

Exploring AIGC Video Quality: A Focus on Visual Harmony, Video-Text Consistency and Domain Distribution Gap

Bowen Qu¹Xiaoyu Liang¹Shangkun Sun^{1,2}Wei Gao^{1,2†}¹ School of Electronic and Computer Engineering, Peking University, China² Peng Cheng Laboratory, China

{bowenqu, 2000017789, sunshk}@stu.pku.edu.cn, gaowei262@pku.edu.cn

Abstract

The recent advancements in Text-to-Video Artificial Intelligence Generated Content (AIGC) have been remarkable. Compared with traditional videos, the assessment of AIGC videos encounters various challenges: visual inconsistency that defy common sense, discrepancies between content and the textual prompt, and distribution gap between various generative models, etc. Target at these challenges, in this work, we categorize the assessment of AIGC video quality into three dimensions: visual harmony, video-text consistency, and domain distribution gap. For each dimension, we design specific modules to provide a comprehensive quality assessment of AIGC videos. Furthermore, our research identifies significant variations in visual quality, fluidity, and style among videos generated by different text-to-video models. Predicting the source generative model can make the AIGC video features more discriminative, which enhances the quality assessment performance. The proposed method was used in the **third-place** winner of the NTIRE 2024 Quality Assessment for AI-Generated Content - Track 2 Video, demonstrating its effectiveness.

1. Introduction

In recent years, the emergence of Artificial Intelligence Generated Content (AIGC), including images, texts and videos, has significantly influenced the digital media production landscape. Numerous video generation models based on different technical routes have been de-

[†]Corresponding author. This work was supported by Natural Science Foundation of China (62271013, 62031013), Guangdong Province Pearl River Talent Program High-Caliber Personnel - Elite Youth Talent (2021QN020708), Shenzhen Science and Technology Program (JCYJ20230807120808017), and Sponsored by CAAI-MindSpore Open Fund, developed on OpenI Community (CAAI-XSJLJJ-2023-MindSpore07).

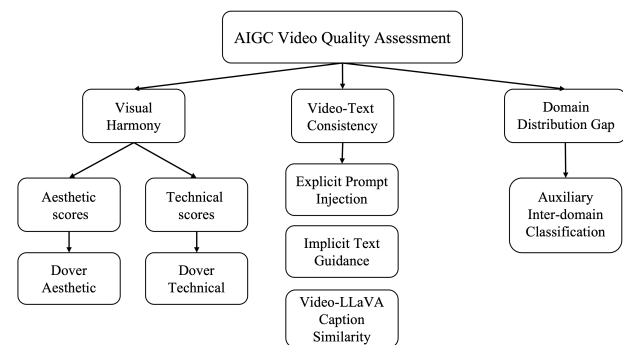


Figure 1. Three Dimensions for AIGC Video Quality Assessment.

veloped, which can be branched into GAN-based [43], autoregressive-based [55] and diffusion-based [48]. As an emerging new video type, AI-generated video needs more comprehensive quality assessment (QA) for the users to get better visual experience.

On the one hand, Inception Score (IS) [37], Fréchet Video Distance (FVD) [41] and Fréchet Inception Distance (FID) [10] are usually employed to evaluate the perceptual quality. On the other hand, CLIPScore [10] is usually used to evaluate the video-text correspondence. However, these metrics are heavily reliant on specific datasets or pretrained models, which is not comprehensive enough.

There are also many advanced Video Quality Assessment (VQA) methods [17, 39, 50, 57] proposed for evaluating natural or UGC (User Generated Content) videos. Nevertheless, due to the limited focus of these methods on issues like visual abnormalities in AIGC videos, their zero-shot effectiveness is not satisfactory. Additionally, there are significant differences in the content distributions generated by different models, posing a strong challenge to the robustness of the framework. Besides, most of them are single-modality based, focus on the technical or aesthetic visual quality. However, AIGC videos generated by text-to-video

model are inherently multimodal entities, each accompanied by a corresponding textual prompt. In reality, these models only take videos as input, which is not sufficient for a comprehensive evaluation due to the lack of understanding of entire textual prompts. Some methods [25, 52] take the text into consideration for more comprehensive assessment, leveraging the strong multi-modality ability of CLIP [32]. They use hard prompts like "a {high, low} quality photo" instead, which is not a good way to evaluate the video-text correspondence.

