NTIRE 2025 XGC Quality Assessment Challenge: Methods and Results Supplementary Material

1. Challenge Methods

1.1. User-generated Video Track

1.1.1 SLCV

Table 1. The performance of the proposed method for track 1 challenge

Model	Test MainScore	Test PLCC	Test SRCC
InternVL2.5_1B	0.8615	0.8617	0.8610
InternVL2.5_2B	0.8617	0.8637	0.8598
InternVL2.5_26B	0.8730	0.8738	0.8723
AVG of 2 models	0.8657	0.8667	0.8645
AVG of 3 models	**0.8731**	**0.8738**	**0.8724**

Team SLCV wins the championship in the user-generated video track. Unlike conventional approaches that rely on regression or classification for video quality assessment (e.g., LIQE [59], Q-Align [53], Fast-VQA [51], and SimpleVQA [40]), their method leverages a multimodal large language model (MLLM) to estimate video quality. In InternVL 2.5 [3], an effective data filtering process was introduced, leveraging large language model (LLM) scoring to evaluate and remove low-quality samples, thereby improving the overall quality of the training data. Inspired by this capability of InternVL 2.5 to assess data quality using LLM-based scoring, they adopt a multimodal large language model (MLLM) for estimating video quality in our work. Specifically, they directly utilize the InternVL 2.5 model as the MLLM to achieve robust and reliable video quality assessment . To overcome the limitation in the spatial domain, they introduce Spatial Window Sampling as a data augmentation strategy. Specifically, they employ a sliding window approach that crops the original video frames with a window size set to 3/4 of the video's longest side. This method effectively triples the amount of training data, thereby enhancing the model's ability to learn fine-grained spatial features. They employ the LoRA (Low-Rank Adaptation) method to efficiently fine-tune the InternVL 2.5 model, enabling it to perform the six fine-grained quality assessments. The overall framework is depicted in Figure 1. During inference, the same data processing strategy used during training is applied to the test videos. Specifically, the model independently predicts quality scores for the three sub-videos generated by the sliding window sampling process. The final prediction is then obtained by averaging the results across these sub-videos. This approach not only ensures robust training but also facilitates accurate and reliable evaluation of fine-grained video quality.

In experiments, they utilize a machine equipped with 8 NVIDIA A100 GPUs (each with 80 GB of memory) and 1 TB of system memory for both the training and inference phases. To address the XGC Quality Assessment Track 1 challenge under limited training data conditions, they propose a hierarchical framework that integrates video preprocessing, parameter-efficient fine-tuning, and multi-level ensembling. Specifically, input videos are divided into three overlapping segments, each determined by 3/4 of their longest spatial dimension, to strike a balance between computational efficiency and contextual preservation. The InternVL2.5 series models (1B, 2B, and 26B) were fine-tuned using Low Rank Adaptation (LoRA) with hyperparameters (rank=16, alpha=64) and a learning rate of 1e-4. Training was performed on 8 NVIDIA A100-80G GPUs with a batch size of 16 for a single epoch to mitigate the risk of overfitting. During inference, predictions for each video segment were generated in parallel across the 8 GPUs. Final scores were obtained through clip-level averaging and model-level ensembling across all three InternVL2.5 variants, ensuring robust and reliable quality assessments. This approach achieved state-of-the-art performance with a score of 0.8731, showcasing the effectiveness of combining spatial segmentation, parameter-efficient adaptation, and hierarchical score aggregation. As demonstrated in Table 1, the ensemble of InternVL2.5-1B and InternVL2.5-2B models achieved an average performance improvement of 0.4% compared to individual models. By further incorporating the InternVL2.5- 26B model, system performance was significantly enhanced, reaching a final score of 0.8731. These results highlight the substantial benefits of multi-model collaboration and the effectiveness of proposed hierarchical framework in achieving state-of-the-art performance. They finally get the averaged main score of 0.8731.

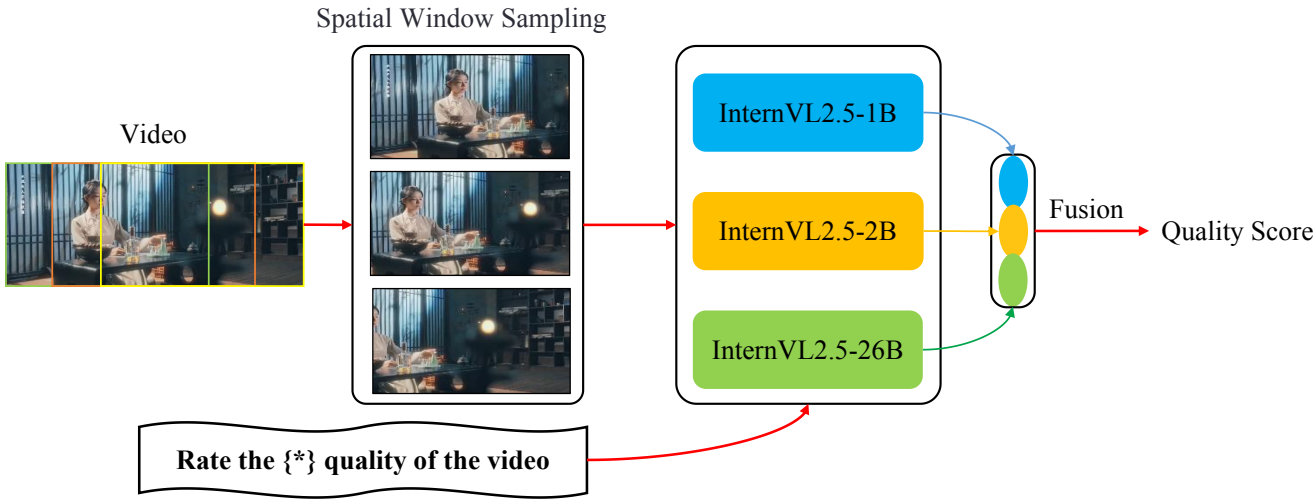

Figure 1. Overview diagram of the proposed method of team SLCV.

1.1.2 SJTU-MOE-AI

Team SJTU-MoE-AI [61] wins second place in the user-generated video track. They propose a multi-dimensional quality assessment for UGC Videos via modular multi-modal vision-language models.They use both single-modal and multiple multi-modal networks for learning. The fine-grained user-generated content (UGC) video quality assessment (VQA) task involves multi-dimensional quality evaluations. Specifically, a VQA model is expected to assess UGC videos across six quality dimensions: color, noise, artifacts, blur, temporal consistency, and overall quality. To tackle this task, they derive two primary insights: first, the employed computational architecture should have sufficient capacity; second, an effective training strategy is needed to optimize the model across all six evaluation aspects.

For the computational architecture, they consider using multi-modal vision-language models trained with both contrastive [35] and autoregressive objectives as the base quality evaluator. Specifically, they employ SigLIP2 [46] and a variant of Q-Align [53]. Given a video, the base quality evaluators generate frame-level quality predictions for eight key frames uniformly sampled from the entire sequence. The final video-level quality score y_b is then obtained by averaging these predictions. For SigLIP-2, they first compute the cosine similarities between the vision representation and textual embedding derived from multiple textual templates: "a photo with {d} {q} quality". where q $\in$ {"bad", "poor", "fair", "good", "perfect"}, represents the five Likert-scale quality levels, and d $\in$ {"color", "noise", "artifact", "blur", "temporal", "overall"} corresponds to different quality dimensions. Following [59], they separately apply a Softmax function to the cosine similarity logits of each quality dimension to obtain the quality distributions over the five quality levels, which are then converted into scalar quality scores via weighted summation (see Figure 2 (b)). They choose the NaFlex version of SigLIP-2, as it is designed to adaptively handle input images (video frames in this task) with varying aspect ratios—a common characteristic of UGC videos. For Q-Align, they follow its default setup for computing quality scores, with one key modification: during training, they alternate conversations associated with different quality dimensions in batches. During inference, they evaluate each quality dimension separately by posing the corresponding question, *e.g.*, *How do you rate the color quality of this video?*, *How do you rate the overall quality of this video?*, etc.

Inspired by [49], they enhance the base quality evaluator with two complementary modules: a degradation perception module and a temporal perception module. The degradation perception module extracts distortion-aware features from sampled video frames within the feature space of ARNIQA [1], which is exposed to diverse distortion types during training. The temporal perception module leverages a SlowFast model [13] to address artifacts resulting from motion anomalies. The representations of both modules are processed through separate multi-layer perceptrons (MLPs). As shown in Fig. 2, each rectifier produces a tuple of scale (α) and shift (β) parameters, which are applied to adjust the base quality predictions y_b as follows:

$$y = \sqrt{\alpha_d \alpha_t} y_b + \frac{\beta_s + \beta_t}{2} \tag{1}$$

Given two videos, their relative quality rankings may differ across various quality dimensions. This motivates us to model the relative quality rankings of each video pair as a joint distribution across all six quality dimensions. Correspondingly, we compute an average of six binary fidelity

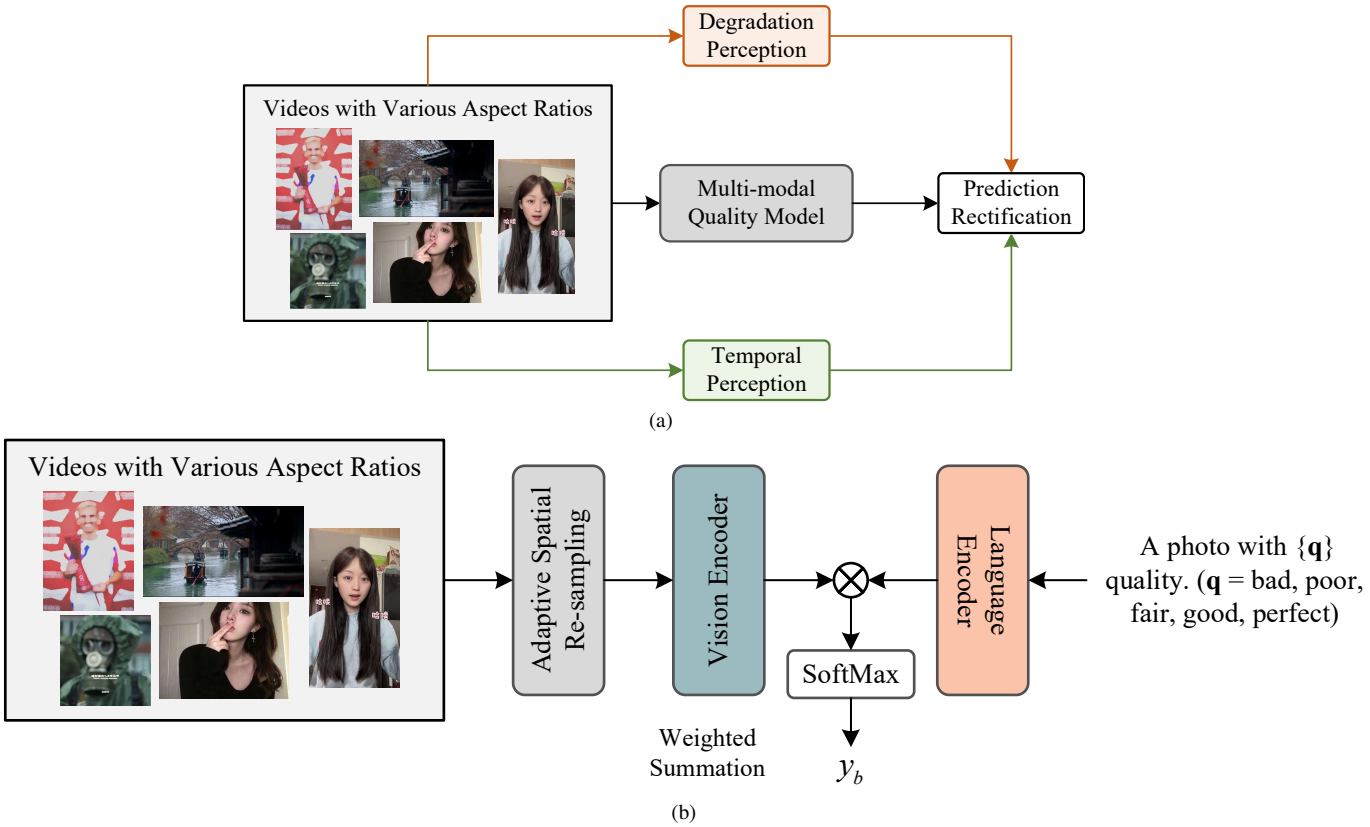

Figure 2. (a) The overall framework of team SJTU-MOE-AI proposed model. (b) SigLIP2 as the base quality evaluator.

