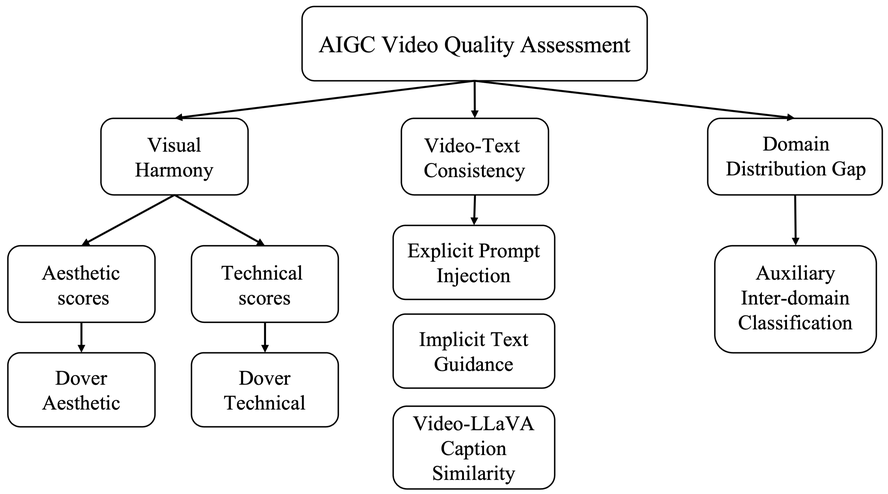
Exploring AIGC Video Quality: A Focus on Visual Harmony, Video-Text Consistency and Domain Distribution Gap

Bowen Qu[1], Xiaoyu Liang[1], Shangkun Sun[1,2], Wei Gao[1,2*]
[1]Peking University & [2]Peng Cheng Laboratory.
* Indicates corresponding author

Code will be available at:
https://github.com/Coobiw/TriVQA

TriVQA: Triple-Dimensional AIGC Video Quality Assessment

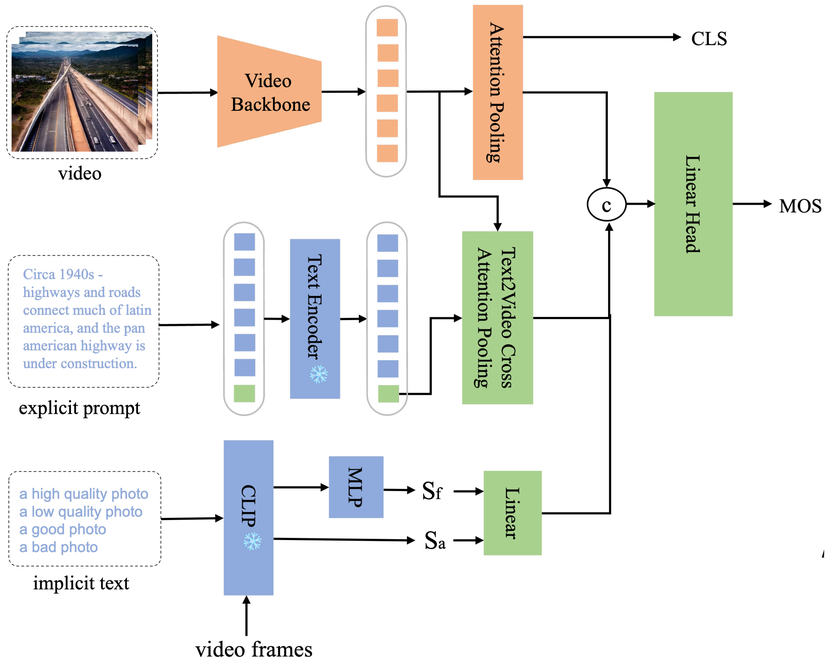


Our contributions:

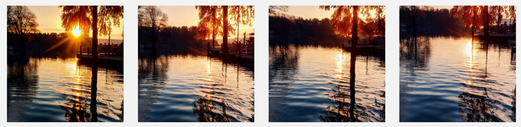
- We propose a new quality assessment framework for AIGC videos, which we decouple **into three aspects: visual harmony, video-text consistency and domain distribution gap**.
- For each aspect, we design specific modeling methods such as LLM and auxiliary inter-domain classifiers, to propose effective solutions.
- Our method shows remarkabale 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.**

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

Framework of TriVQA & Video-LLaVA Enhancement



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.

- Due to the inherent multi-modal nature of AIGC videos, we propose a multi-modal dual-stream framework, integrated with **explicit and implicit textual prompts.**
- We incorporate **an auxiliary inter-domain classification**, predicting the source video generation model.
- We use Video-LLaVA to generate captions and calculate the cosine similairy between generated captions and textual prompts via Sentence-BERT. **We want to leverage the in-context learning ability of Video-LLaVA. So, we use 5-shot inference**

Ablation Study on 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**