In this work, we assess the AIGC videos quality from three dimensions: visual harmony, video-text consistency, and domain distribution gap. The overall framework is shown in Fig. 1. As for visual harmony, we refer to DOVER [53] for the aesthetic and technical evaluation of the videos. To measure video-text consistency, we apply explicit prompt injection, implicit text guidance and caption similarity. We inject the corresponding prompts of the videos into the video features using Text2Video Cross Attention Pooling [30]. We also utilize BVQI's implicit text method [52] and jointly optimize the evaluation network using both implicit text and explicit prompts. Building upon this, we utilize the video-text Multimodal Large Language Model (MLLM), Video-LLaVA [23] to generate additional captions for each video segment. We use SentenceBERT [34] to get the embeddings of generated captions and given prompts, and then calculate cosine similarity between them to further optimize the network. To improve the spatio-temporal modeling capability, we also integrate strongly pretrained video backbones by linear-probing like UniformerV2 [21] for model ensemble to get robust results.

Additionally, considering the domain distribution gaps in the videos generated by different text-to-video models, in supervised learning, we predict not only the final mean opinion score (MOS) but also which text-to-video model generated the video. This additional classification aids the model in better understanding video features. Experiments have shown that this significantly enhances the performance of our model.

To summarize, our contributions are three-fold: **1)** We propose a new quality assessment framework for AIGC videos, which we decouple into three aspects: visual harmony, video-text consistency and domain distribution gap. **2)** For each aspect, we design specific modeling methods such as LLM and auxiliary inter-domain classifiers, to propose effective solutions. **3)** Our method shows remarkable improvements on AIGC videos assessment and is used in the **third-place winner** of the NTIRE 2024 Quality Assessment for AI-Generated Content - Track 2 Video [18, 24].

2. Related work

2.1. No-Reference Video Quality Assessment

Quality Assessment has become a crucial task, and significant progress has been made across multiple domains [4, 8, 30, 44, 45, 53] in recent years. Classic NR-VQA methods adopt handcrafted features as evaluation metrics [3, 15, 16, 29, 35, 40]. These methods can extract useful information like color, motion and temporal-spatial features, while keeping low computational complexity. Some other methods [20, 50, 51] extract video features by deep neural networks. DOVER [53] categorizes video quality evaluation into aesthetic and technical dimensions, leveraging distinct backbones for feature extraction. Scores are assigned individually and then aggregated based on a predetermined ratio. By sampling novel fragments as input for deep neural networks and designing a fragment attention network base on Swin Transformers, Fast-VQA [50] efficiently retains quality-related information. StableVQA [17] measures video stability by obtaining optical flow, semantic, and blur features separately. Recently, some novel methods use visual-language pre-training models to evaluate videos. Q-Align [54] utilizes the comprehension abilities of a Multimodal Large Language Model to transform the video quality evaluation task into the generation of discrete quality level words. BVQI [52] introduces the text-language model CLIP to evaluate video quality by assessing the affinity between positive or negative prompts and extracted frames.

2.2. Video Generation and Quality Assessment

Video generation aims to achieve videos with high visual quality and consistent, smooth movements that closely approximate the real world. Image generators based on Generative Adversarial Networks (GANs) [7] have been extended to be effectively used for video generation. However, these methods [36, 43, 47] often encounter issues with mode collapse, leading to lower quality and stability in the generated content. Additionally, some approaches [6, 33, 49, 55] have proposed the use of auto-regressive models to learn the distribution of video data. These methods are capable of generating high-quality and stable videos, but they require a significantly high computational cost. Recent methods [1, 2, 9, 11, 12, 28, 48] in video generation have predominantly focused on diffusion models, achieving very promising results.

Several quantitative evaluation metrics have been proposed for AI-generated videos, mainly focusing on assessing perceptual quality and the video-text correspondence. For perceptual quality, Inception Score (IS) [37], Fréchet Video Distance (FVD) [41] and Fréchet Inception Distance (FID) [10] are usually employed. CLIPScore [10] is mainly used to evaluate the video-text correspondence, leveraging the capabilities of CLIP [32]. However, these methods are

heavily reliant on specific datasets or pretrained models and sensitive to their calculation parameters, such as batch size. Additionally, they do not take the human visual system into consideration, which mean mis-alignment with human perception in assessing AI-generated video.

3. Method

The overview of our proposed method is shown in Fig. 2. Our approach proposes solutions from various aspects including Visual Harmony, Video-Text Consistency, Domain Distribution Gap, etc., which we will elaborate on in the following sections. In summary, it serves as a dual-stream architecture to simultaneously process the AIGC videos and corresponding textual prompts. We think that the inductive bias of this framework matches the multi-modal nature of AIGC videos better. In order to cooperate with the visual harmony model, modules for explicit prompt and implicit text processing are injected, serving as an incremental enhancement for existing VQA methods. Specifically speaking, the video backbone can be initialized by DOVER [53].