losses [45]:

$$\ell_f(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{S}\sum_{s=1}^{S}\Big(1 - \sqrt{p^{(s)}(\boldsymbol{x},\boldsymbol{y})\hat{p}^{(s)}(\boldsymbol{x},\boldsymbol{y})} - \sqrt{(1-p^{(s)}(\boldsymbol{x},\boldsymbol{y}))(1-\hat{p}^{(s)}(\boldsymbol{x},\boldsymbol{y}))}\Big), \quad (2)$$

where s indexes the six quality dimensions ($S = 6$), $p^{(s)}(\boldsymbol{x},\boldsymbol{y})$ is the binary label of video pair $(\boldsymbol{x},\boldsymbol{y})$ according to the s-th quality dimension, which can be inferred from their ground-truth mean opinion scores (MOSs):

$$p^{(s)}(\boldsymbol{x},\boldsymbol{y}) = \begin{cases} 1 & \text{if } q^{(s)}(\boldsymbol{x}) \geq q^{(s)}(\boldsymbol{y}) \\ 0 & \text{otherwise} \end{cases}. \quad (3)$$

$\hat{p}^{(s)}(\boldsymbol{x},\boldsymbol{y})$ is the estimated probability that the quality of $\boldsymbol{x}$ is higher than $\boldsymbol{y}$ in terms of the s-th dimension. Under the Thurstone's model [43], this can be computed as:

$$\hat{p}^{(s)}(\boldsymbol{x},\boldsymbol{y}) = \Phi\left(\frac{\hat{q}^{(s)}(\boldsymbol{x}) - \hat{q}^{(s)}(\boldsymbol{y})}{\sqrt{2}}\right), \quad (4)$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution function, and the variance is fixed to one. We use ℓ_f along with PLCC loss and mean absoluate error (MAE) loss to optimize the model, aiming to improve both the precision and monotonicity of quality predictions: $\ell_f + \ell_p + 0.1 \times \ell_{mae}$.

In experiments, they train two variants of SigLIP2-based models, based on SigLIP2-base-patch16-naflex, and SigLIP2-so400m-patch16-naflex, respectively. They also train a Q-Align-based model using the same data. They train SigLIP2-based models for 15 epochs, minimizing the hybrid loss function using the Adam optimizer with a batch size of 16 and 6 for SigLIP2-base-patch16-naflex and SigLIP2-so400m-patch16-naflex, respectively. The initial learning rate is set to 5e-6. The maximum number of patches is set to 900 and 576 for SigLIP2-base-patch16-naflex and SigLIP2-so400m-patch16-naflex, respectively. They train Q-Align for 3 epochs with an initial learning rate of 2e-5. After training, they freeze the Q-Align model and incorporate into the modular framework, which is exposed to additional 10 epochs of training. They perform full fine-tuning for SigLIP2-base-patch16-naflex and lora finetuning for SigLIP2-so400m-patch16-naflex along with Q-Align. All the experiments are conducted on a single NVIDIA A5880ada GPU.They use the entire test dataset for evaluation, and the pretrained model is initialized with the saved

weights from the final epoch of the training phase. They finally get the averaged main score of 0.8620.

1.1.3 MiVQA

Team MiVQA wins third place in the user-generated video track. Their method is based on RQ-VQA [42]. As shown in Figure 3, it employs a Swin Transformer-B [25] to learn spatial quality feature and utilizes SlowFast [14] for modeling motion characteristics. To further enhance its capability in perceiving video quality, the model incorporates additional features extracted from LIQE [60], Q-Align [54], and FAST-VQA [52]. These features are subsequently concatenated to form a comprehensive representation of the video's quality. Finally, six distinct two-layer MLP heads are utilized to predict the video's quality across multiple dimensions, including color, noise, artifact, blur, temporal, and overall quality. Notably, only the Swin Transformer-B and the six MLP heads are trainable, while the remaining modules are frozen during training.

They employed the model trained by [42] on LSVQ [57] to initialize the Swin Transformer-B. As for SlowFast, LIQE, Q-Align, and FAST-VQA, they utilized the weights officially provided by their respective developers. For Swin Transformer-B, LIQE, and Q-Align, they extract a key frame from every one-second video segment. For SlowFast, the video resolution is adjusted to 224×224, and the video is divided into one-second segments to capture temporal features. FAST-VQA features are generated from the entire video using fragment sampling [52].

During training, they randomly split the training data into ten different training-validation sets and trained ten models. The video quality score is computed by averaging the quality scores obtained from these models. During testing, to enhance performance, they employ data augmentation during the testing phase. Specifically, each image is horizontally flipped, effectively transforming a single image into two parallel inputs. The final prediction scores are then averaged to produce the final result.

In experiments, they implement their model on the PyTorch framework using NVIDIA RTX 3090 GPUs. The model is trained on 4 GPUs with a batch size of 12 for 30 epochs, which takes 14 hours. The initial learning rate is set to 2e-5. They use the Adam optimizer [19] with a weight decay of 1e-7 and the StepLR scheduler with a step size of 10 and a decay ratio of 0.9. They finally get the averaged main score of 0.8440.

1.1.4 XGC-Go

Initially, the characteristics of the data distribution in five distinct evaluation dimensions are systematically examined, and their mutual correlations are investigated. Subsequently, a weighted, data-independent information loss metric is developed through discriminative feature extraction from each dimension's unique informational aspect. The contributions are listed as follows: 1) methods for data analysis, 2) reasons for using large model features, and 3) reference ideas on how to use large model features and the loss weight design.

Analysis 1: Correlation. All dimensions of labels include: 1) color, 2) noise, 3) artifacts, 4) blur, 5) temporal, and 6) overall.

- It can be seen that the distribution is close to the same, probably on a scale from 0 to 100.
- It is close to the sigmoid function distribution, possibly with machine assignments for labeling, so there should be some training noise.
- In particular, the second dimension, noise, exceeds the minimum value of 0.
- The data score distribution is close to a Gaussian distribution, which is suitable for PLCC correlation training.
- The relationship between the first five dimensions and the sixth dimension is calculated.
- The relation between dimensions 1 and 6 is 0.90102.
- The relation between dimensions 2 and 6 is 0.88396.
- The relation between dimensions 3 and 6 is 0.94065.
- The relation between dimensions 4 and 6 is 0.94059.
- The relation between dimensions 5 and 6 is 0.86511.
- It can be seen that the correlation between the second dimension (noise) and the fifth dimension (temporal) is weak. Therefore, focusing on learning these dimensions is necessary.
- The average of all dimensions against all in the annotation is calculated as 0.9867, indicating that the definition of all dimensions should be the close mean of all dimensions.

Analysis 2: Information independence. The information independence(values for each dimension minus mean of all values for a single video) across samples is then sorted from largest to smallest:

- It can be seen that the second dimension has the least information independence, with special cases fluctuating greatly (up to 37).
- Compared with the sixth dimension, the fluctuation can be seen to be smaller, and the number of information independence is larger.
- Other dimensions have some missing combinations of information. (i.e., a direct combination of different points for each dimension). This will lead to the problem of: a) imbalanced data and imbalanced training samples, which will lead to the learning of unwanted features.b) less training data in different dimension combinations, which will lead to low generalization.

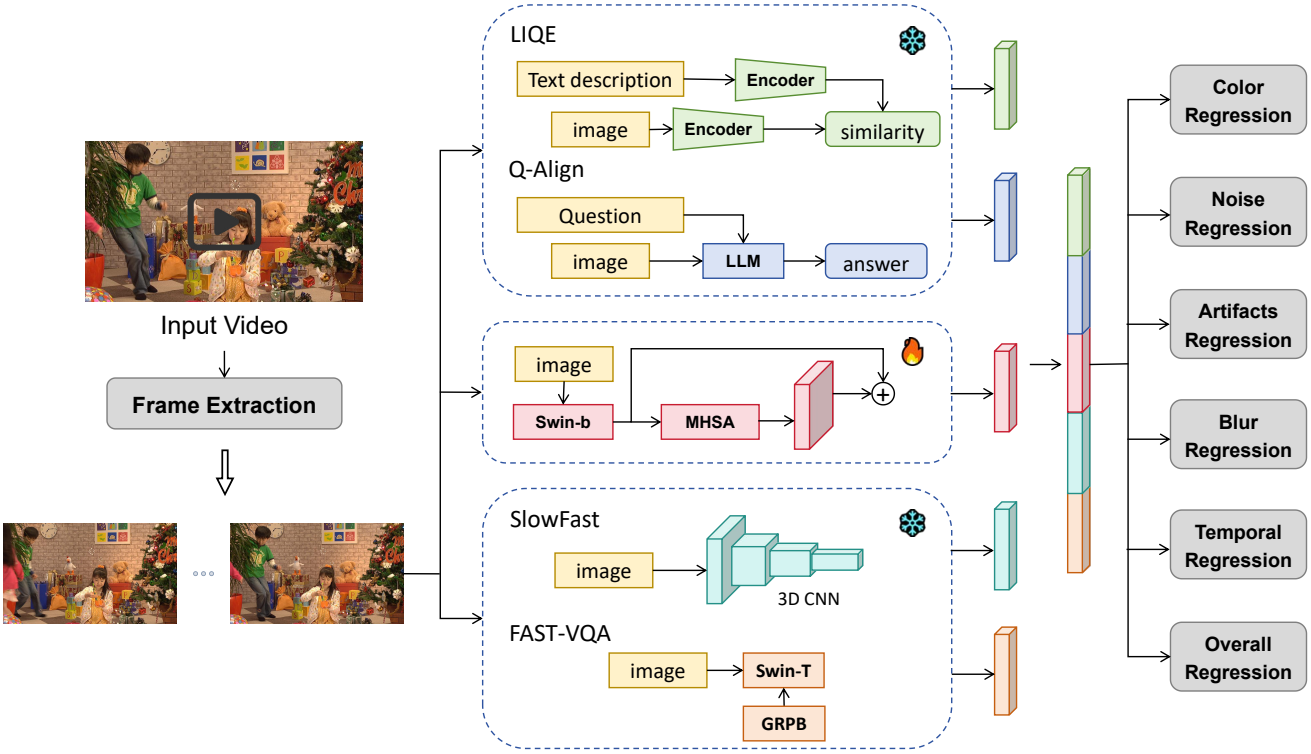

Figure 3. Overview of team MiVQA proposed method.

