitivity on different specific factors. Thus, we visualize the absolute mean responses (AMR_a) in Fig. 5 as the database-wise average of **absolute** MOS_a for each factor a in Maxwell, formulated as $\mathrm{AMR}_a = \frac{\sum_{i=1}^{N} |\mathrm{MOS}_{a,i}|}{N}$, where $\mathrm{MOS}_{a,i}$ is the mean factor opinion score of axis a for video i, and $N = 4,543$ is the number of videos in Maxwell. Furthermore, we visualize the proportion of non-neutral opinions in each dimension as absolute **raw** responses (ARR_a) (gray bars in Fig. 5). From both responses, we observe the especial sensitivity on the two temporal distortions (Flicker ([**T-5**]) and Fluency ([**T-8**]), demonstrating that temporal modeling is important for real-world VQA. Moreover, Content (**A-1**) and Sharpness (**T-1**) are ranked next under both response metrics, suggesting that they are also important quality concerns to be considered for in-the-wild videos.

3.2.4 Qualitative Studies on Specific Factors. In Fig. 4, we further illustrate extreme examples in each dimension, to better qualitatively understand the quality concerns of different dimensions. From these examples, we validate that technical distortions are usually intermixed in real-world videos. Moreover, the semantic-related aesthetic dimensions are also highly associated with overall quality perception, such as Composition, Color, and Lighting (where the *good* and *bad* cases are with obvious higher and lower quality). In summary, different dimensions represent diverged quality concerns

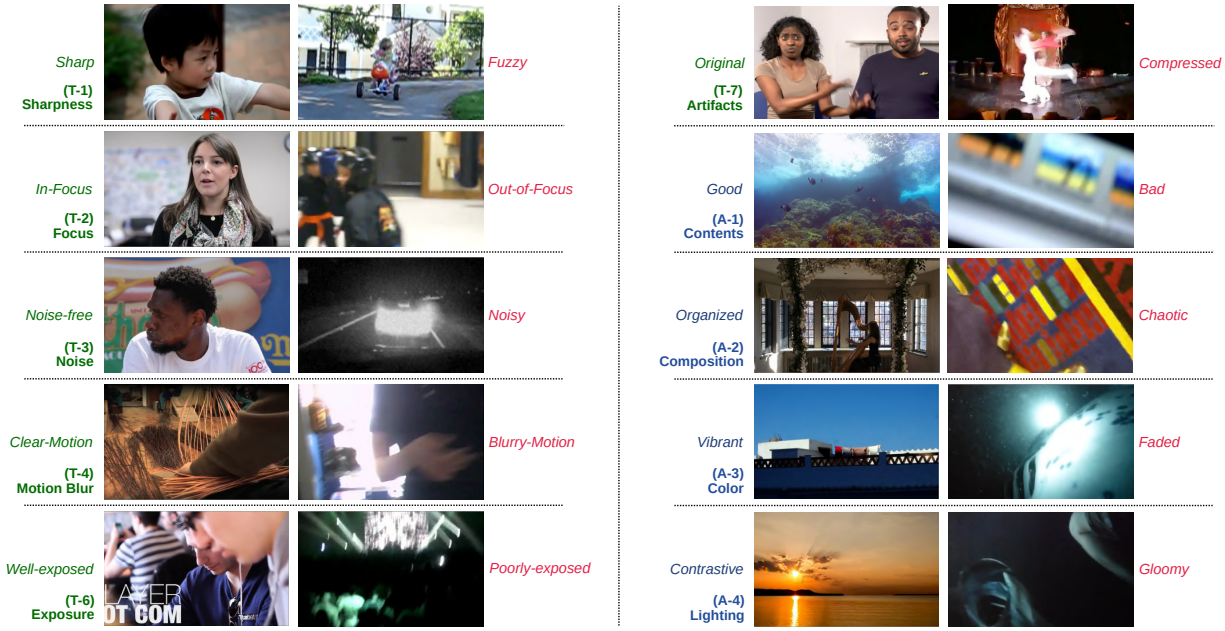

Figure 4: Qualitative studies on different specific factors, with a good video (>0.6) and a bad video (<-0.6) in each dimension of Maxwell; [A-5] Trajectory, [T-5] Flicker, and [T-8] Fluency are focusing on temporal variations and example videos for them are appended in supplementary package. Zoom in for details.

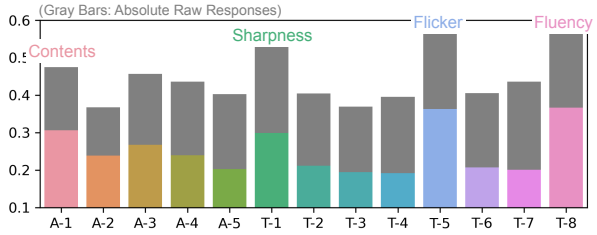

Figure 5: The absolute mean responses (AMR, colored) and absolute raw responses (ARR, gray) on diverse quality factors, where higher absolute response means subjects are more sensitive to the respective quality concerns.

which align with conventional definitions about them, providing reliable supervisions for objective specific quality evaluation.