3.1. Visual Harmony

Due to the impressive performance of DOVER [53] across multiple Video Quality Assessment datasets [13, 14, 38, 56], we select it for visual harmony modeling. DOVER consists of two branches: the aesthetic branch and the technical branch. These utilize the ConvNext [27] and Swin-Transformer [26] backbone respectively, trained on their DIVIDE-3k dataset [53]. The processing of input videos by these two branches differs. Notably, the technical branch exhibits additional patchifying and fragment sample operations, focusing more on patch-wise features and temporal information. We replace the non-learnable Global Average Pooling (GAP) layer with a learnable Attention Pooling layer (as shown in Fig. 2). Specifically, the output of the GAP is treated as the query, while the spatio-temporally flattened visual tokens are considered as keys and values for cross-attention operations, collectively forming the Attention Pooling module.

3.2. Video-Text Consistency

AIGC videos possess multi-modality nature inherently because of their corresponding textual prompts which is the condition for text-to-video generative model. Due to this characteristic, we propose a multi-modal framework, integrated with explicit textual prompt and implicit text with hard template. These operations enable our model to take textual prompts into consideration and acquire more comprehensive multi-modal features, following text-video interaction. In order to incorporate video-text consistency capabilities into our video quality assessment framework more directly, we also employ the strong and robust video-text Multimodal Large Language Model (MLLM), Video-

LLaVA, to generate captions by the input videos. Then we calculate sentence embedding similarity with the respective textual prompts for a direct zero-shot video-text consistency score.

Explicit Prompt Injection. AIGC videos have inherent multimodal natures from birth. The explicit prompt is a specific condition and guidance for the text-to-video model to generate the corresponding video. Meanwhile, video-text consistency is also an important degree of AIGC video quality assessment. So, we produce a CLIP-like dual-stream architecture with two separate encoders to process the video and prompt respectively. Given a video V and the corresponding explicit prompt T , let $f_{\theta_v}(V)$ represent the video embedding, extracted by the video encoder with parameters θ_v , and let $h_{\theta_t}(T)$ represent the text embedding produced by the text encoder with parameters θ_t . The embedding of $[eot]$ (end of text) token to represent the entire prompt. The process is shown as following, where F_v and F_t refer to the video and text features respectively. F_{eot} refers to the $[eot]$ token embedding, which serves as the global encoding of the whole textual prompt.

$$\begin{aligned} F_v &= f_{\theta_t}(I), \\ F_t &= h_{\theta_t}(T, [eot]), \\ F_{eot} &= F_t[:, -1, :]. \end{aligned} \quad (1)$$

After separate feature extractions, explicit prompt embedding needs to interact with the visual embeddings. We design Text2Video Cross Attention Pooling for this, which is based on cross-attention mechanism. The $[eot]$, representation of the explicit prompt, serves as the query. As for the visual embeddings, the output of the video backbone, we flatten them on the spatio-temporal dimensions and serve them as the key and value.

$$\begin{aligned} Q &= W_q \cdot F_{eot}, \\ K &= W_k \cdot F_v, \\ V &= W_v \cdot F_v, \end{aligned} \quad (2)$$

where W_q, W_k, W_v refer to the query, key, value projection matrix and Q, K, V refer to the query, key, value respectively. Then, scaled-dot attention [42] is calculated.

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V, \quad (3)$$

where d_k is the number of hidden state channels. The workflow of Text2Video Cross Attention Pooling module is shown in Fig. 2 (b). After these procedures, we gain the text-video embedding, which can serve as a significant supplement for AIGC video quality assessment pipeline and work well on the video-text consistency evaluation.