Analysis 3: Data Situation. As shown in Figure 5, different colors represent different combinations of information, and the numbers represent richness. In the case of training and test sets, the model tends to learn the information from the yellow features more easily, while it is weaker with the blue features. This leads to weak generalization.

Reason 1: Data Characteristics.

- The training set is noisy.
- The hard and easy samples are unbalanced.
- There is a data information combination balance problem.
- It is difficult to fit a model that generalizes well when trained on some training data already available.

Reason 2: Model Characteristics.

- The aim is to borrow world knowledge to help understand relevant qualities and reasons for using large model features.
- However, the feature information of different large models is not aligned.
- At the same time, different large models have different degrees of performance for different videos.
- Different large models also have varying levels of effectiveness for different dimensional information.

Methods. Based on the above analysis, all dimensional data is consolidated into a 0 to 100 score. At the same time, all six dimensions are trained simultaneously, with weights set according to the dataset analysis. Focused learning is performed on the second and fifth dimensions. The following loss functions are defined:

$$loss_a = W_{11} \cdot \text{PLCC}_1 + W_{12} \cdot \text{PLCC}_2 + \cdots + W_{16} \cdot \text{PLCC}_6 \tag{5}$$

$$loss_b = W_{21} \cdot \text{Rank}_1 + W_{22} \cdot \text{Rank}_2 + \cdots + W_{26} \cdot \text{Rank}_6 \tag{6}$$

$$loss_{all} = loss_a + loss_b \tag{7}$$

The large model reference selection includes:

- 1: Q-ALIGN: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels,
- 2: CLIPVQA: Video Quality Assessment via CLIP,
- 3: Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective,
- 4: Enhancing Blind Video Quality Assessment with Rich Quality-aware Features,
- 5: DepictQA: Depicted Image Quality Assessment with Vision Language Models,
- 6: Teaching Large Language Models to Regress Accurate Image Quality Scores Using Score Distribution,

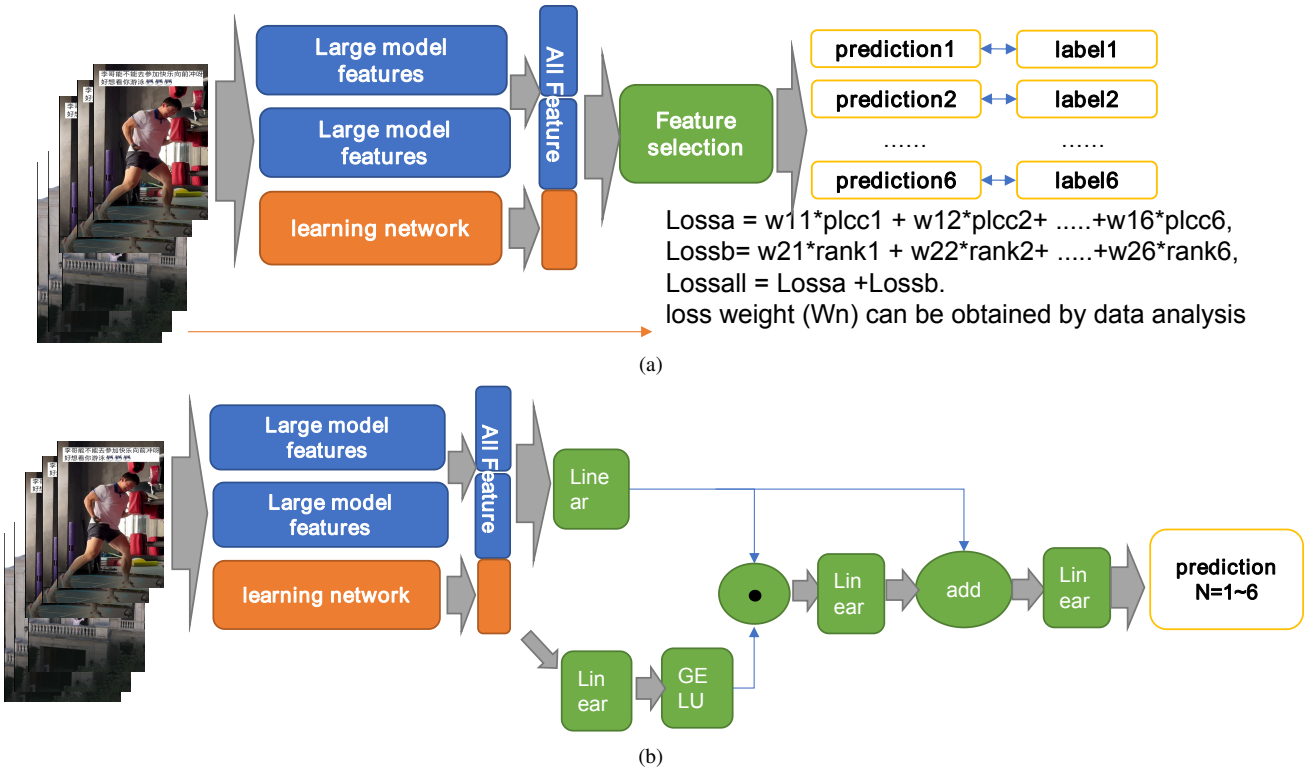

Figure 4. The overall framework of team XGC-Go proposed model.

Figure 5. Example diagram of the data situation

- 7: Descriptive Image Quality Assessment in the Wild,
- 8: Depicting Beyond Scores: Advancing Image Quality Assessment through Multi-modal Language Models, and so on.

The parameters of the large models are frozen, as training large models with less data is difficult. They also design the selection module, as shown in Figure 4, which is effective in some data cases. This will serve as an important reference and idea for better utilization and selection of appropriate large model features. After the activation function, the feature is calculated with the original feature, which has the function of selection. It is desirable to select better large model features. Reference can be made to more complex networks, such as nets for sensor fusion.

In experiments, the optimizer used is Adam with a learning rate of 0.000015 during training. The training was conducted on a GPU (A30) for 48 hours, with no external datasets involved. The training strategies include a two-stage approach, loss weight adjustment, and hyperreference. A comparison is also made during the training process. They finally get the averaged main score of 0.8248.

1.1.5 FoodVQA

Team FoodVQA proposes a multi-level distortion adaptation and spatiotemporal cross-attention fusion framework for VQA, named MACA-VQA. Specifically, a novel multi-level adaptive strategy progressively incorporates distortion information into each Transformer layer of the CLIP model, enabling layer-wise fusion of semantic and distortion features. Furthermore, a newly introduced cross-attention fusion mechanism dynamically integrates spatiotemporal features, capturing complex, multi-dimensional interactions.

As shown in Figure 6, their framework comprises four main components: pre-trained feature extractors, multi-level distortion adaptation module, spatiotemporal cross-attention fusion module, and quality regressor. The pre-trained feature extractors contains three extractors, including a distortion feature extractor, semantic feature extrac-

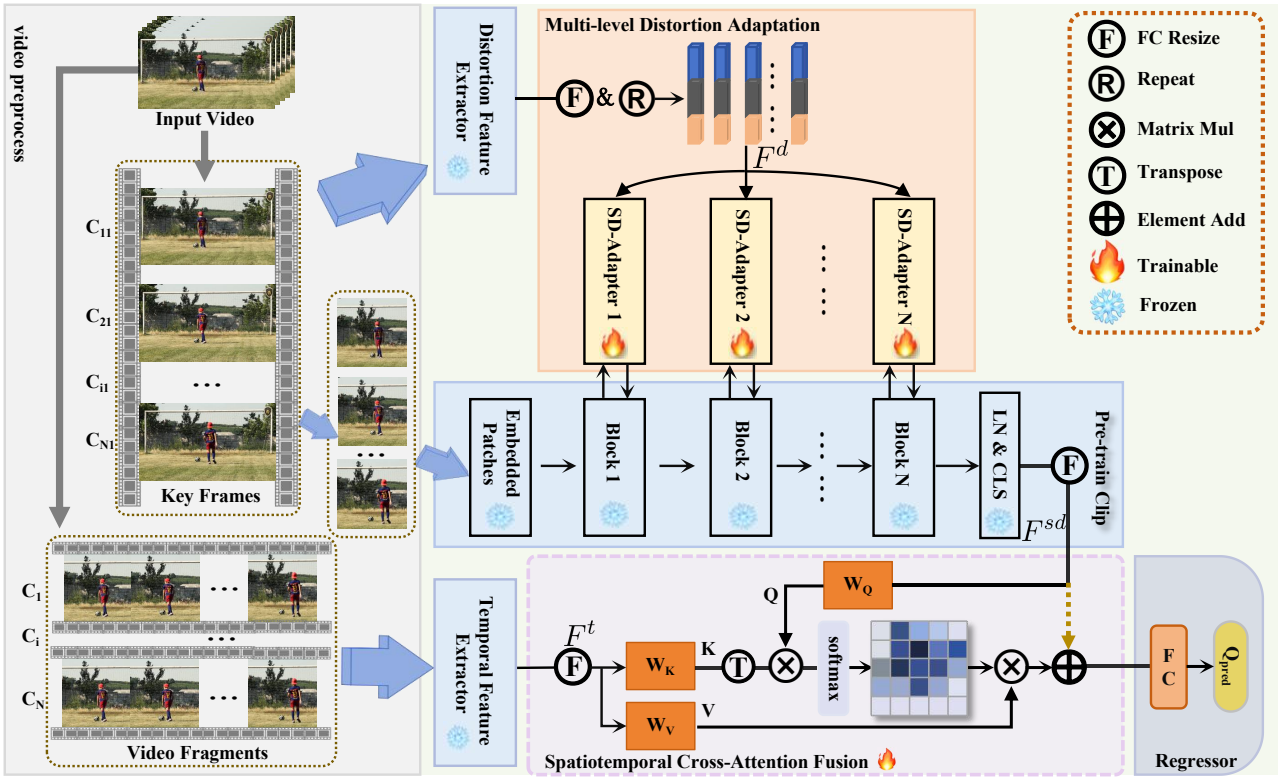

Figure 6. Overview of team FoodVQA proposed method.

tor, and a temporal feature extractor. These extractors yield a robust feature extraction to all kinds of videos. Besides, they propose a layer-by-layer fusion strategy that integrates distortion and semantic features, enhancing the Transformer's capacity to capture quality-relevant distortions while preserving its original feature extraction. An adapter module ensures distortion information does not disrupt the Transformer's inherent mechanisms. By progressively incorporating distortion at each layer, the attention mechanism refines semantic relationships and adapts to distortions, enabling adaptive feature refinement. Meanwhile, the proposed method applies cross-attention to fuse semantic-distortion features F_i^{sd} with temporal information F_i^t. Previous approaches often concatenate spatiotemporal features or use 3D methods (e.g., 3D CNNs or 3D Swin Transformer) directly, failing to capture the potential interactions between them. Finally, the spatiotemporal representation F_i^{sd} is then passed through a regressor consisting of multiple MLP layers to obtain a prediction score q_i for each video fragment. The final predicted quality score Q for the video is calculated as:

$$Q_{pred} = \frac{1}{N}\sum_{i=1}^{N} q_i \, .$$

The differentiable Pearson's Linear Correlation Coefficient (PLCC) loss combined with a rank loss is employed as the objective function. The introduction of rank loss helps the model better distinguish the relative quality levels among videos. The overall loss function is defined as:

$$L = L_{plcc} + \alpha * L_{rank} \, ,$$

where α represents a balancing hyper-parameter, empirically set to 0.3 during training.