4 OUR APPROACH: THE MAXVQA

4.1 Overview: Language-Prompted VQA

4.1.1 Unifying Dimensions via Language Prompts. With large-scale multi-dimensional human opinions collected in the Maxwell database, we would like to design an objective approach that can learn from these opinions and then jointly predict specific quality factors and overall quality scores for in-the-wild videos. Unlike traditional multi-task strategies [8, 13] that regress each training objective independently, we propose to unify these inter-related dimensions via a language-prompted paradigm. Specifically, it is based on CLIP (Contrastive Language Image Pre-training) [38], which is composed of a textual encoder (**CLIP-Textual**, E_t) and a visual encoder (**CLIP-Visual**, E_v). Given any text prompt P and video $\mathcal{V} = \{V_t|_{t=0}^{N-1}\}$ as inputs, CLIP can encode them into the same representation space:

$$f_{V,t} = \mathrm{E}_v(V_t)|_{t=0}^{N-1};\ f_P = \mathrm{E}_t(P) \tag{1}$$

Then, we can calculate the similarity between prompt P and V_t:

$$\mathrm{Sim}(P, V_t) = \frac{f_{V_t} \cdot f_P}{\|f_{V_t}\|\|f_P\|} \tag{2}$$

which allows us to unify quality evaluation on different dimensions by setting different text prompts P. Details are discussed in Sec. 4.4.

4.1.2 Modifying CLIP for in-the-wild VQA. Though CLIP can jointly encode videos and various natural language text prompts, it still has several problems in both its visual and textual parts hindering it from modeling VQA more effectively. **1) For the visual part**, due to downsampling and global pooling operations, CLIP-visual has reduced sensitivity to low-level detail-related factors such as *noise, sharpness, artifacts*, which are most correlated with overall quality scores. Moreover, CLIP-visual does not have any temporal modeling, which neglects the temporal distortions as proved important in our analysis. To solve the problems, we propose to utilize the local CLIP visual features by modifying its attention pooling outputs (Sec. 4.2.1); more importantly, we fuse the CLIP-visual features with detail-aware and temporal-aware FAST-VQA features (Sec. 4.2.2). **2) For the textual part,** previous attempts have shown that unlike general prompts (*e.g. good/bad*) that can effectively match overall quality scores, the prompts for specific factors are poorly aligned with human perception (see Table 5). Therefore, we adopt the learnable contextual prompts to optimize the textual inputs (Sec. 4.3). With these designs, we propose the Multi-axis Video Quality Assessor (**MaxVQA**), discussed as follows.

4.2 Low-level Enhanced Visual Backbone

In this section, we introduce the enhanced visual backbone in MaxVQA (Fig. 6(b)), which extracts the local CLIP features after attention pooling, and further fuses them with FAST-VQA features to enhance CLIP-visual on low-level and temporal distortions.

4.2.1 Local Features from CLIP. For the original CLIP, the visual backbone encodes the global embedding f_{V_t} for each frame through the attention pooling layer. Denote the feature before the specific attention pooling layer in CLIP-ResNet-50 backbone as $f_{V_t}^{\text{Pre-Pool}}$,

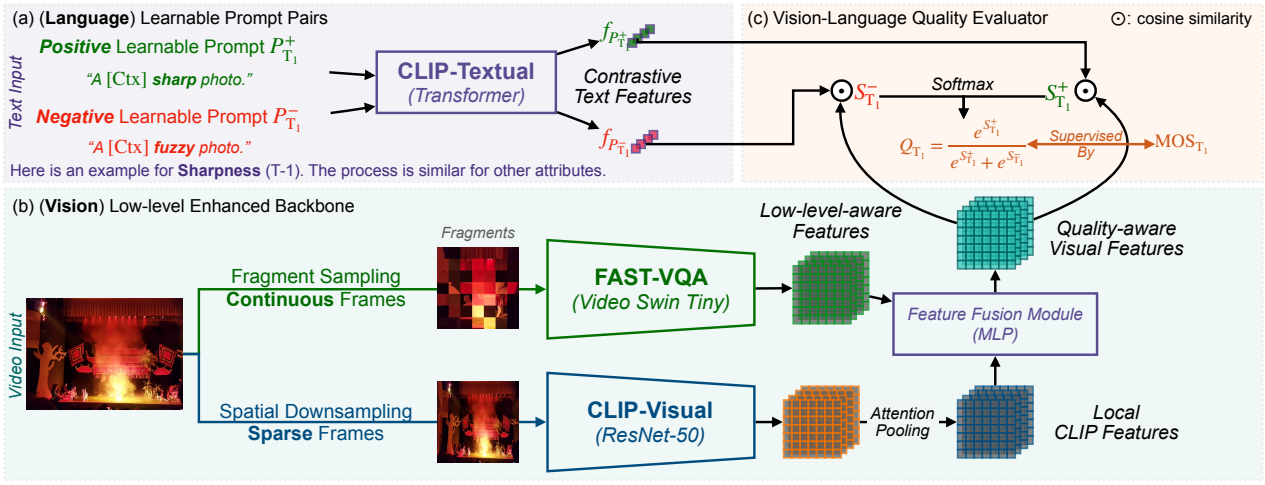

Figure 6: The structure of the proposed Multi-axis Video Quality Assessor (MaxVQA), including (a) Learnable Contrastive Language Prompts to *encode text inputs*, (b) Low-level Enhanced Visual Backbone to *encode videos*, (c) and the final Vision-Language Quality Evaluator to *output multi-axis quality scores*.