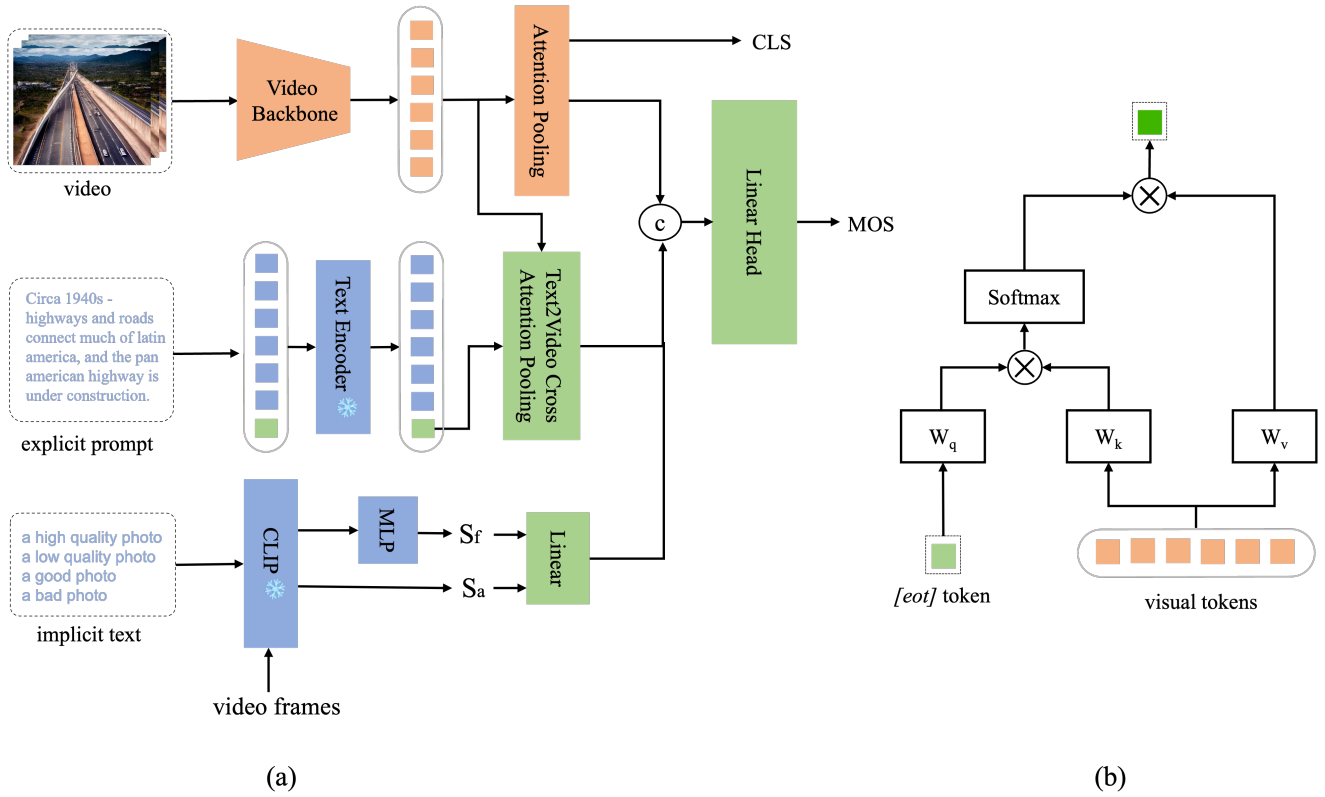


Figure 2. Detailed overview of our framework. (a) illustrates the whole framework, which serves as an incremental enhancement for DOVER. Except for the visual part, our framework also incorporates modules to deal with the explicit prompt and implicit text, enriching the capability in video-text consistency assessment. (b) shows the workflow of the Text2Video Cross Attention Pooling module, which is based on cross-attention mechanism.

Implicit Text Guidance. Inspired by BVQI [52], we use an implicit text module to evaluate video quality. We calculate the affinity scores between a given N frames video (V) and two pairs of texts (T_0, T_1), where each pair consists of one positive text and one negative text, along with the feature score of the video. Subsequently, these scores are aggregated through a linear output.

The sampling pipeline is different from the one proposed in [52]. In order to fully explore the potential information of the video, all frames are utilized and cropped to the size of 224×224 , ensuring the integrity and efficient utilization of the information. Then, following [52], we use CLIP [31] visual (E_v) and textual (E_t) encoders to extract the video feature of frame i ($f_{v,i}$) and the text feature of text pair j ($f_{t,j}$), calculate the affinity and conduct sigmoid remapping to form the final affinity scores $S_{a,0}$ and $S_{a,1}$:

$$f_{v,i} = E_v(V_i), \quad (4)$$

$$f_{t,j} = E_t(T_{j,pos}), E_t(T_{j,neg}), \quad (5)$$

$$S_{a,j} = \text{Sigmoid}\left(\frac{\sum_{i=0}^{N-1} (f_{v,i} \cdot f_{t,j,pos}^T - f_{v,i} \cdot f_{t,j,neg}^T)}{N}\right). \quad (6)$$