In experiments, they pre-trained on LSVQ, containing 39, 000 real-world distorted videos and 117, 000 space-time localized video patches ('v-patches'), and 5.5M human perceptual quality annotations. The pre-training weights are saved, and loaded for training each dimension, then the weights for each dimension are saved, for a total of 6 dimensions. They finally get the averaged main score of 0.8162.

1.2. AI Generated Video Track

1.2.1 SLCV

Team SLCV is the final winner of the AI generated video track. They propose temporal pyramid sampling , as shown Figure 8, to address the unique challenges posed by AI-generated videos in quality assessment. Unlike user generated video, the quality assessment of these AI generated videos primarily focuses on two core aspects: the smooth-

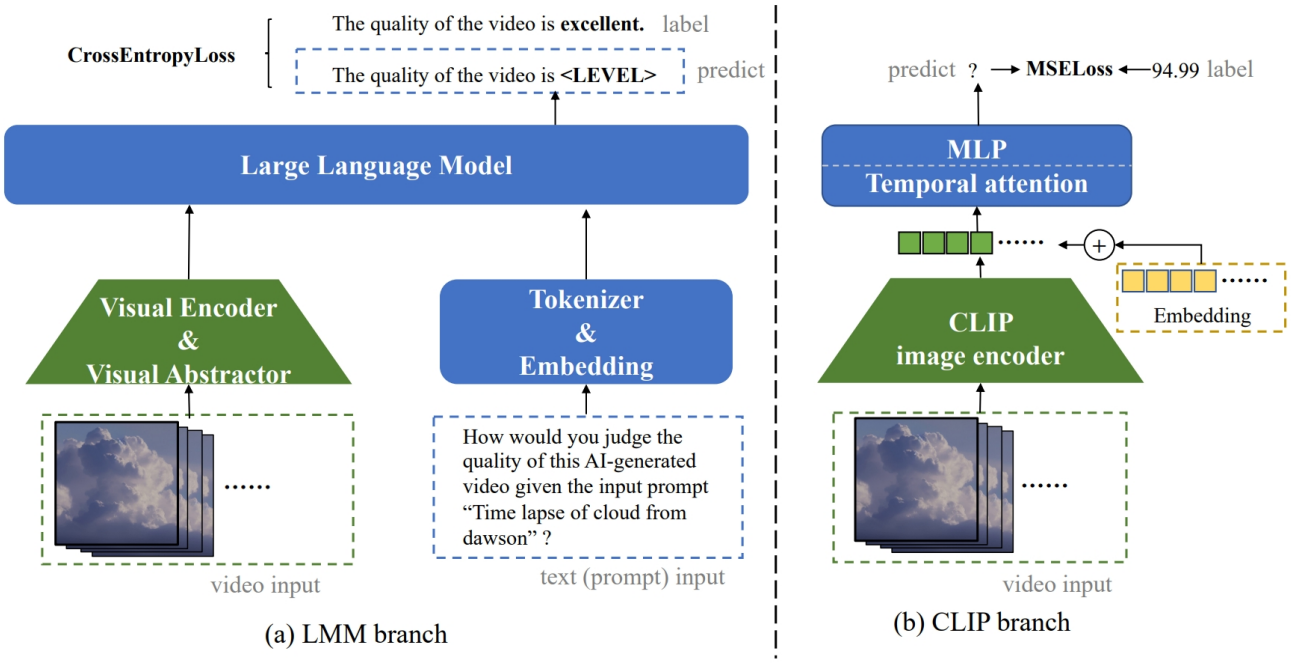

Figure 7. The overview of team Kwai-kaa proposed LMM and CLIP branches.

ness of object motion and the authenticity of the content. To effectively capture these critical metrics, the team design the temporal pyramid sampling method to capture the dynamic characteristics of videos at multiple temporal resolutions. This is achieved by performing multi-scale frame interval sampling at varying frequencies. The original video is sampled at different frame rates and lengths, generating multiple subsets of data with diverse temporal granularities. Each subset is then used to independently train the model, enabling it to learn distinct motion smoothness and content authenticity features at different temporal scales.

During training, they evaluate the performance under varying temporal resolutions and model sizes. First, they test the performance of InternVL2.5 4B [8], InternVL2.5 8B [8], and InternVL2.5 26B [8] using a fixed sampling frame rate of 2 frames per second(fps). Their results demonstrate that larger models consistently outperform smaller ones, indicating a positive correlation between model size and performance. Then, to assess the impact of different temporal resolutions, they use the InternVL2.5 26B model to resample the original video at varying frame rates. The results reveal consistent performance across different frame rates. Finally, they train three assessment models on InternVL2.5 26B at three different sampling frequencies: 2.5 fps, 3 fps, and 4 fps, respectively. The whole trainning phase uses 8 NVIDIA A100 (80G) GPUs with the LowRank Adaptation (LoRA) method for two epochs.

During inferring, they average the results from three models, and observe a notable improvement of 2% in overall performance, achieving a final score of 0.6645.

1.2.2 CUC-IMC

Team CUC-IMC [32] wins second place in the AI generated video track. They propose an LLM-based AI-Generated Video (AGV) visual quality assessment method, which consists of multibranch encoders and an LLaMA2-based [44] decoder, as illustrated in Figure 9. Specifically, they design a multi-branch encoding architecture to comprehensively characterize AGV visual quality by decoupling it into three complementary dimensions: (1) The technical quality encoder employs Swin-3D Transformer [55] to capture technical artifacts such as motion blur, noise, and jitter; (2) The motion quality encoder quantifies video motion characteristics including naturalness, smoothness, and dynamic intensity through a SlowFast [13] network; and (3) The semantic encoder utilizes a BLIP-based [22] visual backbone to represent holistic content semantics. Through a specially designed multimodal prompt engineering framework, they align the features extracted from the multi-branch encoders with the LLM's reasoning space. Guided by feature mapping and semantic anchors, the LLM establishes cross-dimensional correlations among these features. Combined with LoRA fine-tuning technique, this design enables superior adaptation of the LLM for quality assessment tasks.

During the training phase, they employ the Adam optimizer with an initial learning rate of 1e-5 and a decay rate of 0.05. The learning rate is dynamically adjusted using the

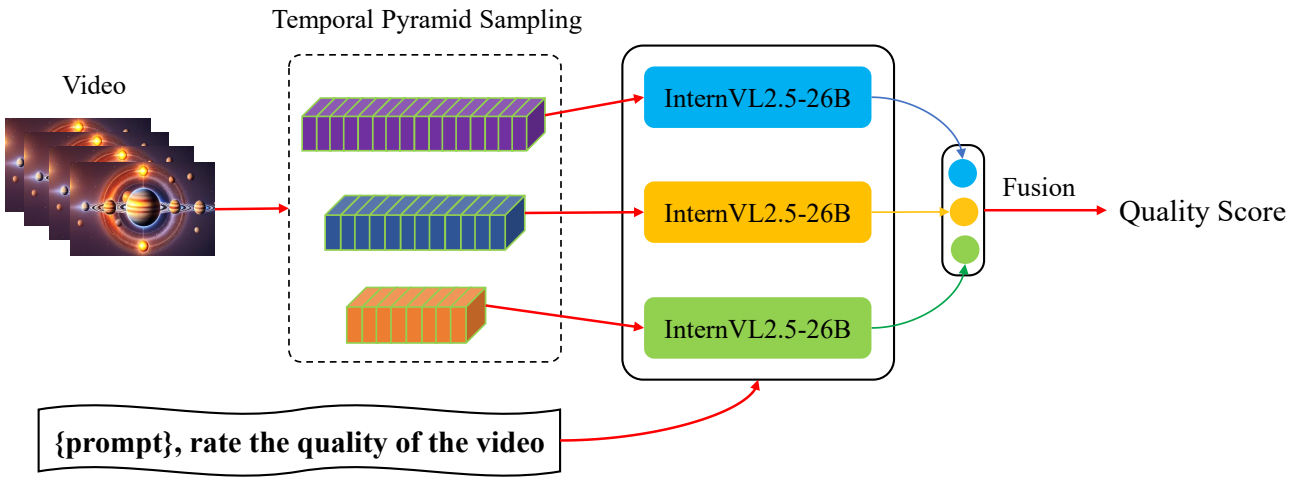

Figure 8. The overview of team SLCV proposed method in track 2.

LambdaLR scheduler, with the training process spanning 10 epochs including 2.5 warmup epochs. In the video frame sampling stage, each video is processed through two distinct sampling strategies: Randomly sample 8 frames with a stride of 2 frames between consecutive samples. This sparse sampling captures representative visual semantics while reducing redundancy. Uniformly sample 32 frames with temporal continuity preserved. Dense sampling enables precise motion pattern analysis through the SlowFast architecture. And during the testing phase, they maintain the same frame sampling strategy as used in the training stage. Their all stages are consistently completed using an NVIDIA A800 GPU, with an inference speed of 1.91 seconds per video during the final testing. Their final score is 0.6310.

1.2.3 opdai

Team opdai wins third place in the AI generated video track. They propose to evaluate the video quality of AIGVs by using two models. They train two separate models based on T2VQA [20] and SAMA, respectively. The T2VQA model was trained with the addition of LoRA training, while the SAMA model was trained using adversarial weight perturbations. Then they use the fused results of the two models as the final result. Their specific training configuration is as follows: the Sama model using 8 $\times$ 32G NVIDIA V100 GPU, with a batch size of 144 (18 $\times$ 8), a learning rate of 1e-4, and trained for 10 epochs. Finally, they get the score of 0.5903.

1.2.4 Magnolia

Team Magnolia designs a cocktail model , as shown in Figure 10, incorporating multiple feature encoders for AI-Generated video quality assessment (AIG-VQA). It consists of four visual encoders (Siglip2-ViT [47], CLIP-ConvNext [26], Swin2 [24], and VideoSwin [27]) and two textual encoders (Siglip2 [47] and CLIP [9]). The visual encoders Siglip-ViT and CLIP-ConvNext are responsible for capturing the semantic content of videos, while Swin2 and VideoSwin focus on encoding textural and motion-related information. To achieve this, Siglip-ViT and CLIP-ConvNext process resized video frames to represent the overall view of the content, whereas Swin2 and VideoSwin operate on cropped frames or video clips to extract features of fine details. The two textual encoders take the prompts used to generate videos as input. Each visual/textual encoder is followed by an adapter (a linear layer) that projects its extracted features into a quality perception space. Notably, textual features are concatenated with their corresponding visual features before projection.

Each visual/textual feature outputted from adapters are mapped into a quality score. They use the PLCC and rank losses in FAST-VQA [52] as the quality prediction loss for each predicted scores. Simultaneously, all the features are concatenated and fed into the quality network (which consists two linear layers) for final quality prediction. The same loss is used. In addition, they also adopted an auxiliary classification task to boost the quality perception. Specifically, they roughly categorized the training videos according to their resolution and fps. Then the visual features are mapped into specific categories. The cross entropy loss is used as the auxiliary loss.