this final attention pooling is achieved by a single-layer multi-head self-attention (MHSA) on these features, as follows:

$$[f_{V_t}, f^{\text{Local}}_{V_t}] = \text{MHSA}([\overline{f^{\text{Pre-Pool}}_{V_t}}, f^{\text{Pre-Pool}}_{V_t}]) \tag{3}$$

while only f_{V_t} is used in naive CLIP, we notice that the output local features $f^{\text{Local}}_{V_t}$ could also be useful as most quality factors as studied in the Maxwell can be localized. Moreover, existing study [39] proves that these local features are more sensitive to object-level recognition (*detection, segmentation*), which are also associated with aesthetic-related dimensions [14, 16]. Thus, we take the $f^{\text{Local}}_{V_t}$ instead of f_{V_t} as output features of CLIP-visual.

4.2.2 Fusion with FAST-VQA Features. Many existing works [23, 45, 46, 56] have pointed out that directly using fixed features from high-level pre-trained deep neural networks [7, 11] may have compromised sensitivity on several texture-related factors (*e.g. noises, artifacts and sharpness*) due to downsampling of visual inputs. To enhance these important low-level perceptual factors and compliment the lack of temporal modeling in CLIP, we fuse the CLIP feature with the FAST-VQA [56, 57] features, the state-of-the-art VQA-specific features which proved excellent performance on several VQA datasets and well distinction on temporal distortions. To get these features, the videos are passed through the fragment sampling (*crop multiple **original resolution** patches and splice them together*, see in Fig. 6(b)), and then fed into a modified Video Swin Transformer [32]. Denote the fragment sampling as **F**, the FAST-VQA features $F^{\text{FAST}}_{\mathcal{V}}$ of $\mathcal{V} = \{V_t|^{N-1}_{t=0}\}$ are extracted as follows:

$$F^{\text{FAST}}_{\mathcal{V}} = \text{Swin}([\{\mathbf{F}(V_t)|^{N-1}_{t=0}\}]) \tag{4}$$

where Swin denotes the modified Video Swin Transformer Tiny, that takes temporally-aligned fragments (concatenated as videos) as inputs. *More details for fragment sampling are in supplementary.*

Finally, the FAST-VQA features are fused with local CLIP-Visual features through a residual multi-layer perceptron (MLP):

$$f^{\text{Final}}_{V_t} = \text{MLP}([f^{\text{Local}}_{V_t}, (f^{\text{FAST}}_{\mathcal{V}})_t]) + f^{\text{Local}}_{V_t} \tag{5}$$

where $(f^{\text{FAST}}_{\mathcal{V}})_t$ denotes t-th feature frame of $f^{\text{FAST}}_{\mathcal{V}}$; $t \in [0, N)$.

4.3 Learnable Language Prompts

To distinguish diverse dimensions, we initialize the text prompts differently for each axis with their respective positive and negative descriptions in Maxwell (see Table 1). Denote positive and negative descriptions for a as D^{Pos}_a and D^{Neg}_a, the initial positive $\hat{P}^+_a$ and negative $\hat{P}^-_a$ prompts for a technical factor a (T-x) are defined as:

$$\hat{P}^+_a = D^{Pos}_a + \text{" photo."}; \ \hat{P}^-_a = D^{Neg}_a + \text{" photo."} \tag{6}$$

For the semantic-related aesthetic factors, the prompts are initialized with their dimension name (denoted as Name_a, *e.g.* Contents for A-1, Color for A-3) for more targeted and accurate modeling:

$$\hat{P}^+_a = D^{Pos}_a + \text{Name}_a + \text{" photo."}; \ \hat{P}^-_a = D^{Neg}_a + \text{Name}_a + \text{" photo."} \tag{7}$$

Though some more abstract initial prompt pairs (*e.g. good/bad*) have proved good performance [51, 60, 69] on overall quality perception, both existing studies and our experiments (Table 5, *row 1*) suggest that the more specific prompts are poorly aligned with respective human perception. Therefore, we choose the simple and efficient contextual prompt [72] to optimize these initial prompts, by inserting a single context token before the initial prompts:

$$P = \text{"A "} + \text{Ctx} + \hat{P} \tag{8}$$

where Ctx is the context token, initialized as *"X"* and optimized during training, and $\hat{P}$ is an overall denotion of P^+_a and P^-_a. Tokens except Ctx and weights of the language encoder are keep **frozen**.

4.4 Unified Vision-Language Quality Evaluator

To finally evaluate quality, the proposed MaxVQA calculates the positive similarity (S^+_a) and negative similarity (S^-_a) in dimension a from its specific positive P^+_a and negative P^-_a prompts,

$$S^+_a = \sum_{t=0}^{N-1} \frac{\text{Sim}(P^+_a, f^{\text{Final}}_{V_t})}{N}; \ S^-_a = \sum_{t=0}^{N-1} \frac{\text{Sim}(P^-_a, f^{\text{Final}}_{V_t})}{N} \tag{9}$$

and perform softmax pooling to obtain the respective quality scores:

$$Q_a = \frac{e^{S^+_a}}{e^{S^+_a} + e^{S^-_a}} \tag{10}$$

Table 2: Multi-Axis benchmark comparison of existing approaches and the proposed MaxVQA on Maxwell. [Temporal] dimensions are labeled in brackets.