To bridge the gap between AIGC video frames and reality images, we generate feature score S_f from the generated video features and project it within the range of $[0,1]$:

$$S_f = \text{Sigmoid}(\text{GELU}(\text{MLPs}(f_v))). \quad (7)$$

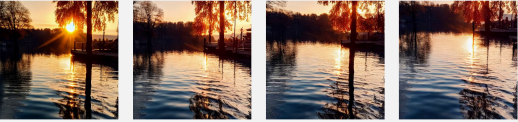
Ultimately, we combine it with affinity scores to produce implicit text score through a linear output.

$$\text{Score} = \text{Linear}(S_f, S_{a,0}, S_{a,1}). \quad (8)$$

Video-LLaVA Caption Similarity. The modeling of video-text consistency above is implicit. We also want to insert it into our assessment explicitly. So we use a video captioning model to translate the AIGC video into a brief description. Considering that MLLM has strong capabilities especially on zero-shot and few-shot learning, we choose Video-LLaVA [23] to generate appropriate caption for the corresponding AIGC video.

However, the style of natural captions is different from the textual prompts of text-to-video model. In order to alleviate this problem, we want to leverage the in-context learning ability of Video-LLaVA. So, we use 5-shot inference. In practice, we randomly choose 5 textual prompts from the train dataset, which serve as the context of MLLM. Fig. 3 shows the workflow of Video-LLaVA inference.

Video Input:



User Query:
 The input video is generated by Deep Learning Model with its corresponding prompt. Please give a description that can be used to generate this image. Here are five examples for you: \n
 1. Circa 1950s - blueprints for the hull of a ship are translated into wooden frames and painted in 1955. steel is cut for the frames.\n
 2. Clouds in the sky. time lapse.\n
 3. Waterfall in fountain.\n
 4. Beautiful shot of sunset ending over water and tree silhouettes.\n
 5. Polonnaruwa, sri lanka asia remains of the ancient city. tourist center and a lot of debris surviving stout buddha. phallic symbol locals childless woman prays.\n
 Please output your prompt here:

Video-LLaVA Output:
 A serene lake with a sunset in the background.

Prompt:
 Beautiful calm sunset or sunrise above the lake in town with sun reflecting in golden color water.

Figure 3. One example used in In-Context-Learning for Video-LLaVA to generate the prompt-like caption.

After that, we use Sentence-BERT [34] to extract the embeddings of generated captions and corresponding textual prompts and calculate cosine similarity. We normalize the output and serve them as the finale caption similarity scores.

3.3. Auxiliary Inter-domain Classification

Video generation models typically follow three technical approaches: GAN-based, auto-regressive based, and diffusion-based methods. Videos produced by these varying models exhibit distinctions in visual quality, fluency and style. Predicting the specific generative model behind AIGC videos can lead to the extraction of more discriminative features. This capability significantly aids in the enhanced assessment of AIGC video quality. Consequently, we have integrated an additional auxiliary inter-domain classification branch. This component predicts the origin of given AIGC videos from among 10 potential video generation models, substantially benefiting the quality evaluation of AIGC videos. We use cross-entropy loss L_{cls}

as the auxiliary objective function, incorporating it into the main loss with a weight of β .

$$L = L_{qual} + \beta \cdot L_{cls}$$

$$= L_{plcc} + \alpha \cdot L_{rank} + \beta \cdot L_{cls}, \quad (9)$$

where L_{qual} refers to the quality loss, composed of PLCC(Pearson Linear Correlation Coefficient) loss L_{plcc} and rank loss [5] L_{rank} with α as its weight. In practice, the values of α and β are set to 0.3 and 0.2 respectively.

4. Experiments

4.1. Dataset

Our experiments utilize the AI-Generated video dataset proposed for the video track of the NTIRE 2024 Quality Assessment for AI-Generated Content [18, 24]. The dataset can be divided into three parts: the validation dataset, the test dataset, and the train dataset. The train dataset comprises 7000 videos, along with their corresponding Mean Opinion Scores (MOS) and textual prompts. The validation and test datasets encompass 2000 and 1000 videos respectively, including only their prompts.

All videos comprise either 15 or 16 frames, possessing a duration of 4 seconds, and exhibit a frame rate of either 3.75 or 4 frames per second. We observe that video filenames in the train dataset adhere to a structured format: "x_y.mp4," where "x" identifies the video's unique number within the dataset, and "y" indicates the distinct generative model employed.