During training phase, they train the model for 10 epochs. During the first 5 epochs, they freeze the parameters of the visual and textual encoders, training only the adapters, fully connected (FC) layers, and the quality network. The batch size is set to 16, and the learning rate is 3e-4. In the last 5 epochs, they finetune the entire model ex-

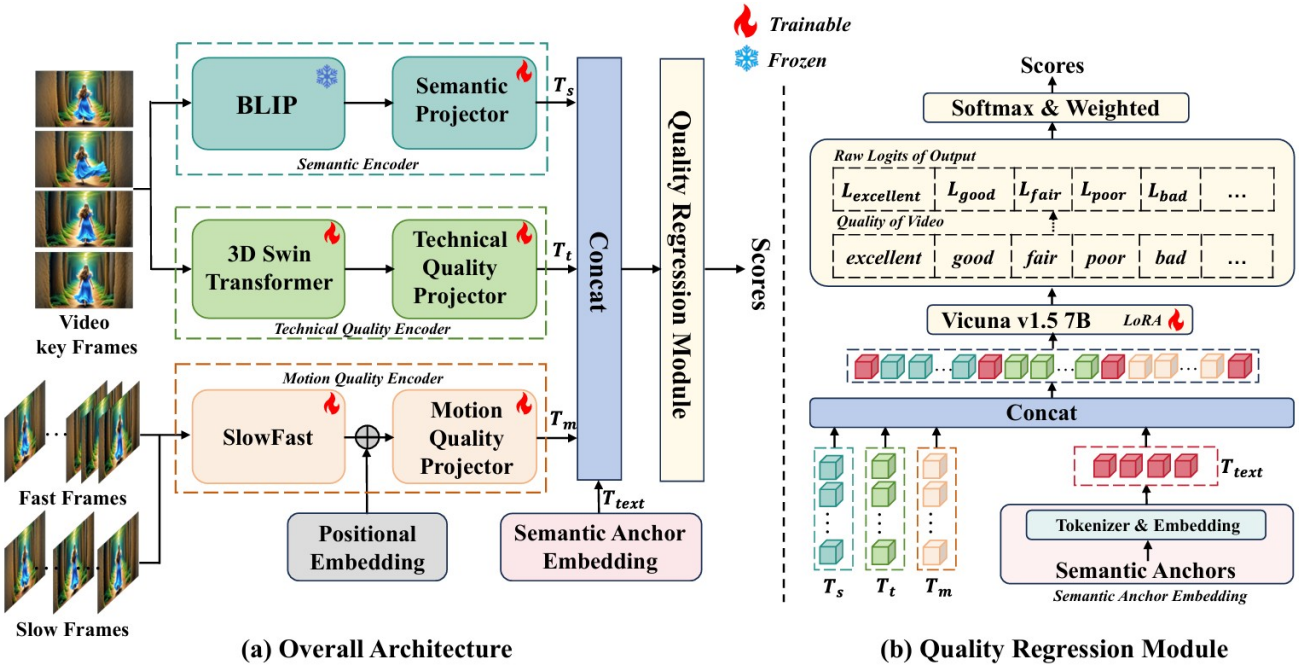

Figure 9. The overview of team CUC-IMC proposed method in track 2.

cept for the textual encoders, using the same batch size but with a reduced learning rate of 3e-5. AdamW is used as the optimizer. They resize and crop the video according to the resolution requirements of different backbones. The entire training process is completed on a 32G vGPU.

During test phase, the team randomly selects eight consecutive frames from each video. The testing process is repeated ten times, and the final quality prediction for the video is obtained by averaging the results of these ten runs. Their final score is 0.5889.

1.2.5 AIGC-VQA

Team AIGC-VQA proposes a simple but effective method, as shown in Figure 11. They use the pre-trained KSVQE [28] model, freeze the backbone, and simply fine-tune the model head on the competition training dataset, using PLCC loss and distortion contrast loss. This model uses 3D Swin Transformer [55] as the backbone, CLIP [33] to extract semantic information, and CONTRIQUE [30] to extract spatial distortion. Finally, their score is 0.5606.

1.2.6 SJTU-MOE-AI

Team SJTU-MOE-AI proposes a quality assessment method using multi-modal vision-language models trained with both contrastive [33] and autoregressive objectives as the base quality evaluator, as shown in Figure 12. Specifically, they employ SigLIP2 [47] and a variant of Q-Align [53]. Given a video, the base quality evaluators generate frame-level quality predictions for eight key frames uniformly sampled from the entire frame sequence. The final video-level quality score y_b is then obtained by averaging these frame-level quality predictions. For SigLIP-2, they first compute the cosine similarities between visual embedding and textual embedding derived from five textual templates: "a photo with $\{q\}$ quality, which matches prompt". where $q \in \{bad, poor, fair, good, perfect\}$, where $\{q\}$ corresponds to the Likert-scale of five quality levels and prompt is the textual prompt used to generate the video. Following [59], they apply a softmax function to the cosine similarity logits to obtain a quality distribution over the five levels, which they then convert into a scalar quality score via weighted summation. As for Q-Align, they adhere to its default setup for computing quality scores, except that they append the video prompt to the question in the conversation template.

Then inspired by [50], they equip the quality evaluator a spatial rectifier and a temporal rectifier, which are implemented by a truncated ResNet-18 [17] and a SlowFast model [14], respectively. The spatial rectifier captures distortions caused by spatial re-sampling, while the temporal rectifier addresses artifacts resulting from motion anomalies. To evaluate text-video alignment, we introduce an alignment rectifier based on FGA-BLIP2 [15] that bridges vision and language modals. The representations from all three rectifiers are processed through separate multi-layer perceptrons (MLPs). As shown in 12, each rectifier pro-

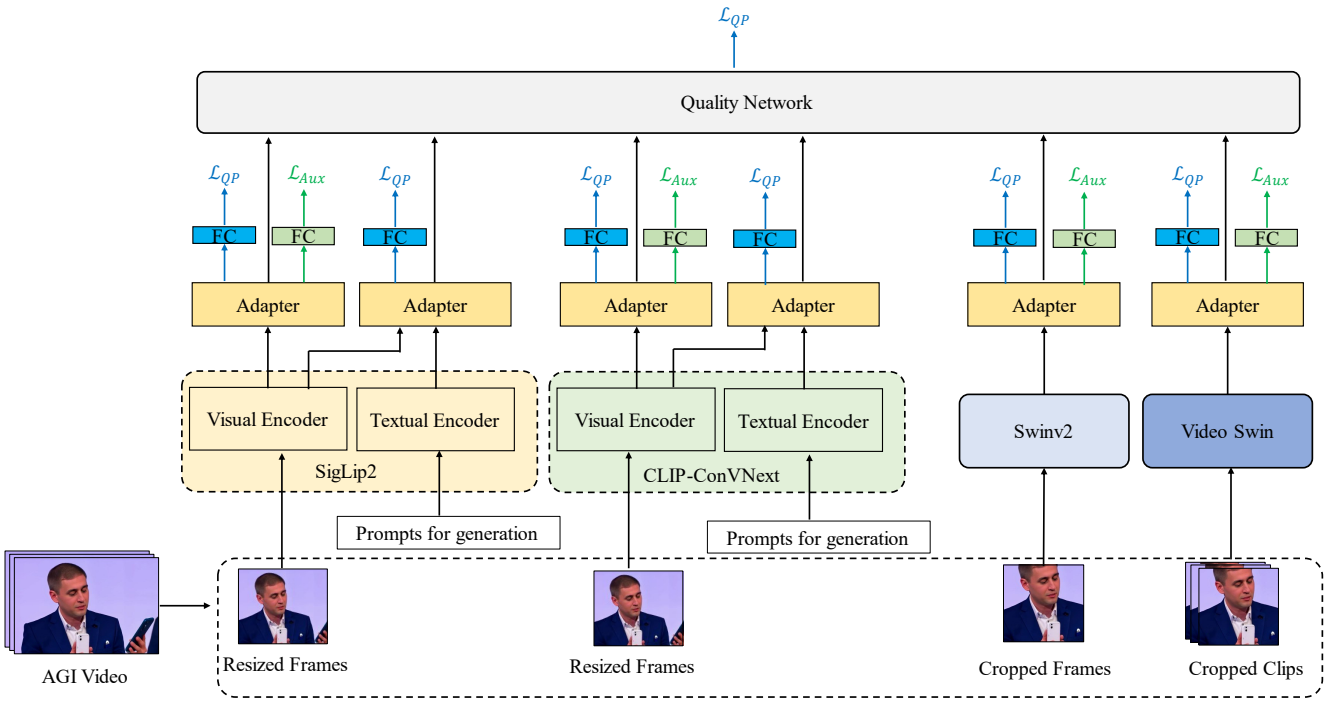

Figure 10. The overview of team Magnolia proposed method in track 2.

duces a tuple of scale (α) and shift (β) parameters, which are applied to rectify the base quality predictions y_b as follows:

$$y = \sqrt[3]{\alpha_s \alpha_t \alpha_a} y_b + \frac{\beta_s + \beta_t + \beta_a}{3} \tag{8}$$

They train their model for 10 epochs, minimizing the PLCC loss function using the Adam optimizer with a batch size of 8. The initial learning rate is set to 1e-5. For SigLIP2, "siglip2-base-patch16-naflex" is used, and the maximum number of patches is set to 768. For the dataset in this track, they sample 8 key frames. For the Q-Align baseline model, they train for 4 hours using four A800 GPUs, while their model requires 25 hours of training with a 4090 GPU. SigLIP2 is trained for 20 hours using a single A800 GPU. They finally get the score of 0.5463.

1.3. Talking Head Track

1.3.1 QA Team

The QA Team [39] thoroughly considered both the visual and auditory aspects of TH (Talking Head), proposing a novel NR video quality assessment model based on multimodal feature representations.As shown in Fig. 13. Their method can be divided into four modules:the spatial feature extraction module,the temporal feature extraction module,the audio feature extraction module,and the audio-visual fusion module.

The types of visual distortions in videos can be roughly divided into two categories: spatial distortion[62] and motion distortion. First, Talking Head videos are split into clips for spatial and temporal feature extraction.Whole clip is utilized for temporal feature extraction with a fixed pretrained 3D-CNN backbone SlowFast[14].The first frame of each clip is used for spatial feature extraction.The spatial feature extraction module utilizes an efficient channel attention module ECA-Net[48], to effectively achieve cross-channel interaction,and then utilize the SwinTransformer-tiny[25] to extract visual features from the first frame.

