🔎 **COVER**: A Comprehensive Video Quality Evaluator

Chenlong He[1] Qi Zheng[1] RuoXi Zhu[1] XiaoYang Zeng[1]
Yibo Fan[1] Zhengzhong Tu[2]

Now faculty of CS at Texas A&M

[1] Fudan University [2] University of Texas at Austin

CVPRW 2024

🏆 Winner of AIS 2024 UGC Video Quality Challenge 🏆

Problem statement

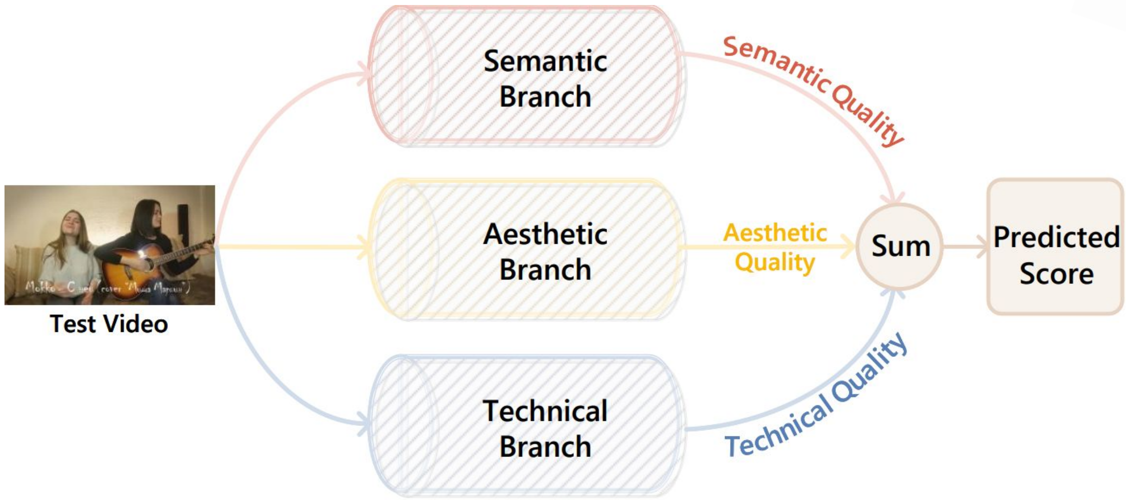
- **UGC** videos often suffer from various unpredictable distortions at different levels, such as low-level technical, mid-to-high-level aesthetic, and high-level semantic, which impact users' quality-of-experience (QoE)

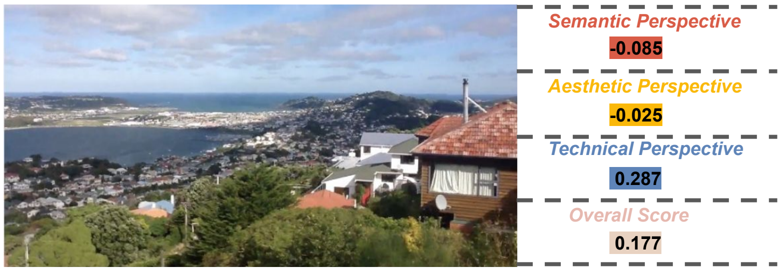
- Existing VQA models are mainly designed to quantify quality from the technical aspect, such as distortions like noise, blur, compression artifacts.

- The demand for **high-resolution** and **high-frame-rate** videos on social media platforms presents new **challenge**s for VQA tasks, as they must ensure effectiveness while also meeting real-time requirements