4.2. Implement Details

Following DOVER [53], we apply temporal and spatial sample to raw videos. During both training and testing phases, frames are sampled comprehensively across all branches. Specifically, in the DOVER technical branch, 7×7 spatial grids are utilized. For other branches, frames are sampled and resized to a resolution of 224×224 .

The whole procedures are implemented using Python programming language, leveraging the PyTorch framework for deep learning. For dover-base branches, we employ the AdamW optimizer with a weight decay of 0.05. Different learning rates are used for the backbones and heads: the backbones are trained with a learning rate of $6.25e-5$, while the heads are trained with a larger learning rate of $6.25e-4$. The training process is conducted on a single NVIDIA GeForce RTX 4090 24GB GPU, with each branch trained for 25 epochs. We divide the 25 training epochs into two phases: 10 linear-probe epochs and 15 end-to-end fine-tuning epochs, following the DOVER [53] approach. Applying this training strategy to the branches results in a training time of approximately 8 hours. Due to the limita-

tion of GPU memory, branches with larger backbones requires 25 linear-probe epochs, which takes approximately 4 hours to complete.

In order for further improvement, we also use model-ensemble tricks, by weighted summation of results from various models. The advanced VQA methods [50, 51, 57] and linear-probing of strongly pretrained backbones [21, 22, 46] are integrated.

4.3. Evaluation Metrics

Like traditional No-Reference Video Quality Assessment, MOS (Mean Opinion Score) is the ground-truth. The mixture of PLCC (Pearson’s Linear Correlation Coefficient) and SROCC (Spearman’s Rank Order Correlation Coefficient) serves as the evaluation metrics.

PLCC is a measure of the linear correlation between predicted scores and MOS. It ranges from -1 to 1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. The formula is shown as following:

$$PLCC = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}, \quad (10)$$

where X_i and Y_i refer to the prediction and target of the i^{th} sample. \bar{X} and \bar{Y} are the means.

SROCC is a measure used to evaluate the strength and direction of association between two ranked variables. Unlike PLCC, which assesses linear relationships, SROCC is used to identify monotonic relationships (whether linear or not). It ranges from -1 to 1, where 1 indicates a perfect positive association, -1 a perfect negative association, and 0 no association.

$$SROCC = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad (11)$$

d_i is the difference between the ranks of corresponding values of predictions and targets, and n is the number of observations.

The overall score, *MainScore*, is obtained by ignoring the sign and reporting the average of absolute values $((PLCC + SROCC)/2)$.

4.4. Experimental Results

We conduct a comparative evaluation of our method against fine-tuned state-of-the-art VQA methods including SimpleVQA [39], BVQA [19], Fast-VQA [50] and DOVER [53] on the validation dataset. To ensure a fair comparison, we report only the performance of a single model and do not employ model ensemble. The experimental results shown in Tab.1 demonstrate that our approach outperforms the existing VQA methods.

Table 1. Results on the NTIRE 2024 Quality Assessment for AI-Generated Content - Track 2 Video Challenge Validation.

Models	PLCC	SROCC	Main Score
SimpleVQA [39]	0.6338	0.6275	0.6306
BVQA [19]	0.7486	0.7390	0.7438
Fast-VQA [50]	0.7295	0.7173	0.7234
DOVER [53]	0.7693	0.7609	0.7651
Ours	0.8099	0.7905	0.8002

The NTIRE 2024 Quality Assessment for AI-Generated Content - Track 2 Video Challenge has the goal of developing a solution for AIGC video quality assessment. 13 teams were involved in the finale submission stage. All these teams have evaluated their proposed methods on the unseen validation and test set, and then submit the fact-sheet. Tab. 2 shows the leaderboard of this challenge, according to the *MainScore* on test set. Our proposed method was used in the third-place winner of the NTIRE 2024 Quality Assessment for AI-Generated Content - Track 2 Video. [18, 24]

Table 2. The leaderboard of the NTIRE 2024 Quality Assessment for AI-Generated Content - Track 2 Video Challenge.

Team name	Main Score
ICML-USTC	0.8385
Kwai-kaa	0.824
SQL	0.8232
musicbeer	0.8231
finnbingo	0.8211
PromptSync	0.8178
QA-FTE	0.8128
MediaSecurity_SYSU&Alibaba	0.8124
IPPL-VQA	0.8003
IVP-Lab	0.7944
Oblivion	0.7869
CUC-IMC	0.7802
UBC DSL Team	0.7531

4.5. Ablation Study

To verify the effectiveness of proposed methods, we conduct ablation study on the validation set. We use the visual harmony model, i.e. fine-tuned DOVER [53] as our baseline. The purpose of the ablation studies is to explore the effectiveness of explicit prompt injection, implicit text guidance, auxiliary inter-domain classification (represented by Aux-CIs), model ensemble and Video-LLaVA [23] caption similarity. Main results are shown in Tab. 3.