For audio feature extraction, they first align the audio to the visual frames according to the timeline and utilize four extraction techniques to extract audio features[4], including the chromagram, CQT, MFCC, and GFCC. These features provide various characteristics and compensate for each other.Subsequently, they stack these features to generate 4 channels of time-frequency audio features and feed them into a separable convolution network to obtain more discriminative and distinguishable audio features.The separable convolution network consists of three blocks (frequency block, time block, and fusion block), which can handle time and frequency domain characteristics for audio feature representation and output more distinct and complimentary audio features. Each block consists of Conv2D-Conv2D-Conv2D-BatchNorm-Maxpool with different numbers of

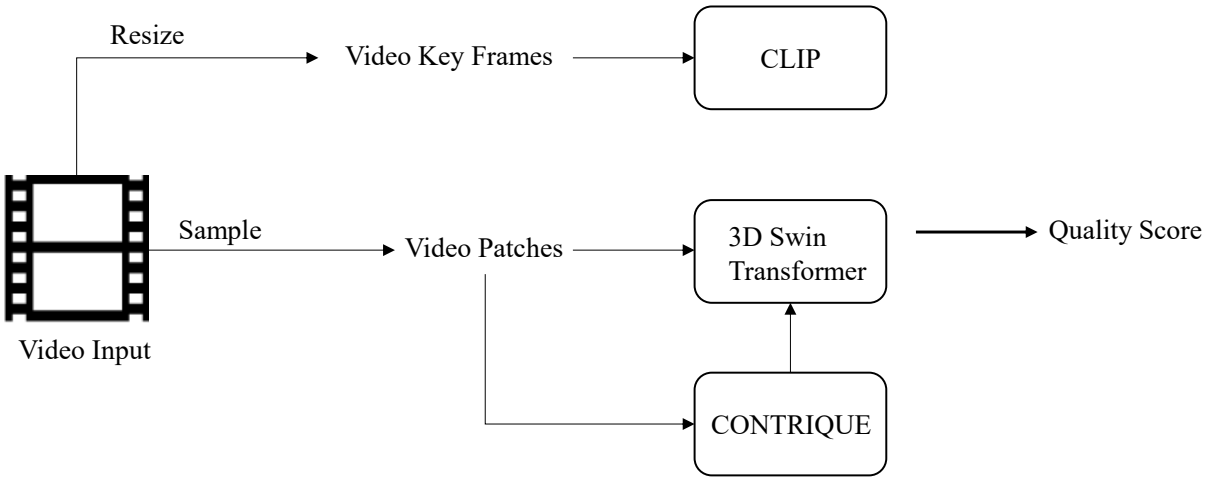

Figure 11. The overview of team AIGC-VQA proposed method in track 2.

kernels and kernel sizes. Specifically, the kernel sizes of the frequency block and time block are $1 \times m$ and $n \times 1$ in the frequency domain and time domain, respectively, which employ spatially separable convolutions and reduce the parameters drastically.Then learns the temporal information through Bi-LSTM, which has a good effect on processing timing signals. Bi-LSTM, a particular LSTM, receives information in both forward and backward directions at the same time, which makes the characteristics of the time sequence richer, because each signal possesses context message from the past and future.Finally, they fuse the features into a final quality score using the FC layers.

The videos are divided into 1s clips.they employ 6 clips from each video with cyclic sampling.If a video has less than 6 clips,the existing clips are expanded with cyclic sampling until 6 clips are selected.For videos lasting for more than 6 seconds,the first 6 clips are used.The Swin Transformer is utilized as the spatial feature extractor and patches with a resolution of 3*224*224 are cropped as input.The SlowFast is utilized as the temporal feature extractor and the clips are resized to 224*224 for training.

The Adam optimizer is utilized, with an initial learning rate of 1e-5. The default number of epochs and batch size are set as 50 and 1, using one RTX3080 10GB GPU for talking head dataset. The total training duration was 15 hours. Specifically, after every 2 epochs, the learning rate is multiplied by 0.95, ensuring that the learning rate progressively diminishes over time.

1.3.2 MediaForensics

The MediaForensics team proposed integrating image and video pre-trained models for enhanced multiframe video assessment. As shown in Fig. 14, The method ensemble consists of two types of models: The first type of models involve using the image pre-trained model "eva02 large patch14 448. mim m38m ft in22k in1k" [12] as the backbone, combined with a feature regression layer (NeXtVLAD [23]), to build a multi-frame video assessment model (training two separate models with 4 frames and 8 frames, respectively). The Second type of models involves using the visual encoder part of the video pre-trained model "microsoft/xclip-large-patch14-16-frames" [31] to construct a multi-frame video quality analysis model (training two separate models with 16 frames and 32 frames, respectively). Finally, the average of the prediction scores from the four models is calculated to obtain the final prediction score.

During training, they firstly train the model for 27 epochs using 90% of the training set, and the remaining 10% of the training set is used as a validation set. Subsequently, they finetune the model for another 7 epochs using the entire training set.

1.3.3 AutoHome AIGC

The AutoHome AIGC team developed an enhanced solution based on SimpleVQA [40] for assessing the quality of generated Talking Head videos. The team's methodology revolved around extracting both spatial and motion features to improve video quality assessment. Spatial features were obtained using ResNet50 [16]or RegNetY-8G [36], while motion features were extracted via a pre-trained SlowFast action recognition network. Multi-resolution processing was enhanced by concatenating spatial features from different backbone stages along with motion features. The final video scores were computed using a Multi-Layer Perceptron (MLP) combined with Mean Pooling.

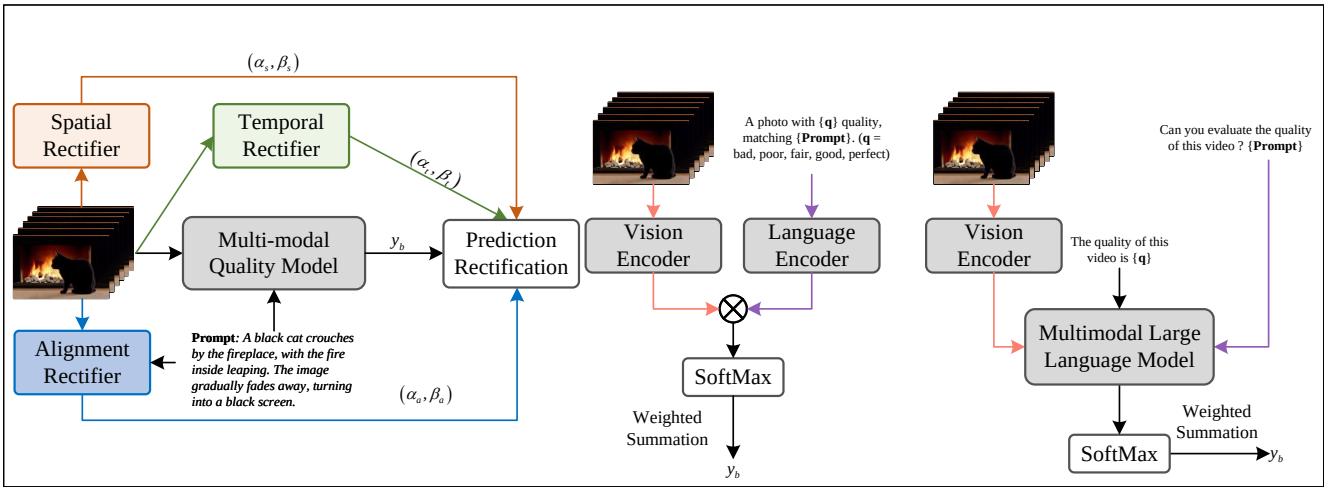

Figure 12. The overview of team SJTU-MOE-AI proposed method in track 2.

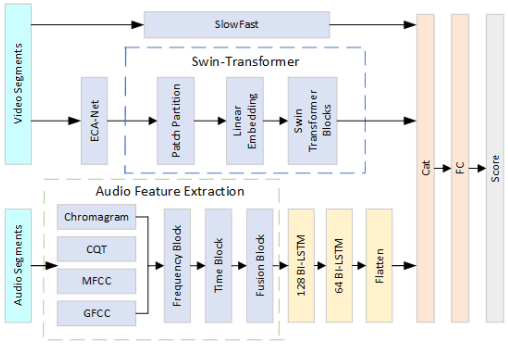

Figure 13. The framework of the QA team method.

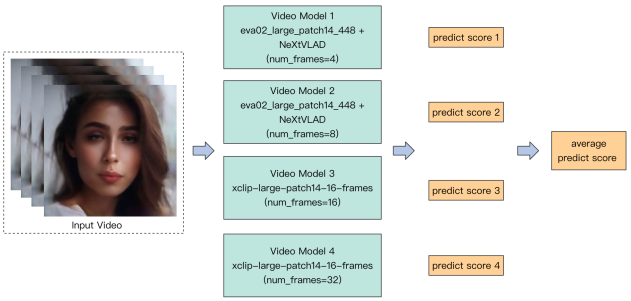

Figure 14. The framework of the MediaForensics team method.

The team implemented online performance fitting during the testing phase. While the original SimpleVQA model applied performance fitting only in training, extending this approach to inference further optimized the model's predictions, resulting in a MainScore increased to 0.799761. To further enhance feature extraction, the team replaced the original ResNet50 backbone with RegNetY-8G, a more powerful model with superior Top-1 accuracy on ImageNet. This modification boosted the MainScore to 0.801045. The MainScore improvements on the validation dataset are shown in Table 2.

	MainScore
BVQA Baseline	0.26
Fintune on training dataset	0.780976
Increase Learning Rate to 1e-4	0.799159
On-line performance fit	0.799761
RegNetY-8G Backbone	0.801045

Table 2. MainScore Improvements on Validation Dataset

Beyond these architectural changes, the team also optimized the testing process to improve efficiency. Offline-extracted frames and SlowFast features were used instead of real-time computation, reducing computational costs while maintaining accuracy. The incorporation of online performance fitting during inference further refined predictions, which led to the highest test-phase MainScore of 0.807383. The MainScore Improvements on Validation Dataset shows in Table 3

	MainScore
ResNet50	0.803407
RegNetY-8G	0.807383

Table 3. MainScore on Test Dataset

For training, RegNetY-8G, pre-trained on ImageNet, was used as the initial weight. The training strategy remained consistent with SimpleVQA, with the exception of an increased learning rate. The team utilized an Adam optimizer with a StepLR scheduler, applying a learning rate decay factor of 0.9 every two epochs. During data preprocess-

ing, one frame per second was extracted from each video, and for videos shorter than 8 seconds, the last frame was repeated to maintain consistency. SlowFast was also used for motion feature extraction, and all frames were resized to 224×224 to reduce computational costs.These optimizations allowed AutoHome AIGC to achieve 1st place on the test leaderboard, demonstrating significant improvements over the baseline SimpleVQA model.

1.3.4 USTA-AC

The USTC-AC team proposes a dual-branch network for video quality assessment. To address the domain gap between 2D and 3D video clips, they split the dataset into two parts. The videos are processed at 25 fps and cropped to 512x512 portrait format, with face regions detected using MediaPipe.

The team built a dual-branch network to extract the frame feature and the face image feature, where they use ViT [11] and SwinTransformer-b [25]. In the first, they use the DSL-FIQA [6] model as shown in the Figure 15 to extract features from randomly sampled frames and crop out the face region, applying the SwinTransformer-b to extract face-specific features. Additionally, the team uses to extract artifact features.

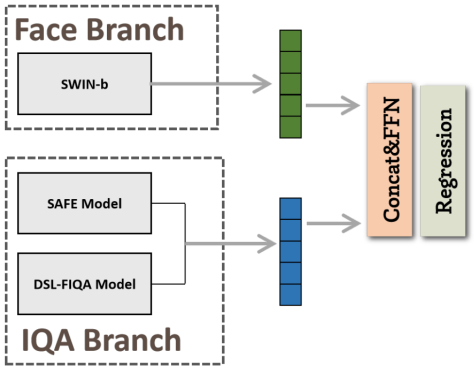

Figure 15. The framework of the proposed solution.

These features are then concatenated into a final feature representation, which is passed through a two-layer MLP network to predict the final quality score. The model is pretrained on the CGFIQA-40k dataset [6]. During training, the team randomly samples frames to compute MOS, while in the testing phase, they sample 10 frames and calculate their mean score as the final MOS.

1.3.5 SJTU-MOE-AI

The SJTU-MOE-AI team proposed a modular multi-Modal quality assessment for talking head videos via spatial and temporal rectification. The model is adapted from [50] with

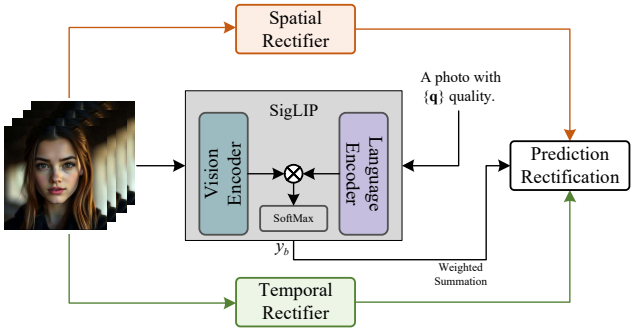

Figure 16. The schematic of SJTU-MOE-AI team proposed.

