NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

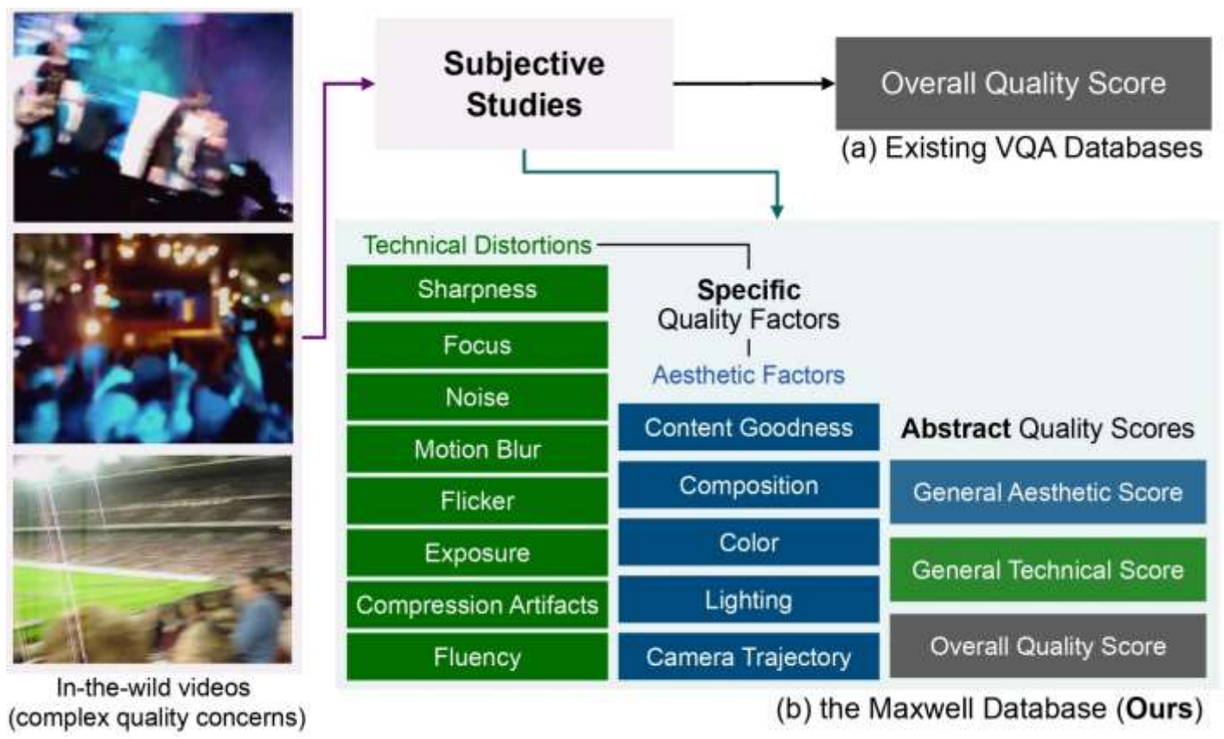
Multi-axis Video Quality Assessor (MaxVQA)

Vision-Language-Model-based Video Quality Assessment

Student: Zhang Erli Supervisor: Prof Lin Weisi

BACKGROUND

As video content consumption rises, the need to assess video quality accurately is paramount. Current video quality assessment (VQA) databases do not fully encompass the intricate quality concerns present in in-the-wild videos, often lacking explainability in their evaluation metrics.

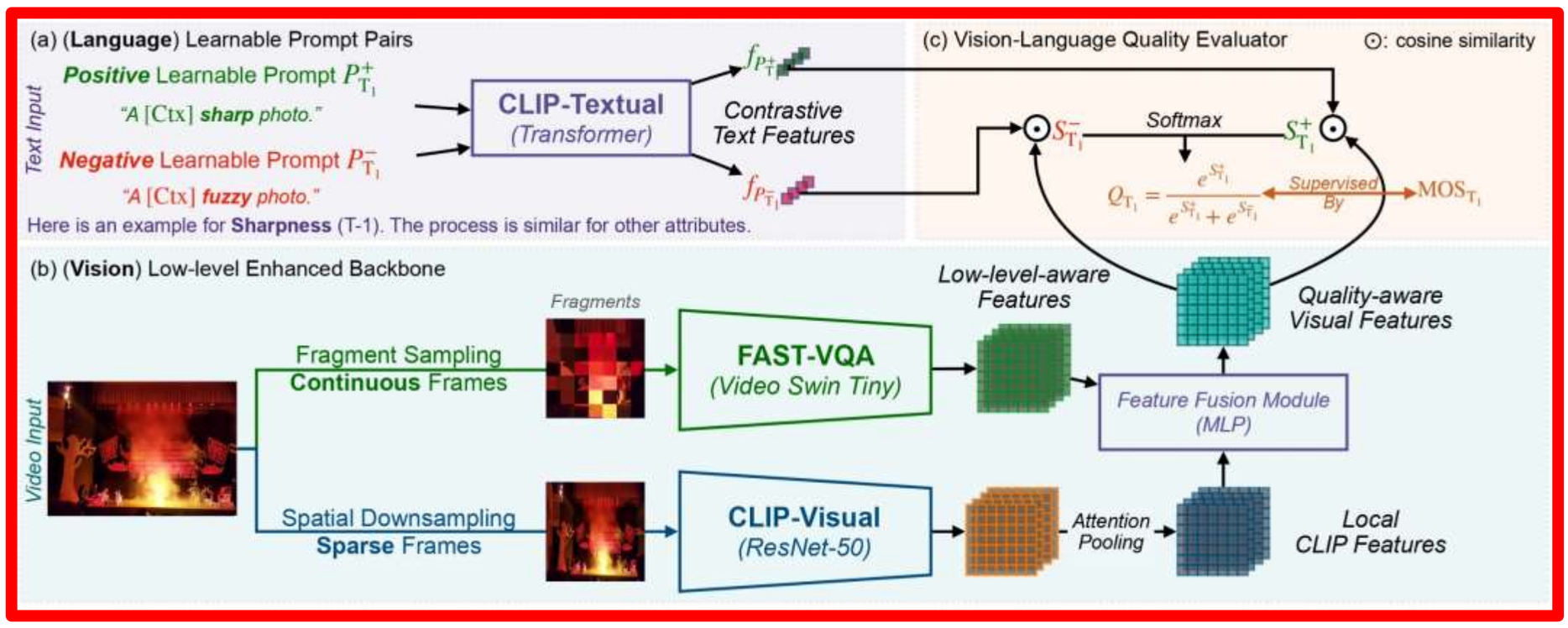


METHODOLOGY

1. Develop a 16-Dimension VQA Database encompassing a broad spectrum of quality concerns from technical distortions to general aesthetics.
2. Propose a Language-Prompted Approach using CLIP-Textual (Transformer) and CLIP-Visual (ResNet-50) to create a Vision-Language Quality Evaluator.
3. Implement FAST-VQA (Video Swin Tiny) for continuous and sparse frame processing, with attention to low-level and quality-aware visual features.
4. Fuse features using a Feature Fusion Module (MLP) for comprehensive quality assessment.

CONCLUSION & FUTURE WORK

The integration of language models and an extensive VQA database advances the field of VQA, providing granular insights into video quality factors. This method shows potential for bridging the gap between technical quality metrics and human perception, with a clear pathway towards full-scale deployment in diverse in-the-wild contexts.



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

RESULT (multi-axis benchmark comparison with existing VQA models)

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