→ **To develop a highly efficient and comprehensive evaluator for UGC**

Our model: COVER

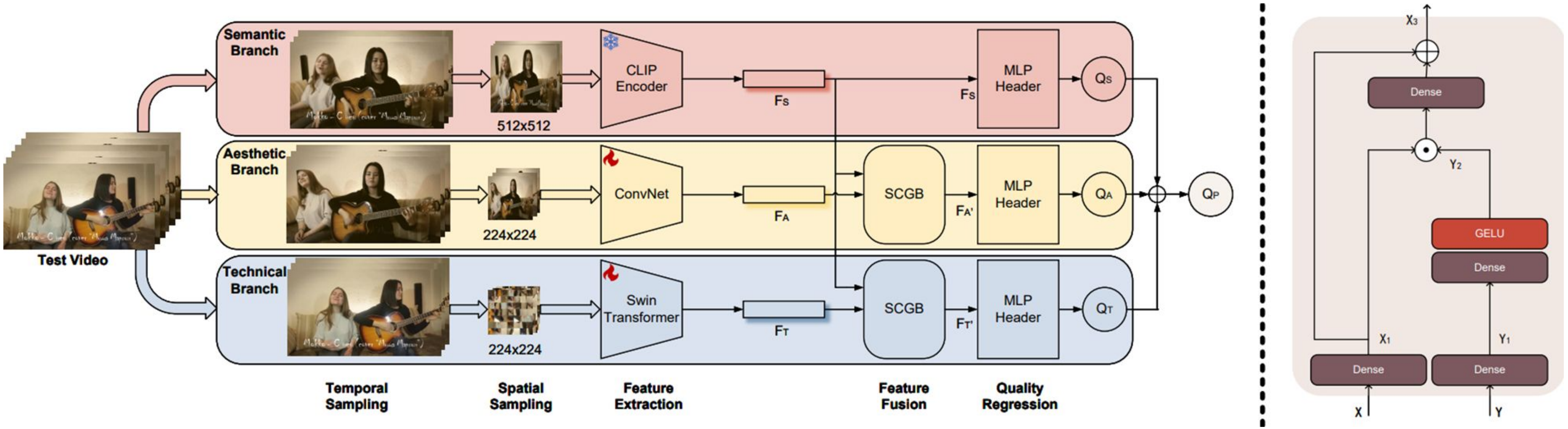




🤗 Space

🌟Code

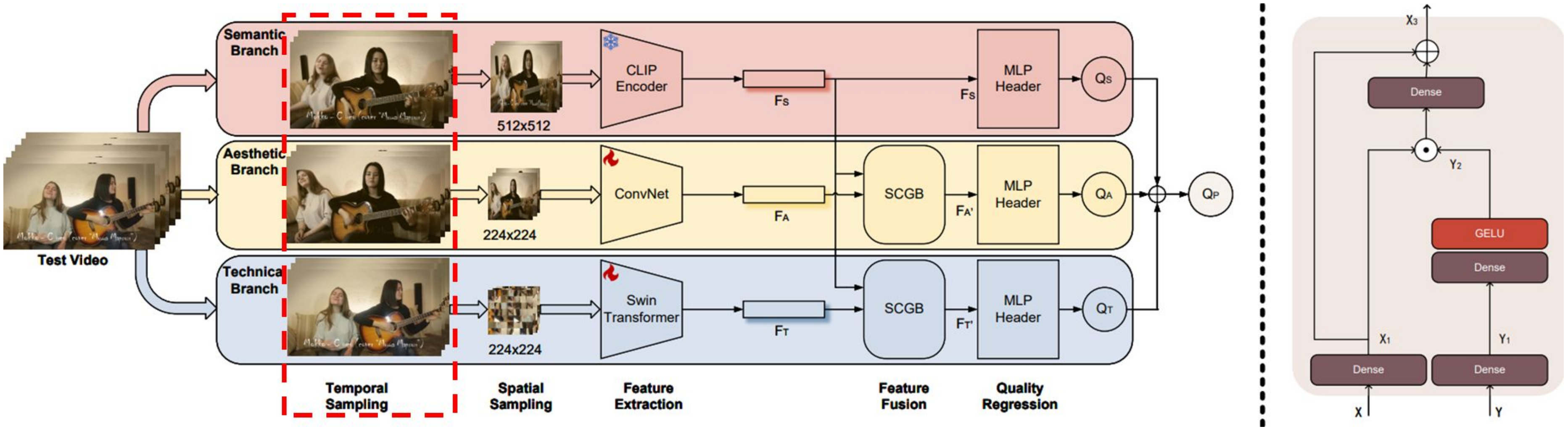
Our model: COVER



COVER processes the input video in **five steps**:

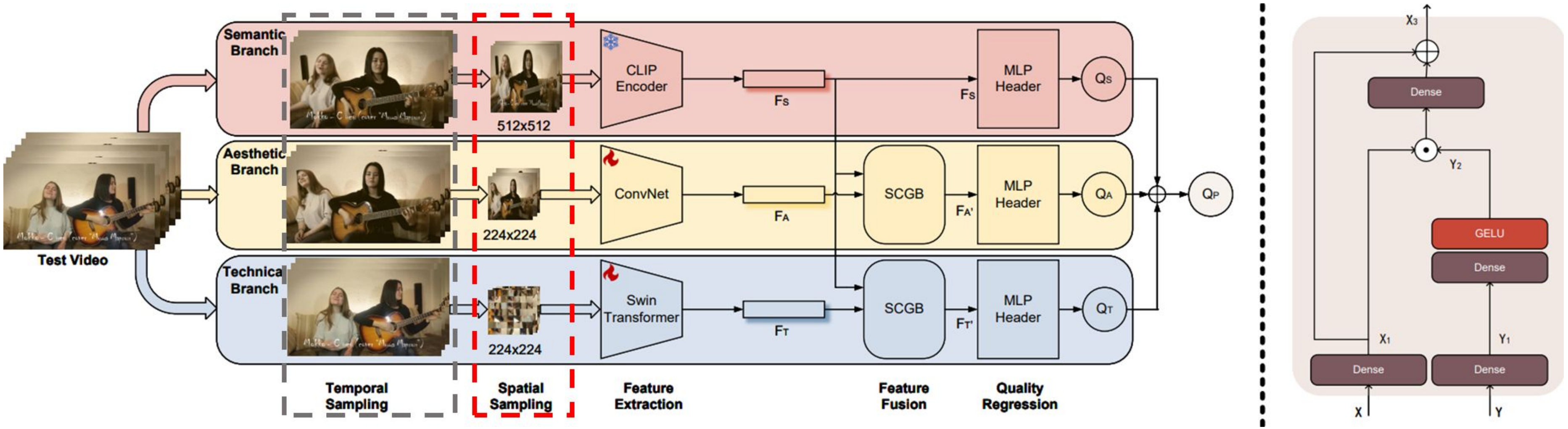
I. **Temporal Sampling**
II. **Spatial Sampling**
III. **Feature Extraction**
IV. **Feature Fusion**
V. **Quality Regression**

Step I: Temporal Sampling



- The **semantic** branch randomly samples **one** picture from **every second** of an input video
- The **aesthetic** branch randomly samples **two** picture from **every second** of an input video
- The **technical** branch randomly samples **two** picture from **every second** of an input video