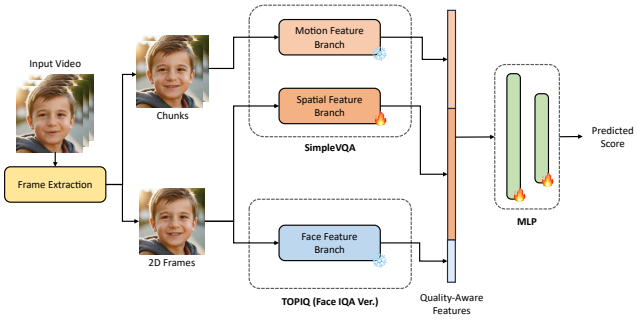

Figure 17. Overall Pipeline of the Face-IQA-Enhanced Evaluator (FIEE) proposed by the FocusQ team.

two modifications (see Fig. 16). First, we use SigLIP-base-patch16-512 [58] (rather than CLIP-ViT-base-patch16 [34] in the original implementation) the base quality predictor. Second, instead of introducing an additional multi-layer perceptron (MLP) to compute quality predictions on top of vision representation, we follow [60] to compute the cosine similarities between the visual embedding and textual embedding derived from five textual templates: "a photo with {q} quality.", where q ∈ {"bad", "poor", "fair", "good", "perfect"}, corresponding to the Likert-scale of five quality levels. We then apply a Softmax function to the cosine similarity logits to obtain a quality distribution over the five levels, which are converted into a scalar quality score via weighted summation. We also enhance the base quality predictor with a spatial rectifier and a temporal rectifier as in [50], which are based on a truncated ResNet-18 [18] and a SlowFast model [14], respectively. The representations from both rectifiers are processed through separate multi-layer perceptrons (MLPs), producing a tuple of scale (α) and shift (β) parameters, which are applied to rectify the base quality predictions y_b as follows:

$$y = \sqrt{\alpha_s \alpha_t} y_b + \frac{\beta_s + \beta_t}{2} \tag{9}$$

1.3.6 FocusQ

The FocusQ team proposes Face-IQA-Enhanced Evaluator (FIEE), an innovative framework for the quality assessment of talking head videos. Unlike other general video generation [37, 38], this track focuses on the quality assessment of facial local details. Specifically, FIEE first extracts key frames for each videos. Then, FIEE uses two key components to obtain quality-aware features: (i) a TOPIQ model [5] pretrained with face IQA dataset (GFIQA) [7] to extract static face-specific features, and (ii) a SimpleVQA model [41] learning to extract dynamic motion and spatial features. Finally, FIEE fuses these features and adopts an MLP to predict the final MOS scores. Experimental results have shown that our method can achieve competitive performance. Specifically, the overall pipeline of FIEE is illustrated in Figure 17. To improve the training/inference efficiency, the FocusQ team follows SimpleVQA [41] to extract chunks and 2D frames from each video. As a result, the image size of chunks and 2D frames is $3 \times 224 \times 224$ and $3 \times 448 \times 448$, respectively. The extracted chunks are processed by the motion feature branch, while the 2D frames are fed into both the spatial feature branch and the face feature branch. Through these branches, FIEE obtains three types of 1D features with lengths of 2304, 7168, and 256, respectively. To further improve efficiency, FIEE pre-extracts motion features using the motion feature branch with a SlowFast-R50 framework [14, 18], following the methodology of SimpleVQA [41]. Next, FIEE concatenates these features to formulate the quality-aware features, which are subsequently fed into a two-layer MLP with the number of hidden nodes set to 128. Finally, the MLP is expected to predict the MOS score aligned with human perception.

In terms of training strategy, FIEE employs L1 loss, Pearson linear correlation coefficient (PLCC) loss, and Spearman rank-order correlation coefficients (SROCC) loss, with their loss weight all set to 1. FIEE only trains the spatial feature branch with ResNet-50 [18] as the backbone and the MLP and keeps other components frozen. Note that for the face feature branch, FIEE also adopt ResNet-50 [18] as the backbone. Moreover, the batch size is set to 8. FIEE adopts AdamW as the optimizer, with the learning rate set to 1e-5 and the weight decay ratio set to 0.9. The implementation is based on Python 3.10 and PyTorch 2.6, and all experiments are conducted on an NVIDIA H100 GPU. Importantly, the FocusQ team did not use any additional external datasets for training the model.

1.3.7 NJST-KMG

The NJUST-KMG team proposes a Multi-Granularity Fusion strategy to enhance video quality assessment for talking head videos, addressing the challenge of integrating both qualitative and quantitative evaluations to improve prediction robustness. Specifically, the method employs two distinct large language models: one trained for 5-level classification (bad, poor, fair, good, excellent) and another for 10-level regression (integer scores 0–9). During inference, outputs from both models are fused to leverage complementary insights from different granularities, enhancing the overall assessment accuracy.

The team implemented this approach using pre-trained vision-language models QwenVL2.5-7B [2] and MiniCPM-V-2_6 [56], which were fine-tuned on competition-specific data formatted as user instructions. Training utilized the Adam optimizer for 2 epochs on 2× NVIDIA RTX 3090 GPUs, completing within 24 hours. No external data or quantization techniques were employed. The fused inference achieved a processing speed of 1.38 videos per second, demonstrating computational efficiency despite the model scale.

Experiments on the competition dataset highlighted the method's advantages, including improved robustness to subjective variations, generalization from pre-trained models, and enhanced accuracy through multi-granularity fusion. The approach outperformed baseline methods without requiring additional training data, validating its applicability in resource-constrained scenarios.

1.3.8 XIDIAN-VQATeam

XIDIAN-VQATeam proposed the Adaptive Talking Head Quality Assessment (ATHQA) method, which is an optimized approach based on short video quality assessment techniques, specifically designed for the NTIRE 2025 talking head video quality assessment task.An example of ATHQA shows in Figure 18. ATHQA integrates both Content Understanding and Distortion Understanding modules to effectively identify key regions affecting video quality and distinguish various types of distortions, thereby accurately predicting subjective video quality. The model employs a Fragment Sampling Strategy to enhance training efficiency, while Content-adaptive Modulation (CaM) and Distortion-aware Modulation (DaM) improve the model's understanding of content and distortions.

During the training phase, the model leverages CLIP [33] and CONTRIQUE [29] as pre-trained models, with a learning rate set to 3e-5. The loss function is defined as 0.7LMSE + 0.3LRank, and data augmentation strategies include Temporal Reversal, FiveCrop, and Random Frame Sampling (8 frames per video). Additionally, Multi-Stage Training and Multi-Scale Feature Fusion are utilized to enhance the model's adaptability to different distortion patterns.

In the testing phase, the team maintained the same hyperparameter settings and data augmentation strategies as

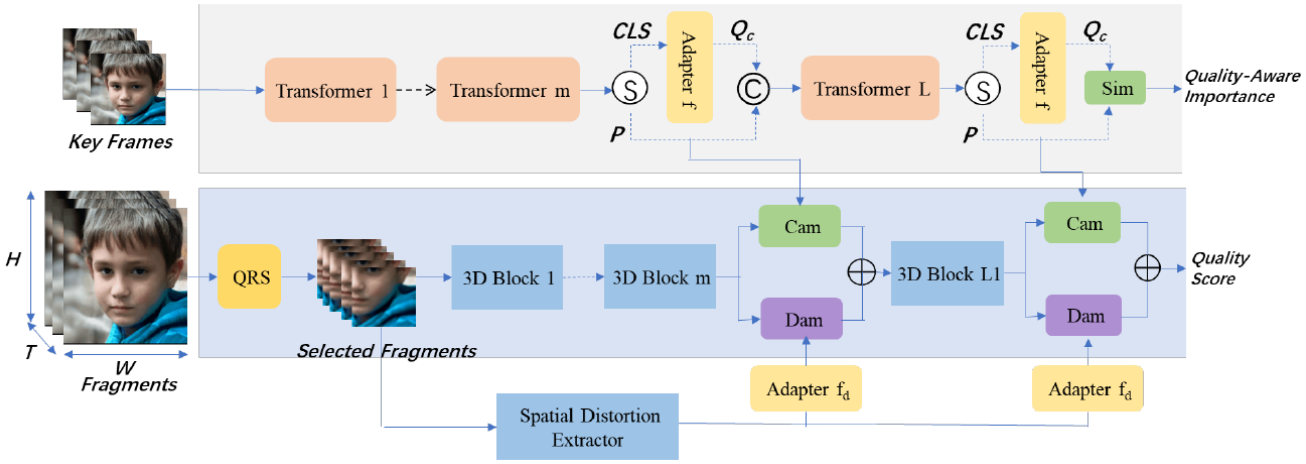

Figure 18. Representative image: Proposed network architecture.

Input res	Training time	Epochs	Extra data	Attention	Quantization	Params (M)	GPU
448×448	34h	100	No	Swin-T	FP16	26	NVIDIA TITAN RTX

Table 4. Key technical parameters

used during training. ATHQA introduces several innovative components, including the Lip-sync Consistency Module [10], which analyzes lip synchronization in talking head videos; Dynamic Motion Artifact Detection [21], based on optical flow to identify motion distortions; and a Hybrid Ensemble of Spatial and Temporal Predictions, which integrates spatial and temporal information to improve evaluation accuracy. Table 4 shows key technical parameters.

Regarding computational complexity, the model consists of 26 million parameters (26M), with an inference time of 0.54 seconds per video on an NVIDIA TITAN RTX. During training, the AdamW optimizer is employed($\beta_1 = 0.9$,$\beta_2 = 0.999$, weight decay $= 0.05$), with a batch size of 4 and a 2.5-epoch warmup. Moreover, the team optimized computational efficiency by incorporating Depth-wise Separable Convolutions and Operator Fusion (e.g., Conv+BN+ReLU) to accelerate inference.

A. NTIRE 2025 Organizers

Title:
NTIRE 2025 Quality Assessment of AI-Generated Content Challenge
Members:
Xiaohong Liu[1] *(xiaohongliu@sjtu.edu.cn)*, Xiongkuo Min[1], Qiang Hu[1], Xiaoyun Zhang[1], Jie Guo[2], Guangtao Zhai[1,2], Shushi Wang[1], Yingjie Zhou[1,2], Lu Liu[1], Jingxin Li[3], Liu Yang[1], Farong Wen[1], Li Xu[2], Yanwei Jiang[1], Xilei Zhu[1], Chunyi Li[1], Zicheng Zhang[1], Huiyu Duan[1], Xiele Wu[1], Yixuan Gao[1], Yuqin Cao[1], Jun Jia[1], Wei Sun[1], Jiezhang Cao[4], Radu Timofte[5,6]
Affiliations:
[1] Shanghai Jiao Tong University, China
[2] Peng Cheng Laboratory, China
[3] China National Information Technology Standardization Committee Multimedia Subcommittee, China
[4] Harvard University, USA
[5] ETH Zürich, Switzerland
[6] University of Würzburg, Germany

B. Teams and Affiliations in User-generated Video Track

SLCV

Title:
Multimodal Large Language Model for Video Quality Assessment
Members:
Baojun Li[1] *(libaojun2@shopline.com)*, Jiamian Huang [1], Dan Luo[1], Tao Liu[1]
Affiliations:
[1] Shopline AI Research

SJTU-MOE-AI

Title:
Multi-Dimensional Quality Assessment for UGC Videos via Modular Multi-Modal Vision-Language Models
Members:

Weixia Zhang[1] *(zwx8981@sjtu.edu.cn)*, Bingkun Zheng[1], , Junlin Chen[1]
Affiliations:
[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

MiVQA

Title:
Multi-dimensional UGC Video Quality Assessment Method Based on RQ-VQA
Members:
Ruikai Zhou[1] *(zhouruikai12@163.com)*, Meiya Chen[1],Yu Wang[1], Hao Jiang[1]
Affiliations:
[1] Xiaomi Technology Co., Ltd.