Dimensions (*in codes*)	A-1	A-2	A-3	A-4	[A-5]	A-all	T-1	T-2	T-3	T-4	[T-5]	T-6	T-7	[T-8]	T-all	O
Methods	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC
Zero-shot Methods: (not fine-tuned on any dimensions)																
(c, spatial) NIQE [33]	0.317	0.281	0.329	0.321	0.211	0.338	0.189	0.255	0.174	0.217	0.136	0.199	0.156	0.178	0.255	0.301
(c, temporal) TPQI [30]	0.246	0.293	0.210	0.225	0.360	0.319	0.223	0.293	0.239	0.374	**0.463**	0.225	0.244	**0.410**	0.363	0.361
(CLIP-based) SAQI [60]	**0.388**	**0.410**	**0.453**	**0.504**	**0.393**	**0.515**	**0.560**	**0.500**	**0.524**	**0.509**	0.344	**0.482**	**0.497**	0.311	**0.554**	**0.559**
Supervised Methods: (for existing approaches, we adopt naive multi-task training on all dimensions)																
(classical) TLVQM[22]	0.477	0.523	0.437	0.471	0.601	0.590	0.537	0.571	0.538	0.606	0.664	0.503	0.539	0.530	0.653	0.652
(classical) VIDEVAL[45]	0.469	0.533	0.501	0.513	0.533	0.564	0.578	0.534	0.548	0.557	0.664	0.467	0.543	0.393	0.595	0.601
(c+d) RAPIQUE[46]	0.490	0.538	0.520	0.559	0.560	0.651	0.610	0.618	0.588	0.621	0.563	0.568	0.566	0.406	0.695	0.708
(deep) VSFA[28]	0.512	0.556	0.611	0.634	0.515	0.624	0.719	0.625	0.642	0.612	0.555	0.645	0.643	0.406	0.672	0.678
(deep) BVQA-*Li*[26]	0.553	0.607	0.659	0.668	0.678	0.671	0.746	0.686	0.694	0.682	0.781	0.653	0.677	0.659	0.759	0.739
(deep) FAST-VQA[56]	0.614	0.630	0.696	0.709	0.646	0.721	0.800	0.724	0.755	0.731	0.751	0.695	0.736	0.654	0.803	0.782
MaxVQA (Ours)	**0.681**	**0.701**	**0.757**	**0.749**	**0.712**	**0.775**	**0.825**	**0.748**	**0.776**	**0.761**	**0.782**	**0.748**	**0.763**	**0.684**	**0.827**	**0.813**

Table 3: Evaluation of MaxVQA on existing in-the-wild VQA datasets. All experiments are conducted under 10 train-test splits with mean results reported.

Dataset	LIVE-VQC		KoNViD-1k		YouTube-UGC	
Methods	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
TLVQM[22]	0.799	0.803	0.773	0.768	0.669	0.659
VIDEVAL[45]	0.752	0.751	0.783	0.780	0.779	0.773
RAPIQUE[46]	0.755	0.786	0.803	0.817	0.759	0.768
CNN+TLVQM[23]	0.825	0.834	0.816	0.818	0.809	0.802
VSFA[28]	0.773	0.795	0.773	0.775	0.724	0.743
PVQ[64]	0.827	0.837	0.791	0.786	NA	NA
CoINVQ[53]	NA	NA	0.767	0.764	0.816	0.802
BVQA-*Li*[26]	0.834	0.842	0.834	0.836	0.818	0.826
DisCoVQA[58]	0.820	0.826	0.846	0.849	0.809	0.808
FAST-VQA[56]	0.849	0.862	0.891	0.892	0.855	0.852
MaxVQA (Ours)	**0.854**	**0.873**	**0.894**	**0.895**	**0.894**	**0.890**

5 EXPERIMENTAL ANALYSIS

In this section, we benchmark existing VQA approaches on the Maxwell database (Sec. 5.2), and evaluate the performance and generalization ability of the proposed MaxVQA (Sec. 5.2&5.3). We also conduct ablation studies (Sec. 5.4) and qualitative studies (Sec. 5.5) to further analyze the effectiveness of the proposed MaxVQA.

5.1 Implementation Details

5.1.1 Experimental Setups. We evaluate the proposed MaxVQA with **frozen** visual and textual encoders, and pre-extract the visual features to reduce computational cost. The only optimizable parameters are the MLP module for visual feature fusion and the contextual prompt Ctx, therefore the training requires only 3GB graphic memory cost at batch size 16. Following original CLIP, the videos are downsampled to 224×224 before fed to E_v. Our code is based on OpenCLIP [18]. The CLIP-visual backbone is ResNet-50.

5.1.2 Database Settings. Following recent practices [2, 64], we split the Maxwell database with two parts, the open **training** set with 3,634 videos, and the reserved **test** set with 909 videos (*We will maintain a test server for future methods to evaluate on the test set*). Moreover, we evaluate a single-dimension variant for MaxVQA on three common in-the-wild VQA datasets with only overall quality scores available: **KoNViD-1k** (1200 videos), **LIVE-VQC** (585 videos), **YouTube-UGC** (1147 videos). Results reported for existing databases are the mean results of 10 random train-test splits.