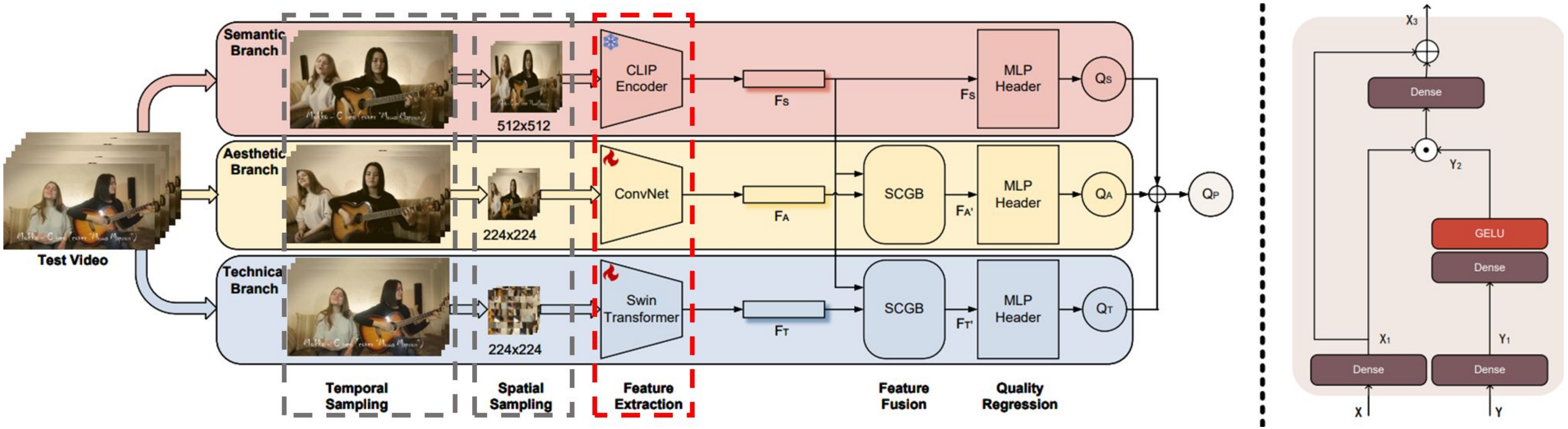
Step II: Spatial Sampling



- The **semantic** branch **resizes** the frames from their original size to **512x512** (CLIP pre-trained input)
- The **aesthetic** branch **resizes** the frames from their original size to **224x224** (Aesthetics is robust to size)
- The **technical** branch **sampled fragments*** the pictures to **224x224** (inspired from Fast-VQA*)

* H. Wu et al, FAST-VQA: Efficient End-to-end Video Quality Assessment with Fragment Sampling, ECCV 2022

Step III: Feature Extraction

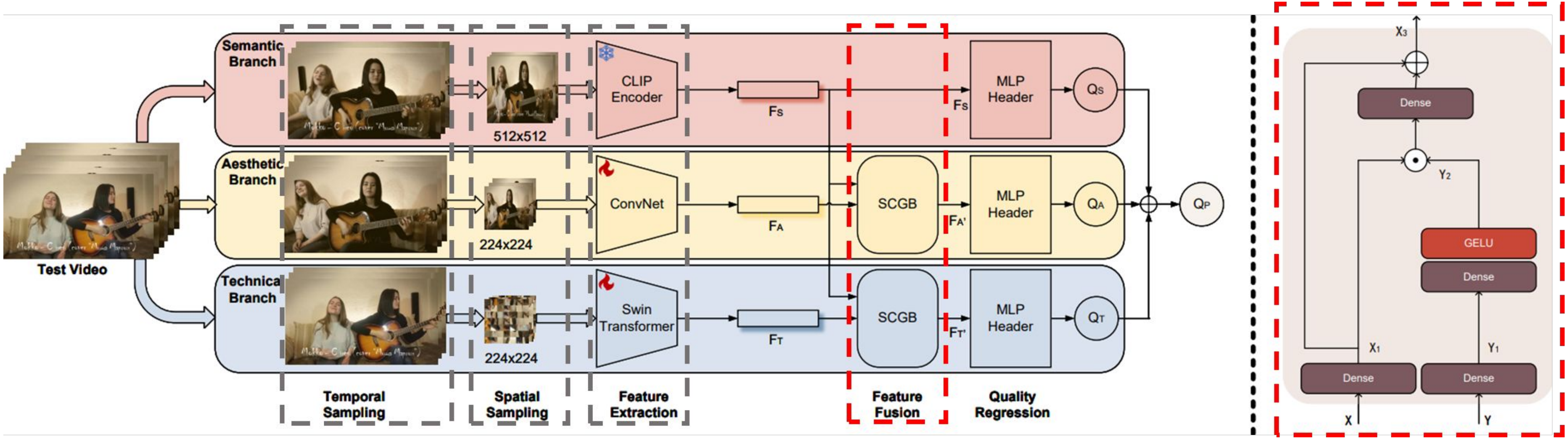


- The **image encoder** of CLIP* is used as the backbone of the **semantic branch;** A **ConvNet** is used for the **aesthetic branch**; a **Swin Transformer** is used for the **technical branch**