Impact of Explicit Prompt Injection. The *Explicit Prompt Injection* is designed to get multi-modal text-video interaction features for better assessment on video-text consistency. With *Explicit Prompt Injection*, the performance in-

Table 3. The ablation results on the validation set.

Explicit-Prompt	Implicit-Text	Aux-Cls	Model-Ensemble	Video-LLaVA	PLCC	SROCC	MainScore
					0.7649	0.7417	0.7533
✓					0.7888	0.7676	0.7782
	✓				0.7843	0.7631	0.7737
✓	✓				0.7991	0.7803	0.7897
✓		✓			0.8020	0.7814	0.7917
✓	✓	✓			0.8099	0.7905	0.8002
✓	✓	✓	✓		0.8317	0.8153	0.8235
✓	✓	✓	✓	✓	0.8341	0.8165	0.8253

creases from 0.7533 to 0.7782. This result shows that video-text consistency is important for AIGC video quality assessment and our *Explicit Prompt Injection* can enhance the multi-modal understanding.

Impact of Implicit Text Guidance. The *Implicit Text Guidance* module leverages the multi-modal alignment and understanding capabilities of CLIP [32] to enhance the features of video frames. With *Implicit Text Guidance*, the performance increases 0.0204, compared with the visual harmony baseline.

Impact of Auxiliary Inter-domain Classification. *Auxiliary Inter-domain Classification* is an auxiliary task to predict the video generation model. AIGC videos generated by different text-to-video generative models have different visual quality, fluency and style. So, It is a good way to make the AIGC video features more discriminative. According to Tab. 3 Line.2 and Line.5, using *Auxiliary Inter-domain Classification* improves the *MainScore* from 0.7782 to 0.7917. Additionally, comparing Tab. 3 Line.4 with Line.6, there is an increase from 0.7897 to 0.8002. These results shows that our *Auxiliary Inter-domain Classification* task benefits for AIGC video quality assessment.

Impact of Model Ensemble. Model ensemble is a good way to make our results more robust. In order for further enhancement, we leverage the train pipeline shown in Fig. 2, initializing the video backbone by UniformerV2 [21] etc. and linear-probing without auxiliary inter-domain classification loss. These video backbones are pretrained on huge database with a large number of videos. So, they have strong capability on spatio-temporal modeling. As shown in Tab. 3, *Model Ensemble* can increase the performance from 0.8002 to 0.8235.

Impact of Video-LLaVA Caption Similarity. In order to integrate the video-text consistency modeling explicitly, we use Video-LLaVA [23] to generate captions and calculate the cosine similarity between generated captions and textual prompts via Sentence-BERT [34]. We also give five prompts from the train dataset as context for Video-LLaVA to generate prompt-like caption, which is shown in Fig. 3. By this operation, *MainScore* is increased from 0.8235 to 0.8253. We believe that the strong capabilities on in-context learning and zero-shot inference of MLLM will help a lot on the test dataset and other open scenes.

5. Conclusion

We decouple AIGC videos quality assessment into three dimensions: visual harmony, video-text consistency and domain distribution gap. According to this, we design corresponding models or modules respectively for comprehensive AIGC videos quality assessment. Due to the inherent multi-modal nature of AIGC videos, we propose a multi-modal framework, integrated with explicit and implicit textual prompts. During this research, we also find that videos generated by different text-to-video models have different visual quality, style and temporal fluency. Therefore, we incorporate an auxiliary inter-domain classification, predicting the source video generation model. This operation makes the features of AIGC videos more discriminative and benefits the quality assessment. Our method was used in the **third-place winner** of the NTIRE 2024 Quality Assessment for AI-Generated Content - Track 2 Video. Experimental results show the effectiveness of our proposed method. We believe that AIGC videos quality assessment can give a beneficial feedback to text-to-video generation and a larger AIGC videos dataset with samples from more recent T2V models is needed.

