CAMP-VQA: Caption-Embedded Multimodal Perception for No-Reference Quality Assessment of Compressed Video
Supplementary Material for WACV 2026

Xinyi Wang Angeliki Katsenou Junxiao Shen David Bull
School of Computer Science, University of Bristol
Bristol, United Kingdom
{xinyi.wang, angeliki.katsenou, junxiao.shen, dave.bull}@bristol.ac.uk

Abstract

This document is the supplementary material for our WACV submission. We provide further explanation and details on frame difference fragmentation, quality-aware caption generation, spatio-temporal feature extraction, and loss functions. Additional figures are included to illustrate the quality-related prompt settings and to demonstrate the effectiveness of the proposed CAMP-VQA in predicting scores across different quality dimensions.

A. Frame Difference Fragmentation

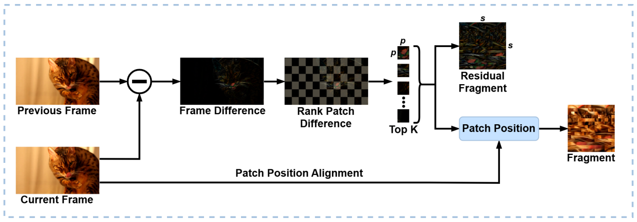


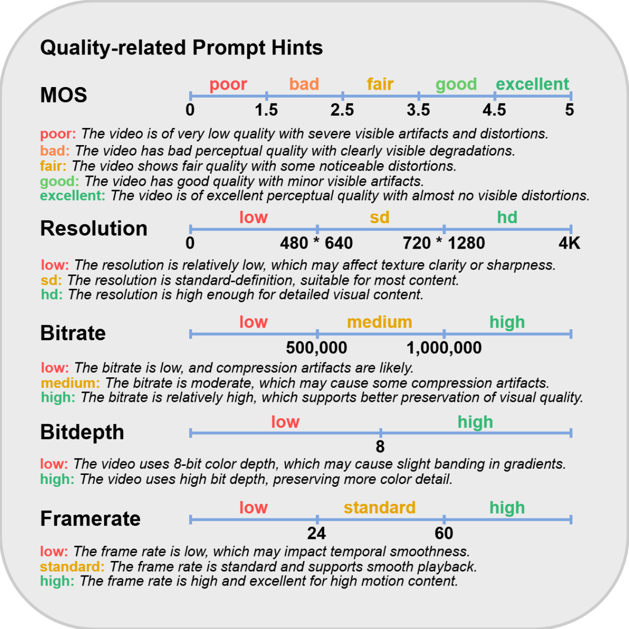
Figure 1. Frame difference fragmentation (FDF) module.

Regarding patch selection, we compute the inter-frame absolute differences for all patches and rank them accordingly. The fragmentation details are shown in (Fig. 1). The number of selected patches K is calculated based on the target model input size $s \times s$ and each patch size $p \times p$:

$$K = \frac{s^2}{p^2}. \tag{1}$$

We then select the top K patches from the ranked list based on the sum of absolute inter-frame differences. To ensure spatial consistency between the extracted frame and residual fragments, each selected patch in the residual is paired with a corresponding fragment extracted from the same pixel position in F_t. For a target size of $s = 224$, we use a patch size of $p = 16$, resulting in $K = 196$. This method generalizes across resolutions, as fragmentation operates directly on the original frames. It is important to note that similar residual intensities can arise from different types of motion, which are likely to be associated with distinct distortion types.

B. Semantic-aligned Feature Extraction



poor: The video is of very low quality with severe visible artifacts and distortions.
bad: The video has bad perceptual quality with clearly visible degradations.
fair: The video shows fair quality with some noticeable distortions.
good: The video has good quality with minor visible artifacts.
excellent: The video is of excellent perceptual quality with almost no visible distortions.

low: The resolution is relatively low, which may affect texture clarity or sharpness.
sd: The resolution is standard-definition, suitable for most content.
hd: The resolution is high enough for detailed visual content.

low: The bitrate is low, and compression artifacts are likely.
medium: The bitrate is moderate, which may cause some compression artifacts.
high: The bitrate is relatively high, which supports better preservation of visual quality.

low: The video uses 8-bit color depth, which may cause slight banding in gradients.
high: The video uses high bit depth, preserving more color detail.

low: The frame rate is low, which may impact temporal smoothness.
standard: The frame rate is standard and supports smooth playback.
high: The frame rate is high and excellent for high motion content.