5.1.3 Baseline Methods for Benchmark Study. To better increase the diverse of our benchmark study, we choose representative state-of-the-art VQA methods with different characteristics:

- TLVQM (2019): Representative handcraft *classical* method.
- VIDEVAL (2021): Representative handcraft *classical* method.
- RAPIQUE (2021): Representative method that combines *deep* CNN features with handcraft *classical* features.
- VSFA (2019): The first *deep* CNN-feature-based method.
- BVQA-*Li* (2022): An improved *deep* CNN-feature-based method, with an additional **temporal** 3D-CNN backbone.
- FAST-VQA (2022): The first ***end-to-end*** *deep* VQA method.

and several representative zero-shot VQA methods to explore between these objective metrics and the dimensions in the Maxwell:

- NIQE (2013): NSS-based **spatial** zero-shot quality evaluator.
- TPQI (2022): State-of-the-art zero-shot **temporal** VQA method.
- SAQI (2023): Zero-shot CLIP-based VQA method with only **abstract** (*i.e. high/low quality; good/bad*) prompt design.

We evaluate all baseline methods with official implementations.

5.2 Benchmarking on the Maxwell

5.2.1 Zero-shot Approaches. We first benchmark representative zero-shot quality indices on Maxwell. Specifically, we do not fit the scores with any dimensions, but evaluate how these indices match specific dimensions. The CLIP-based SAQI shows far better performance than NSS-based NIQE, proving the potential of CLIP on VQA. Still, without temporal modeling, it shows notably lower (-25%) accuracy on two temporal distortions: Flicker (**[T-5]**) and Fluency (**[T-8]**) than the specific temporal VQA index, TPQI. This suggests that we need to include temporal modeling ability for a CLIP-based approach to better solve in-the-wild VQA problem.

5.2.2 Existing Supervised Methods, and MaxVQA. We also benchmark existing supervised methods with a multi-task training strategy. Classical methods even struggle to surpass zero-shot approaches on certain dimensions. For deep methods, while FAST-VQA reaches champions on overall performance and most dimensions, BVQA-*Li* is more competitive on temporal-related dimensions (**A-5, T-5, T-8**). The Maxwell benchmark can then suggest that an independent temporal backbone may improve temporal modeling in VQA, besides only concluding that FAST-VQA is more effective. For the proposed **MaxVQA**, with CLIP-visual features and language-prompted modeling, it can further notably outperform FAST-VQA especially on semantic-related dimensions (**A-x**, +9.4% in-average). The results suggest that accurately assessing multiple quality factors is a more challenging task than only predicting overall quality scores.

Table 4: Prediction of which dimension is closer to subjective quality scores in existing databases? Top-5 dimensions are highlighted with (ranks) in parenthesis.

Dimension	A-1	A-2	A-3	A-4	[A-5]	A-all	T-1	T-2	T-3	T-4	[T-5]	T-6	T-7	[T-8]	T-all	O
Existing Database	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC	PLCC
LIVE-VQC	0.681	0.708	0.733	0.747	0.777	0.772	0.767	0.821 (3)	0.767	0.821 (3)	0.700	0.765	0.763	0.816 (5)	**0.830** (1)	0.822 (2)
KoNViD-1k	0.751	0.776	0.759	0.778	0.772	0.825 (4)	0.819 (5)	0.859 (2)	0.813	0.824	0.645	0.768	0.800	0.737	0.859 (2)	**0.865** (1)
YouTube-UGC	0.681	0.713	0.728	0.756	0.671	0.794	0.819 (4)	0.819 (4)	0.821 (2)	0.793	0.553	0.714	0.814	0.746	0.821 (2)	**0.830** (1)

Table 5: Ablation Studies: Effects of Low-level-enhanced Visual Backbone and Contextual Prompts on Maxwell. The metric is PLCC (same as other tables).

Variants / Dimension	A-1	A-2	A-3	A-4	[A-5]	A-all	T-1	T-2	T-3	T-4	[T-5]	T-6	T-7	[T-8]	T-all	O
Baseline A: *Zero-shot* CLIP	0.143	0.322	0.411	-0.028	0.192	0.153	0.498	0.414	0.058	0.241	0.346	0.254	0.174	0.053	0.218	0.467
A+Ctx	0.551	0.598	0.637	0.662	0.582	0.569	0.746	0.664	0.627	0.662	0.591	0.670	0.634	0.414	0.634	0.670
A+Ctx+MLP (*w/o* $f_{\mathcal{V}}^{\text{FAST}}$)	0.580	0.626	0.606	0.615	0.618	0.683	0.766	0.662	0.675	0.670	0.593	0.636	0.673	0.433	0.722	0.735
Baseline B: FAST-VQA	0.614	0.630	0.696	0.709	0.646	0.721	0.800	0.724	0.755	0.731	0.751	0.695	0.736	0.654	0.803	0.782
B+CLIP visual (*w/o* textual)	0.644	0.678	0.699	0.723	0.698	0.765	0.815	0.734	0.772	0.748	0.767	0.733	0.743	0.648	0.809	0.802
MaxVQA (all)	**0.681**	**0.701**	**0.757**	**0.749**	**0.712**	**0.775**	**0.825**	**0.748**	**0.776**	**0.761**	**0.782**	**0.748**	**0.763**	**0.684**	**0.827**	**0.813**

5.3 Evaluation on Existing Databases

5.3.1 Training and Testing on Existing Databases. We also train and evaluate the proposed language-prompted MaxVQA on existing VQA databases, with a variant with only single objective on overall score (**O**). As compared in Table 3, the proposed MaxVQA can significantly outperform baseline methods on all datasets, proving the robust excellent performance of its visual-language-based design.