References

- [1] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models, 2023. [2](#)
- [2] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction, 2023. [2](#)
- [3] Joshua Peter Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C. Bovik. Chipqa: No-reference video quality prediction via space-time chips. *IEEE Transactions on Image Processing*, 30:8059–8074, 2021. [2](#)
- [4] Songlin Fan and Wei Gao. Screen-based 3d subjective experiment software, 2023. [2](#)
- [5] Fei Gao, Dacheng Tao, Xinbo Gao, and Xuelong Li. Learning to rank for blind image quality assessment, 2019. [5](#)
- [6] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. *arXiv preprint arXiv:2204.03638*, 2022. [2](#)
- [7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7, 2022. [2](#)
- [8] Zixuan Guo, Wei Gao, Haiqiang Wang, Junle Wang, and Songlin Fan. No-reference deep quality assessment of compressed light field images. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. [2](#)
- [9] William Harvey, Saïed Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos, 2022. [2](#)
- [10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint*, 2021. [1](#), [2](#)
- [11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. [2](#)
- [12] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. [2](#)
- [13] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2017. [3](#)
- [14] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database, 2017. [3](#)
- [15] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019. [2](#)
- [16] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019. [2](#)
- [17] Tengchuan Kou, Xiaohong Liu, Wei Sun, Jun Jia, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Stablevqa: A deep no-reference quality assessment model for video stability. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1066–1076, 2023. [1](#), [2](#)
- [18] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dataset and metric for text-to-video quality assessment, 2024. [2](#), [5](#), [6](#)
- [19] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception, 2022. [6](#)
- [20] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. pages 2351–2359, 2019. [2](#)
- [21] Kunchang Li, Yali Wang, Yanan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer, 2022. [2](#), [6](#), [7](#)
- [22] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yanan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models, 2023. [6](#)
- [23] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. [2](#), [4](#), [6](#), [7](#)
- [24] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, Xiele Wu, Radu Timofte, et al. NTIRE 2024 quality assessment of AI-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. [2](#), [5](#), [6](#)
- [25] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tiejong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023. [2](#)
- [26] 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, 2021. [3](#)
- [27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. [3](#)
- [28] Kangfu Mei and Vishal M. Patel. Vidm: Video implicit diffusion models, 2022. [2](#)
- [29] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. [2](#)

- [30] Bowen Qu, Haohui Li, and Wei Gao. Bringing textual prompt to ai-generated image quality assessment, 2024. [2](#)
- [31] 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. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. [4](#)
- [32] 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 *ICML*, pages 8748–8763. PMLR, 2021. [2](#), [7](#)
- [33] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer, 2020. [2](#)
- [34] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. [2](#), [5](#), [7](#)
- [35] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, 23(3):1352–1365, 2014. [2](#)
- [36] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2849–2858, 2016. [2](#)
- [37] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 29, 2016. [1](#), [2](#)
- [38] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2019. [3](#)
- [39] 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](#), [6](#)
- [40] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *arXiv preprint arXiv:2005.14354*, 2020. [2](#)
- [41] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric challenges, 2019. [1](#), [2](#)
- [42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. [3](#)
- [43] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Neural Information Processing Systems*, 2016. [1](#), [2](#)
- [44] Jilong Wang, Wei Gao, and Ge Li. Applying collaborative adversarial learning to blind point cloud quality measurement. *IEEE Transactions on Instrumentation and Measurement*, 72:1–15, 2023. [2](#)
- [45] Jilong Wang, Wei Gao, and Ge Li. Zoom to perceive better: No-reference point cloud quality assessment via exploring effective multiscale feature. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2024. [2](#)
- [46] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *CVPR*, 2023. [6](#)
- [47] Yaohui Wang, Piotr Tadeusz Bilinski, François Brémond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5263–5272, 2019. [2](#)
- [48] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models, 2023. [1](#), [2](#)
- [49] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models, 2020. [2](#)
- [50] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fastvqa: Efficient end-to-end video quality assessment with fragment sampling. In *Proceedings of European Conference of Computer Vision (ECCV)*, 2022. [1](#), [2](#), [6](#)
- [51] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment, 2022. [2](#), [6](#)
- [52] Haoning Wu, Liang Liao, Jingwen Hou, Chaofeng Chen, Erli Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring opinion-unaware video quality assessment with semantic affinity criterion. *IEEE International Conference on Multimedia and Expo (ICME)*, 2023. [2](#), [4](#)
- [53] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *International Conference on Computer Vision (ICCV)*, 2023. [2](#), [3](#), [5](#), [6](#)
- [54] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching Imms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. [2](#)
- [55] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers, 2021. [1](#), [2](#)
- [56] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq: ‘patching up’ the video quality problem. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. [3](#)
- [57] Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pre-trained models for blind image quality assessment. *arXiv preprint arXiv:2303.00521*, 2023. [1](#), [6](#)