Figure 2. Quality-related prompt hints derived from video metadata.

Based on video metadata across multiple dimensions, we derive a set of quality-related prompt hints to guide the input prompt, as illustrated in Fig. 2. The figure shows how metadata such as resolution, bitrate, and frame rate are mapped to qualitative levels based on their values, each accompanied by descriptive hints. These mappings provide a structured way to integrate perceptual video quality factors into prompt construction for specific videos. Notably, we avoid disclosing ground-truth MOS scores. During training, quality-related hints are generated from quantized quality levels (ranging from poor to excellent) derived from metadata, guiding prompt text generation without including explicit scores in the descriptions. During inference, the quality level is set to prediction mode, eliminating the need for MOS.

Upon these hints, we design different quality-aware prompt settings, as illustrated in Fig. 3. The quality prompt is used to extract general perceptual captions at the frame level across multiple visual attributes. The fragment prompt focuses on localized degradations, and the residual prompt captures temporal differences between consecutive frames, jointly providing artifact captions at the fragment level. In the ablation study, we also include a content prompt to analyze the role of content descriptions in low-level video quality assessment. Each prompt is tailored to capture a different aspect of perceptual quality.

C. Spatio-temporal Feature Extraction

C.1. Spatial Vision Extractor

The Swin Transformer (SwinT) [3] employs a hierarchical sliding window mechanism and utilizes local self-attention to capture spatial features and long-range dependencies within images. To make full use of its strong spatial modeling capacity, we discard the classification head and retain only the backbone as our Spatial Vision Extractor (SVE) module.

Input frames of a video clip in the form of a tensor $\mathbf{X}_{\text{frame}} \in \mathbb{R}^{N\times C\times H\times W}$, where N indicates the number of frames, C is the number of channels, and H and W are the spatial dimensions. Each frame tensor is encoded by the SwinT backbone network ϕ_s, and compact spatial representations are generated through GAP:

$$z_{\text{swint}} = \text{GAP}(\phi_{\text{swin}}(\mathbf{X}_{\text{frame}})) \in \mathbb{R}^{N\times d_s}, \tag{2}$$

where d_s denotes the dimension of the spatial features. SVE module captures the spatial structural information of video frames, enabling spatial awareness for subsequent temporal modeling and multimodal fusion.

C.2. Temporal Motion Extractor

To enhance the model's capacity for spatio-temporal feature modeling in videos, we designed a Temporal Motion Extractor (TME) module based on the SlowFast [2] architecture. TME uses a dual pathway to extract features across different temporal scales, enabling effective capture of quality fluctuations in videos while maintaining efficiency.

We construct two pathways with different temporal divisions by using an input frame sequence tensor $\mathbf{X}_{\text{sequence}} \in \mathbb{R}^{B\times C\times T\times H\times W}$, where B is batch size, C is channels, T is temporal length, and H, W are the height and width. The slow pathway extracts low-frequency changes over longer temporal spans by downsampling at a rate of $r = \frac{1}{4}$, denoted as $X_{\text{slow}} = \text{Sample}(\mathbf{X}_{\text{sequence}}, r)$. The fast pathway retains the original frame rate, i.e., $X_{\text{fast}} = \mathbf{X}_{\text{sequence}}$, and serves to capture rapidly high-frequency details. The two pathways are fed into feature extractors ϕ_{s} and ϕ_{f}, respectively. Spatio-temporal GAP is then applied to produce compact feature representations:

$$\mathbf{z}_{\text{slow}} = \text{GAP}(\phi_{\text{s}}(\mathbf{X}_{\text{slow}})),\ \mathbf{z}_{\text{fast}} = \text{GAP}(\phi_{\text{f}}(\mathbf{X}_{\text{fast}})). \tag{3}$$

Finally, the two feature vectors are concatenated to form the temporal feature:

$$\mathbf{z}_{\textbf{slowfast}} = [\mathbf{z}_{\text{slow}}; \mathbf{z}_{\text{fast}}] \in \mathbb{R}^{B\times(d_s+d_f)}, \tag{4}$$

where d_s and d_f are the feature dimensions of the slow and fast channels, respectively. This module extracts temporal motion features at different time scales, enhancing the model's sensitivity to temporal quality variations.