5.3.2 Analyzing Existing Databases with Multi-axis Predictions. In Table 4, we compare between multi-dimensional MaxVQA predictions trained with Maxwell and overall scores in existing datasets. First, we prove that MaxVQA can generalize well from Maxwell to existing datasets on overall quality (**O**) prediction. Furthermore, we observe that unlike KoNViD-1k and YouTube-UGC, the subjective scores in LIVE-VQC are more correlated with predictions of the technical perspective (**T-all**) than overall quality (**O**). This result suggests that subjective studies conducted for different databases might not follow the same scoring standards implicitly. Among the specific factors, we notice that Focus (T-2) is important for all databases, while LIVE-VQC is especially concerned about Motion Blur (T-4) and Fluency (T-8) as all videos are from hand-held smartphones. Moreover, KoNViD-1k is obviously more concerned about pure aesthetic predictions (A-all) and contents (A-1) than others.

5.4 Ablation Studies

We conduct ablation studies to investigate the effects of proposed modules in MaxVQA, as listed in Table 5. Zero-shot CLIP generally performs poorly on specific quality factors, validating our second claim in Sec. 4.1.2 and the necessity of learnable text prompts.

5.4.1 Effects of Enhanced Visual Backbone. Considering the context-prompted CLIP as baseline, we discuss the effects of the proposed enhanced visual backbone. Without the FAST-VQA features, either the original or adapted [9] CLIP features lead to less accuracy, especially for *temporal distortions* ([T-5] Flicker and [T-8] Fluency), where the FAST-VQA features can improve **58%** and **32%**, proving the vitality of including temporal modeling upon baseline CLIP.

5.4.2 Effects of CLIP. From another perspective, we treat the FAST-VQA as our baseline method, and investigate the effects of pre-trained CLIP in the MaxVQA. Directly integrating CLIP visual features can lead to significant improvements, especially on aesthetic dimensions. The text-prompted modeling further notably improve the performance than the naive multi-task variant *without* textual modeling, proving the rationality of our multi-modal design.

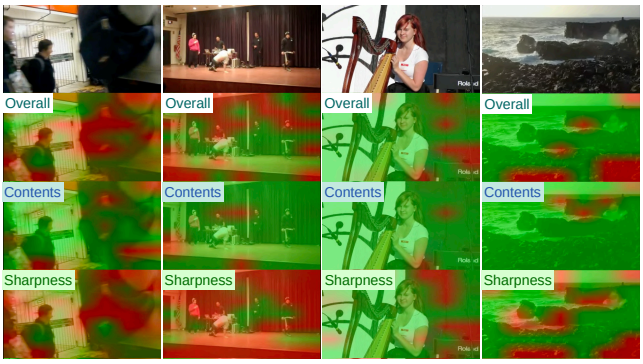

Figure 7: Multi-axis local quality maps for dimensions Overall (O), Contents (A-1) and Sharpness (T-1) of the proposed MaxVQA, showing their differences.

5.5 Multi-Axis Local Quality Maps

As $f_{V_t}^{\text{Final}}$ are localized features, we are able to generate local quality maps (similar as [56]) from different dimensions, so as to qualitatively examine the their quality concerns. As illustrated in Fig. 7, the proposed MaxVQA can comprehensively detect *degraded* or *unappealing* local areas as worse overall quality (O); moreover, the Contents (A-1) axis predicts that human-related areas in videos are with better quality than backgrounds; and the Sharpness (T-1) axis can well-distinguish fuzzy areas (*e.g. backpack in the leftmost video*), suggesting that MaxVQA can distinguish differences among axes.

6 CONCLUSION

Our study significantly expands the scope of in-the-wild Video Quality Assessment (VQA) by explaining subjective quality scores with specific factors. A large-scale in-the-wild VQA database, named **Maxwell**, is created to gather more than two million human opinions across 13 specific quality-related factors, including technical distortions *e.g. noise, flicker* and aesthetic factors *e.g. contents*. With the Maxwell database, we investigate the relationships between various perceptual factors and examine how they influence overall quality opinions. Moreover, the Maxwell establishes a novel multi-dimensional benchmark for objective VQA methods to assess their strengths and weaknesses in capturing various quality issues, providing more detailed guidance for future methods. Additionally, the study introduces the **MaxVQA**, a language-prompted VQA approach that can jointly evaluate multiple specific quality factors and overall perceptual quality scores, achieving state-of-the-art results with excellent generalization abilities. We hope that our efforts will bring along new insights and advancements in the VQA field.

REFERENCES

[1] Recommendation 500-10: Methodology for the subjective assessment of the quality of television pictures. ITU-R Rec. BT.500, 2000.

[2] Antsiferova, A., Lavrushkin, S., Smirnov, M., Gushchin, A., Vatolin, D. S., and Kulikov, D. Video compression dataset and benchmark of learning-based video-quality metrics. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2022).

[3] Chan, K. C., Wang, X., Yu, K., Dong, C., and Loy, C. C. Basicvsr: The search for essential components in video super-resolution and beyond. In *CVPR* (2021).