XGC-Go

Title:
Multi-Dimensional Quality Assessment for UGC Videos via Large Model Features
Members:
Xiantao Li[1]*(631719224@qq.com)*, Yuxiang Jiang[1], Jun Tang[1]
Affiliations:
[1] JD.com

FoodVQA

Title:
Quality Assessment of UGC Videos via Multi-level Distortion Adaptation and Spatiotemporal Cross-Attention Fusion
Members:
Yimeng Zhao[1] *(s230231169@stu.cqupt.edu.cn)*, Bo Hu[1]
Affiliations:
[1] Chongqing University of Posts and Telecommunications

C. Teams and Affiliations in AI Generated Video Track

SLCV

Title:
Multimodal Large Language Model for Video Quality Assessment
Members:
Baojun Li[1] *(libaojun2@shopline.com)*, Jiamian Huang[1], Dan Luo[1], Tao Liu[1]
Affiliations:
[1] Shopline AI Research

CUC-IMC

Title:
Towards Holistic Visual Quality Assessment of AI-Generated Videos: A LLM-Based Multi-Dimensional Evaluation Model
Members:
Zelu Qi[1] *(theoneqi2001@cuc.edu.cn)*, Chaoyang Zhang[1], Fei Zhao[1], Ping Shi[1]
Affiliations:
[1] Communication University of China

opdai

Members:
Lingzhi Fu[1] *(gandor@qq.com)*, Heng Cong[1], Shuai He[1], Rongyu Zhang[1], Jiarong He[1]
Affiliations:
[1] netease

Magnolia

Title:
A Cocktail Model for AI Generated Video Quality Assessment
Members:
Zongyao Hu[1] *(zyhu@bit.edu.cn)*
Affiliations:
[1] Beijing Institute of Technology

AIGC-VQA

Members:
Wei Luo[1] *(lw21@mail.ustc.edu.cn)*, Zihao Yu, Fengbin Guan[1], Yiting Lu[1], Xin Li[1], Zhibo Chen[1]
Affiliations:
[1] University of Science and Technology of China

SJTU-MOE-AI

Title:
Multi-Modal Quality Assessment for AI-Generated Videos: Integrating Spatial, Temporal, and Alignment Rectifiers
Members:
Bingkun Zheng[1] *(0713bkhun@sjtu.edu.cn)*, Weixia Zhang[1], Junlin Chen[1]
Affiliations:
[1] Shanghai Jiao Tong University

D. Teams and Affiliations in Talking Head Track

QA team

Title:
Quality Assessment for Talking Head Videos via Multi-modal Feature Representation
Members:
Mengjing Su[1] *(sumj63@163.com)*, Yi Wang[1], Tuo Chen[1], Chunxiao Li[1], Shuaiyu Zhao[1], Jiaxin Wen[1], Chuyi Lin[1], Sitong Liu[1], Ningxin Chu[1], Jing Wan[1], Yu Zhou[1]
Affiliations:
[1] China University of Mining and Technology

MediaForensics

Title:
Integrating Image and Video Pre-trained Models for Enhanced Multiframe Video Assessment
Members:
Baoying Chen[1] *(1900271059@email.szu.edu.cn)*,Jishen Zeng[1], Jiarui Liu[1], Xianjin Liu[1]
Affiliations:
[1] Alibaba Group

AutoHome AIGC

Title:
Generated Talking Head Video Quality Assessment based on SimpleVQA
Members:
Xin Chen[1] *(chenxin14069@autohome.com.cn)*, Lanzhi Zhou[2], Hangyu Li[1], You Han[1], Bibo Xiang[1]
Affiliations:
[1] AutoHome Inc
[2] Beihang University

USTC-AC

Title:
Video Quality Assessment for Talking Head Based on Dual-branch Network
Members:
Zhenjie Liu[1] *(liuzhenjie@mail.ustc.edu.cn)*,Jianzhang Lu[1], Jialin Gui[1], Renjie Lu[1], Shangfei Wang[1]
Affiliations:
[1] University of Science and Technology of China

SJTU-MOE-AI

Title:
Modular Multi-Modal Quality Assessment for Talking Head Videos via Spatial and Temporal Rectification
Members:
Junlin Chen[1] *(chenjunlin233@sjtu.edu.cn)*, Weixia Zhang[1], Bingkun Zheng[1]
Affiliations:
[1] Shanghai Jiao Tong University

FocusQ

Title:
Face-IQA-Enhanced Evaluator for Talking Head Video
Members:
Donghao Zhou[1] *(dhzhou@link.cuhk.edu.hk)*, Jingyu Lin[2], Quanjian Song[2], Jiancheng Huang[3], Yufeng Yang[4], Changwei Wang[5]
Affiliations:
[1] The Chinese University of Hong Kong
[2] Monash University
[3] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[4] Southern University of Science and Technology
[5] Shandong Computer Science Center

NJUST KMG

Title:
Multi-Granularity Fusion for Large Modelbased Video Quality Inference
Members:
Shupeng Zhong[1] *(zspnjlgdx@gmail.com)*, Yang Yang[1]
Affiliations:
[1] Nanjing University of Science and Technology

XIDIAN-VQATeam

Title:
Adaptive Talking Head Quality Assessment
Members:
Lihuo He[1] *(lhhe@mail.xidian.edu.cn)*, Jia Liu[1], Yuting Xing[1], Tida Fang[1], Yuchun Jin[1]
Affiliations:
[1] School of Electronic Engineering, Xidian University

References

[1] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 189–198, 2024. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen

Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 15

[3] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024. 1

[4] Yuqin Cao, Xiongkuo Min, Wei Sun, and Guangtao Zhai. Subjective and objective audio-visual quality assessment for user generated content. *IEEE Transactions on Image Processing*, 32:3847–3861, 2023. 11

[5] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024. 15

[6] Wei-Ting Chen, Gurunandan Krishnan, Qiang Gao, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. Dsl-fiqa: Assessing facial image quality via dual-set degradation learning and landmark-guided transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2931–2941, 2024. 14

[7] Wei-Ting Chen, Gurunandan Krishnan, Qiang Gao, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. Dsl-fiqa: Assessing facial image quality via dual-set degradation learning and landmark-guided transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2931–2941, 2024. 15

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, and Lewei Lu. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 8

[9] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 9

[10] Phil Cryer. https://github.com/philcryer/lipsync, 2013. 16

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 14

[12] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. 12

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019. 2, 8

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 4, 10, 11, 14, 15

[15] Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation. In *arXiv preprint arXiv:2412.18150*, 2024. 10

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 12

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 10

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 14, 15

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[20] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dateset and metric for text-to-video quality assessment. *arXiv preprint arXiv:2403.11956*, 2024. 9

[21] Thomas Küstner, Annika Liebgott, Lukas Mauch, Petros Martirosian, Fabian Bamberg, Konstantin Nikolaou, Bin Yang, Fritz Schick, and Sergios Gatidis. Automated reference-free detection of motion artifacts in magnetic resonance images. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 31:243–256, 2018. 16

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 8

[23] Rongcheng Lin, Jing Xiao, and Jianping Fan. Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 12

[24] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie,

Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 9

[25] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4, 11, 14

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 9

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 9

[28] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 10

[29] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using contrastive learning. *arXiv:2110.13266*, 2021. 15

[30] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Kvq: Kwai video quality assessment for short-form videos. In *IEEE Transactions on Image Processing*, 2022. 10

[31] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 12

[32] Zelu Qi, Ping Shi, Chaoyang Zhang, Shuqi Wang, Fei Zhao, Da Pan, and Zefeng Ying. Towards holistic visual quality assessment of ai-generated videos: A llm-based multi-dimensional evaluation model. In *CVPR Workshop*, 2025. 8

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 10, 15

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 14

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[36] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 12

[37] Quanjian Song, Mingbao Lin, Wengyi Zhan, Shuicheng Yan, Liujuan Cao, and Rongrong Ji. Univst: A unified framework for training-free localized video style transfer. *arXiv preprint arXiv:2410.20084*, 2024. 15

[38] Quanjian Song, Zhihang Lin, Zhanpeng Zeng, Ziyue Zhang, Liujuan Cao, and Rongrong Ji. A light and tuning-free method for simulating camera motion in video generation. *arXiv preprint arXiv:2503.06508*, 2025. 15

[39] Mengjing Su, Yi Wang, Tuo Chen, et al. Quality assessment for talking head videos via multi-modal feature representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025. 11

[40] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 856–865, 2022. 1, 12

[41] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 856–865, 2022. 15

[42] Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Zhichao Zhang, Linhan Cao, Qiubo Chen, Xiongkuo Min, Weisi Lin, and Guangtao Zhai. Enhancing blind video quality assessment with rich quality-aware features. *arXiv preprint arXiv:2405.08745*, 2024. 4

[43] Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34:273–286, Jul. 1927. 3

[44] Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. In *arXiv preprint arXiv:2204.14047*, 2022. 8

[45] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. FRank: A ranking method with fidelity loss. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 383–390, 2007. 3

[46] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 2

[47] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. In *arXiv preprint arXiv:2502.14786*, 2025. 9, 10

[48] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11531–11539, 2020. 11

[49] Wen Wen, Mu Li, Yabin Zhang, Yiting Liao, Junlin Li, Li Zhang, and Kede Ma. Modular blind video quality assessment, 2024. 2

[50] Wen Wen, Mu Li, Yabin Zhang, Yiting Liao, Junlin Li, Li Zhang, and Kede Ma. Modular blind video quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 10, 14

[51] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling, 2022. 1

[52] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 538–554. Springer, 2022. 4, 9

[53] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Corresponding Authors: Zhai, Guangtao and Lin, Weisi. 1, 2, 10

[54] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 4

[55] Yuqi Yang, Yuxiao Guo, Jianyu Xiong, Yang Liu, Hao Pan, Pengshuai Wang, Xin Tong, and Baining Guo. Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. In *arXiv preprint arXiv:2304.06906*, 2023. 8, 10

[56] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 15

[57] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq:'patching up'the video quality problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14019–14029, 2021. 4

[58] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 14

[59] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023. 1, 2, 10

[60] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081, 2023. 4, 14

[61] Weixia Zhang, Bingkun Zheng, Junlin Chen, and Zhihua Wang. Multi-dimensional quality assessment for ugc videos via modular multi-modal vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 2

[62] Yu Zhou, Weikang Gong, Yanjing Sun, Leida Li, Jinjian Wu, and Xinbo Gao. Pyramid feature aggregation for hierarchical quality prediction of stitched panoramic images. *IEEE Transactions on Multimedia*, 25:4177–4186, 2023. 11