D. Loss Functions

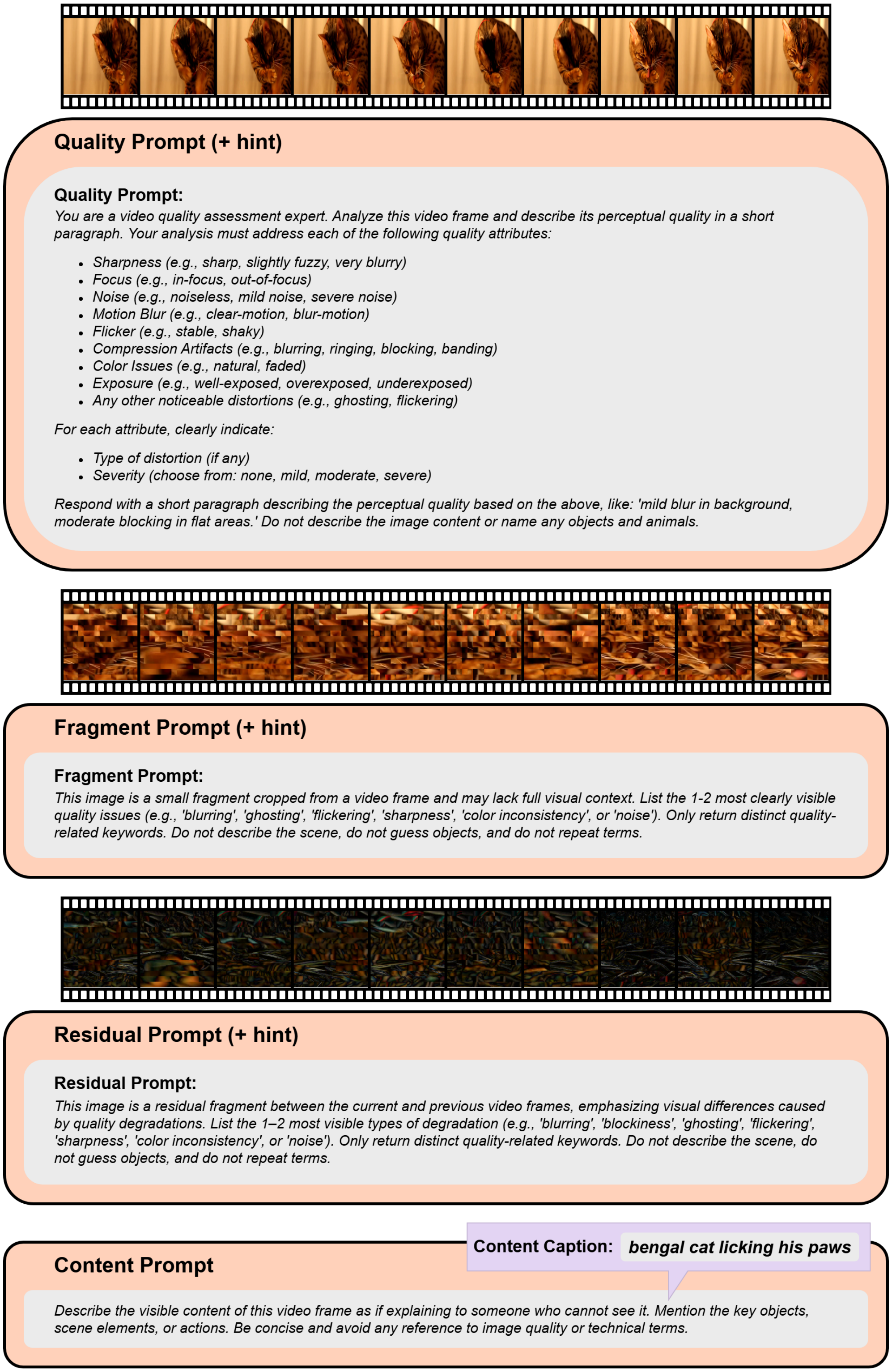
We adopted a composite loss function [4] to enhance accuracy and ranking consistency simultaneously. The first part of the loss function is the precision loss $\mathcal{L}_{\text{p}}$, which quantifies the mean absolute difference between the predicted regression score $\hat{y}\_i$ and the ground truth (MOS) y_i. It is defined as:

$$\mathcal{L}_{\text{p}} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| \,. \tag{5}$$

The second is the ranking loss $\mathcal{L}_{\text{r}}$, which preserves ordinal consistency by comparing the relative rankings and computing the pairwise differences between sample pairs. It is defined as:

$$\mathcal{L}_{\text{r}} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \max\left(0, |y_i - y_j| - s_{ij}\cdot(\hat{y}_i - \hat{y}_j)\right), \tag{6}$$

where N is the number of videos, and i and j represent the indices of video samples. The sign weight s_{ij} represents the ranking relationship inferred from the ground truth scores and is defined as $s_{ij} = 1$ if $y_i \geq y_j$, and $s_{ij} = -1$ if $y_i < y_j$. The composite loss $\mathcal{L}_c$ is formulated as a weighted combination of the two components: $\mathcal{L}_c = \lambda_1 \cdot \mathcal{L}_{\text{p}} + \lambda_2 \cdot \mathcal{L}_{\text{r}}$. Here, $\lambda \in \mathbb{R}$ balances the two loss terms: λ_1 improves prediction accuracy, while λ_2 enhances ordinal consistency. The loss is a weighted combination with $\lambda_1 = 0.6$ and $\lambda_2 = 1$.



Quality Prompt (+ hint)

Quality Prompt:

You are a video quality assessment expert. Analyze this video frame and describe its perceptual quality in a short paragraph. Your analysis must address each of the following quality attributes:

- *Sharpness (e.g., sharp, slightly fuzzy, very blurry)*
- *Focus (e.g., in-focus, out-of-focus)*
- *Noise (e.g., noiseless, mild noise, severe noise)*
- *Motion Blur (e.g., clear-motion, blur-motion)*
- *Flicker (e.g., stable, shaky)*
- *Compression Artifacts (e.g., blurring, ringing, blocking, banding)*
- *Color Issues (e.g., natural, faded)*
- *Exposure (e.g., well-exposed, overexposed, underexposed)*
- *Any other noticeable distortions (e.g., ghosting, flickering)*

For each attribute, clearly indicate:

- *Type of distortion (if any)*
- *Severity (choose from: none, mild, moderate, severe)*

Respond with a short paragraph describing the perceptual quality based on the above, like: 'mild blur in background, moderate blocking in flat areas.' Do not describe the image content or name any objects and animals.

Fragment Prompt (+ hint)

Fragment Prompt:

This image is a small fragment cropped from a video frame and may lack full visual context. List the 1-2 most clearly visible quality issues (e.g., 'blurring', 'ghosting', 'flickering', 'sharpness', 'color inconsistency', or 'noise'). Only return distinct quality-related keywords. Do not describe the scene, do not guess objects, and do not repeat terms.