[4] Chen, B., Zhu, L., Li, G., Lu, F., Fan, H., and Wang, S. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. *IEEE TCSVT* (2021).

[5] Chen, P., Li, L., Ma, L., Wu, J., and Shi, G. Rirnet: Recurrent-in-recurrent network for video quality assessment. *ACM MM* (2020).

[6] Dong, Y., Liu, X., Gao, Y., Zhou, X., Tan, T., and Zhai, G. Light-vqa: A multi-dimensional quality assessment model for low-light video enhancement. In *n Proceedings of the 31st ACM International Conference on Multimedia* (2023).

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[8] Fang, Y., Zhu, H., Zeng, Y., Ma, K., and Wang, Z. Perceptual quality assessment of smartphone photography. In *CVPR*.

[9] Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544* (2021).

[10] Ghadiyaram, D., Pan, J., Bovik, A. C., Moorthy, A. K., Panda, P., and Yang, K.-C. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE TCSVT 28*, 9 (2018), 2061–2077.

[11] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR* (2016), pp. 770–778.

[12] Hosu, V., Hahn, F., Jenadeleh, M., Lin, H., Men, H., Szirányi, T., Li, S., and Saupe, D. The konstanz natural video database (konvid-1k). In *QoMEX* (2017), pp. 1–6.

[13] Hosu, V., Lin, H., Sziranyi, T., and Saupe, D. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE TIP 29* (2020), 4041–4056.

[14] Hou, J., Ding, H., Lin, W., Liu, W., and Fang, Y. Distilling knowledge from object classification to aesthetics assessment. *IEEE TCSVT* (2022).

[15] Hou, J., Lin, W., Fang, Y., Wu, H., Chen, C., Liao, L., and Liu, W. Towards transparent deep image aesthetics assessment with tag-based content descriptors. *IEEE Transactions on Image Processing* (2023).

[16] Hou, J., Yang, S., and Lin, W. Object-level attention for aesthetic rating distribution prediction. In *ACM MM* (2020), p. 816–824.

[17] Huang, Z., Zhang, T., Heng, W., Shi, B., and Zhou, S. Real-time intermediate flow estimation for video frame interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2022).

[18] Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. OpenCLIP, July 2021.

[19] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML* (2021).

[20] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, A., Suleyman, M., and Zisserman, A. The kinetics human action video dataset. *ArXiv abs/1705.06950* (2017).

[21] Ke, J., Ye, K., Yu, J., Wu, Y., Milanfar, P., and Yang, F. Vila: Learning image aesthetics from user comments with vision-language pretraining, 2023.

[22] Korhonen, J. Two-level approach for no-reference consumer video quality assessment. *IEEE TIP 28*, 12 (2019), 5923–5938.

[23] Korhonen, J., Su, Y., and You, J. Blind natural video quality prediction via statistical temporal features and deep spatial features. In *ACM MM* (2020), p. 3311–3319.

[24] Kou, T., Liu, X., Jia, J., Sun, W., Zhai, G., and Liu, N. Stablevqa: A deep no-reference quality assessment model for video stability. In *Proceedings of the 31st ACM International Conference on Multimedia* (2023).

[25] Lee, Y.-C., Tseng, K.-W., Chen, Y.-T., Chen, C.-C., Chen, C.-S., and Hung, Y.-P. 3d video stabilization with depth estimation by cnn-based optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 10621–10630.

[26] Li, B., Zhang, W., Tian, M., Zhai, G., and Wang, X. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE TCSVT* (2022).

[27] Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., Zhai, G., and Lin, W. Agiqa-3k: An open database for ai-generated image quality assessment, 2023.

[28] Li, D., Jiang, T., and Jiang, M. Quality assessment of in-the-wild videos. In *ACM MM* (2019), p. 2351–2359.

[29] Li, D., Jiang, T., and Jiang, M. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision 129*, 4 (2021).

[30] Liao, L., Xu, K., Wu, H., Chen, C., Sun, W., Yan, Q., and Lin, W. Exploring the effectiveness of video perceptual representation in blind video quality assessment. In *ACM MM* (2022).

[31] Liu, X., Van De Weijer, J., and Bagdanov, A. D. Exploiting unlabeled data in cnns by self-supervised learning to rank. *IEEE TPAMI* (2019), 1–1.

[32] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. Video swin transformer. In *CVPR* (2022).

[33] Mittal, A., Soundararajan, R., and Bovik, A. C. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters 20*, 3 (2013), 209–212.

[34] Murray, N., Marchesotti, L., and Perronnin, F. Ava: A large-scale database for aesthetic visual analysis. In *CVPR* (2012), pp. 2408–2415.

[35] Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., and Ling, H. Expanding language-image pretrained models for general video recognition. *ECCV* (2022).

[36] Nuutinen, M., Virtanen, T., Vaahteranoksa, M., Vuori, T., Oittinen, P., and Häkkinen, J. Cvd2014—a database for evaluating no-reference video quality assessment algorithms. *IEEE TIP 25*, 7 (2016).

[37] Park, D., Kim, J., and Chun, S. Y. Down-scaling with learned kernels in multi-scale deep neural networks for non-uniform single image deblurring. *arXiv preprint arXiv:1903.10157* (2019).

[38] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

[39] Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., and Lu, J. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR* (2022).

[40] Seshadrinathan, K., Soundararajan, R., Bovik, A. C., and Cormack, L. K. Study of subjective and objective quality assessment of video. *IEEE TIP 19*, 6 (2010), 1427–1441.

[41] Sinno, Z., and Bovik, A. C. Large-scale study of perceptual video quality. *IEEE TIP 28*, 2 (2019), 612–627.

[42] Sun, W., Min, X., Lu, W., and Zhai, G. A deep learning based no-reference quality assessment model for ugc videos. *arXiv preprint arXiv:2204.14047* (2022).

[43] Tassano, M., Delon, J., and Veit, T. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020).

[44] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. Yfcc100m: The new data in multimedia research. *Commun. ACM 59*, 2 (2016), 64–73.

[45] Tu, Z., Wang, Y., Birkbeck, N., Adsumilli, B., and Bovik, A. C. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE TIP 30* (2021), 4449–4464.

[46] Tu, Z., Yu, X., Wang, Y., Birkbeck, N., Adsumilli, B., and Bovik, A. C. Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing 2* (2021), 425–440.

[47] Vonikakis, V., Subramanian, R., Arnfred, J., and Winkler, S. A probabilistic approach to people-centric photo selection and sequencing. *IEEE Transactions on Multimedia 19*, 11 (2017), 2609–2624.

[48] Vu, P. V., and Chandler, D. M. Vis3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging 23* (2014).

[49] Wallace, G. K. The jpeg still picture compression standard. *Commun. ACM 34*, 4 (1991), 30–44.

[50] Wang, H., Li, G., Liu, S., and Kuo, C.-C. J. Icme 2021 ugc-vqa challenge.

[51] Wang, J., Chan, K. C. K., and Loy, C. C. Exploring clip for assessing the look and feel of images, 2022.

[52] Wang, Y., Inguva, S., and Adsumilli, B. Youtube ugc dataset for video compression research. In *2019 MMSP* (2019).

[53] Wang, Y., Ke, J., Talebi, H., Yim, J. G., Birkbeck, N., Adsumilli, B., Milanfar, P., and Yang, F. Rich features for perceptual quality assessment of ugc videos. In *CVPR* (June 2021), pp. 13435–13444.

[54] Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR* (2022).

[55] Wiegand, T. Draft itu-t recommendation and final draft international standard of joint video specification.

[56] Wu, H., Chen, C., Hou, J., Liao, L., Wang, A., Sun, W., Yan, Q., and Lin, W. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *ECCV* (2022).

[57] Wu, H., Chen, C., Liao, L., Hou, J., Sun, W., Yan, Q., Gu, J., and Lin, W. Neighbourhood representative sampling for efficient end-to-end video quality assessment.

[58] Wu, H., Chen, C., Liao, L., Hou, J., Sun, W., Yan, Q., and Lin, W. Discovqa: Temporal distortion-content transformers for video quality assessment.

[59] Wu, H., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., and Lin, W. Disentangling aesthetic and technical effects for video quality assessment of user generated content.

[60] Wu, H., Liao, L., Hou, J., Chen, C., Zhang, E., Wang, A., Sun, W., Yan, Q., and Lin, W. Exploring opinion-unaware video quality assessment with semantic affinity criterion. In *ICME* (2023).
[61] Wu, H., Liao, L., Wang, A., Chen, C., Hou, J. H., Zhang, E., Sun, W. S., Yan, Q., and Lin, W. Towards robust text-prompted semantic criterion for in-the-wild video quality assessment, 2023.
[62] Xu, J., Li, J., Zhou, X., Zhou, W., Wang, B., and Chen, Z. Perceptual quality assessment of internet videos. In *ACM MM* (2021).
[63] Yang, Y., Xu, L., Li, L., Qie, N., Li, Y., Zhang, P., and Guo, Y. Personalized image aesthetics assessment with rich attributes. In *CVPR* (2022), pp. 19861–19869.
[64] Ying, Z., Mandal, M., Ghadiyaram, D., and Bovik, A. Patch-vq: 'patching up' the video quality problem. In *CVPR* (2021).
[65] Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., and Bovik, A. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *CVPR* (2020).
[66] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models.
[67] Zhang, B., Niu, L., and Zhang, L. Image composition assessment with saliency-augmented multi-pattern pooling. *arXiv preprint arXiv:2104.03133* (2021).
[68] Zhang, W., Ma, K., Yan, J., Deng, D., and Wang, Z. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE TCSVT 30,* 1 (2020), 36–47.
[69] Zhang, W., Zhai, G., Wei, Y., Yang, X., and Ma, K. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE Conference on Computer Vision and Pattern Recognition* (2023).
[70] Zhang, Z., Sun, W., Zhou, Y., Wu, H., Li, C., Min, X., and Liu, X. Advancing zero-shot digital human quality assessment through text-prompted evaluation, 2023.
[71] Zhang, Z., Wu, W., Sun, W., Tu, D., Lu, W., Min, X., Chen, Y., and Zhai, G. Md-vqa: Multi-dimensional quality assessment for ugc live videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).
[72] Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)* (2022).
[73] Zhu, H., Li, L., Wu, J., Dong, W., and Shi, G. MetaIQA: deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jun. 2020), pp. 14143–14152.