Residual Prompt (+ hint)

Residual Prompt:

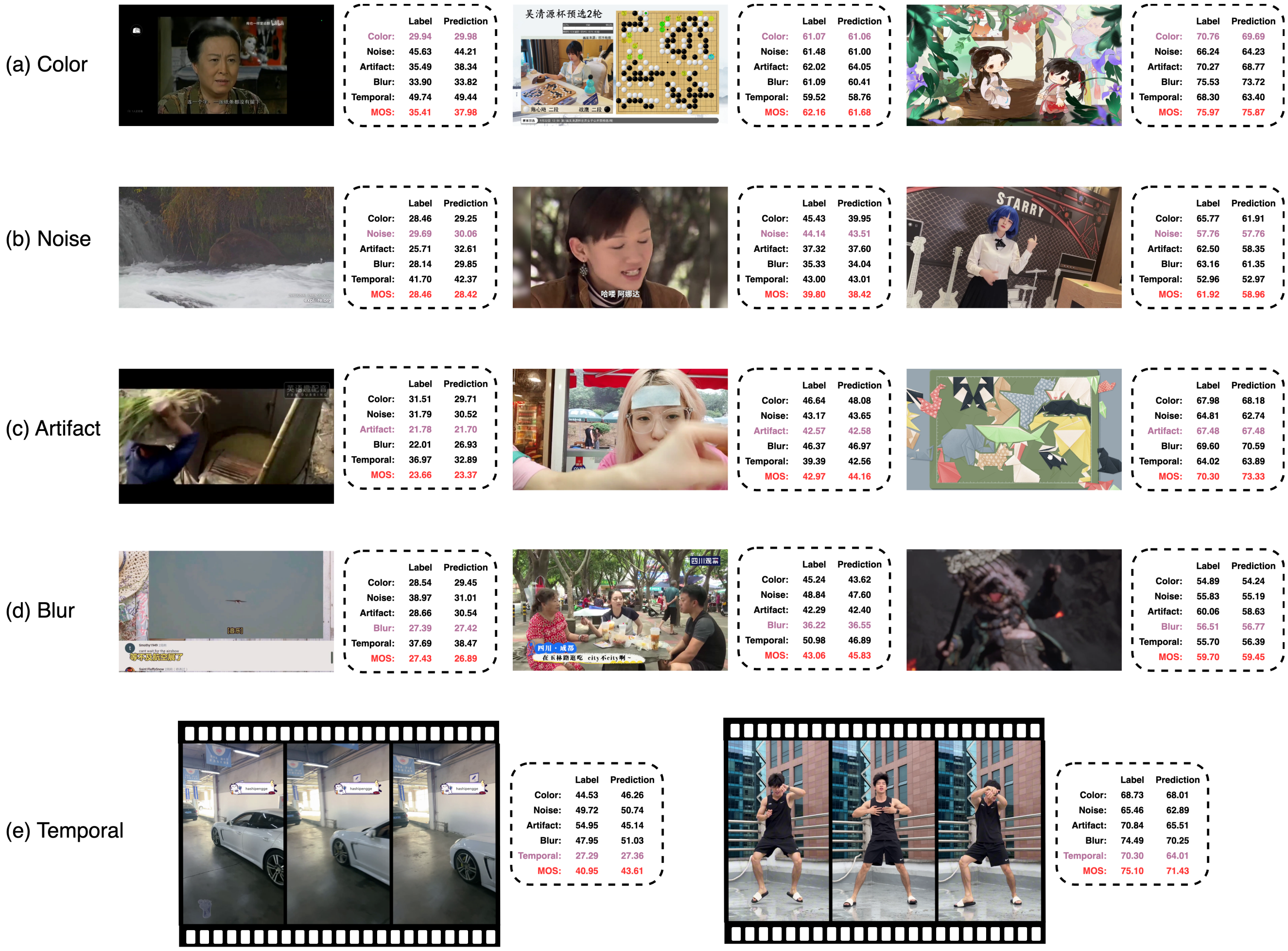
This image is a residual fragment between the current and previous video frames, emphasizing visual differences caused by quality degradations. List the 1–2 most visible types of degradation (e.g., 'blurring', 'blockiness', 'ghosting', 'flickering', 'sharpness', 'color inconsistency', or 'noise'). Only return distinct quality-related keywords. Do not describe the scene, do not guess objects, and do not repeat terms.

Content Prompt

Content Caption: ***bengal cat licking his paws***

Describe the visible content of this video frame as if explaining to someone who cannot see it. Mention the key objects, scene elements, or actions. Be concise and avoid any reference to image quality or technical terms.

Figure 3. Different quality-aware prompt settings: quality prompt, fragment prompt, and residual prompt. We also include a content prompt for the ablation study.



(a) Color

	Label	Prediction
Color:	29.94	29.98
Noise:	45.63	44.21
Artifact:	35.49	38.34
Blur:	33.90	33.82
Temporal:	49.74	49.44
MOS:	35.41	37.98

	Label	Prediction
Color:	61.07	61.06
Noise:	61.48	61.00
Artifact:	62.02	64.05
Blur:	61.09	60.41
Temporal:	59.52	58.76
MOS:	62.16	61.68

	Label	Prediction
Color:	70.76	69.69
Noise:	66.24	64.23
Artifact:	70.27	68.77
Blur:	75.53	73.72
Temporal:	68.30	63.40
MOS:	75.97	75.87

(b) Noise

	Label	Prediction
Color:	28.46	29.25
Noise:	29.69	30.06
Artifact:	25.71	32.61
Blur:	28.14	29.85
Temporal:	41.70	42.37
MOS:	28.46	28.42

	Label	Prediction
Color:	45.43	39.95
Noise:	44.14	43.51
Artifact:	37.32	37.60
Blur:	35.33	34.04
Temporal:	43.00	43.01
MOS:	39.80	38.42

	Label	Prediction
Color:	65.77	61.91
Noise:	57.76	57.76
Artifact:	62.50	58.35
Blur:	63.16	61.35
Temporal:	52.96	52.97
MOS:	61.92	58.96

(c) Artifact

	Label	Prediction
Color:	31.51	29.71
Noise:	31.79	30.52
Artifact:	21.78	21.70
Blur:	22.01	26.93
Temporal:	36.97	32.89
MOS:	23.66	23.37

	Label	Prediction
Color:	46.64	48.08
Noise:	43.17	43.65
Artifact:	42.57	42.58
Blur:	46.37	46.97
Temporal:	39.39	42.56
MOS:	42.97	44.16

	Label	Prediction
Color:	67.98	68.18
Noise:	64.81	62.74
Artifact:	67.48	67.48
Blur:	69.60	70.59
Temporal:	64.02	63.89
MOS:	70.30	73.33

(d) Blur

	Label	Prediction
Color:	28.54	29.45
Noise:	38.97	31.01
Artifact:	28.66	30.54
Blur:	27.39	27.42
Temporal:	37.69	38.47
MOS:	27.43	26.89

	Label	Prediction
Color:	45.24	43.62
Noise:	48.84	47.60
Artifact:	42.29	42.40
Blur:	36.22	36.55
Temporal:	50.98	46.89
MOS:	43.06	45.83

	Label	Prediction
Color:	54.89	54.24
Noise:	55.83	55.19
Artifact:	60.06	58.63
Blur:	56.51	56.77
Temporal:	55.70	56.39
MOS:	59.70	59.45

(e) Temporal

	Label	Prediction
Color:	44.53	46.26
Noise:	49.72	50.74
Artifact:	54.95	45.14
Blur:	47.95	51.03
Temporal:	27.29	27.36
MOS:	40.95	43.61

	Label	Prediction
Color:	68.73	68.01
Noise:	65.46	62.89
Artifact:	70.84	65.51
Blur:	74.49	70.25
Temporal:	70.30	64.01
MOS:	75.10	71.43

Figure 4. Predicted scores of CAMP-VQA and ground-truth MOS of FineVD [1] across different quality dimensions

E. Quality Scoring On Different Dimensions

Building on our quality-aware captioning, we examined CAMP-VQA's impact on video quality scoring across different dimensions of the FineVD [1] dataset. Example videos with predicted CAMP-VQA scores and ground-truth FineVD MOS are shown in Figs. 4.

References

[1] Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, 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. 4

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6202–6211, 2019. 2

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[4] Shaoguo Wen and Junle Wang. A strong baseline for image and video quality assessment. *arXiv preprint:2111.07104*, 2021. 2