- During **training**, the backbone of the **semantic branch** is ❄️**frozen**, while the backbones of the **aesthetic** branch and the **technical** branch are 🔥**fine-tuned**

* OpenAI, Contrastive Language–Image Pre-training

Step IV: Feature Fusion

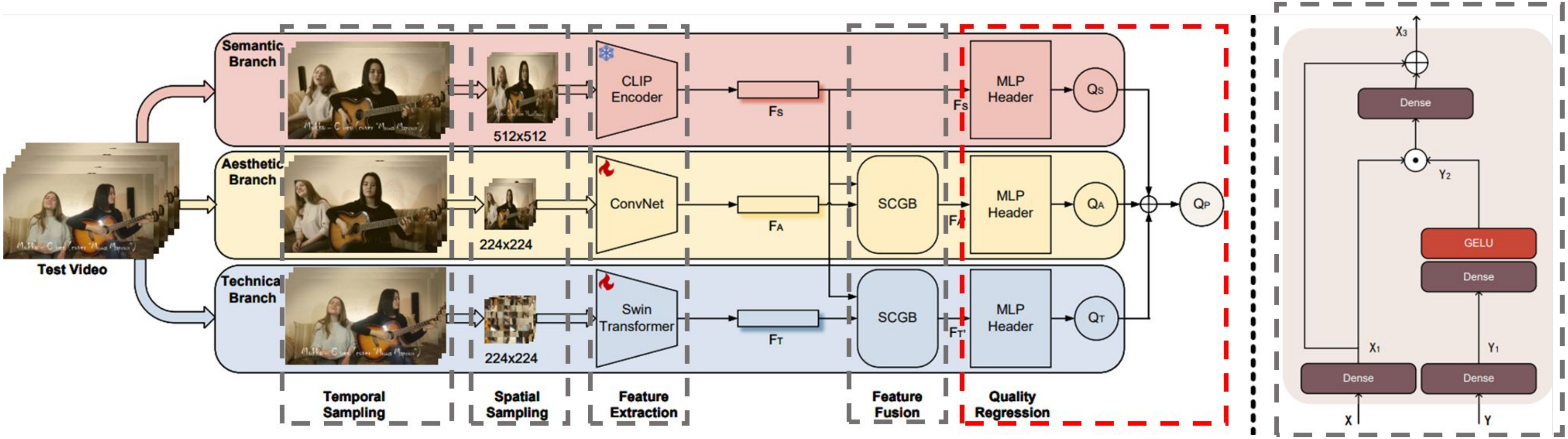


- To enable feature interactions, the **semantic** feature is used to perform **channel-wise cross-gating*** on both the **technical** feature and the **aesthetic** feature

- **Simple Cross-Gating Block (SCGB)** retains only **one gating pathway** and channel-wise interactions, eliminating operations related to spatial interactions

* Z. Tu et al, MAXIM: Multi-Axis MLP for Image Processing, CVPR 2022

Step V: Quality Regression



- The features from each branch are individually fed into an **MLP header** to predict quality scores. The **final** predicted quality **score** is the **sum** of the quality scores from the **three branches**

- While **training** MLP headers, COVER **minimizes** the relative **loss** between the predictions of each branch and the overall opinion MOS

$$\mathcal{L}_{all} = \mathcal{L}_{rel}(Q_S, \mathrm{MOS}) + \mathcal{L}_{rel}(Q_A, \mathrm{MOS}) + \mathcal{L}_{rel}(Q_T, \mathrm{MOS})$$

Experiments

Metadata	YouTube-UGC [38]
Publication year	2019
Source content	YouTube
Number of contents	1,380
Resolution	4k-360p
Framerate	15,20,24,25,30,50,60 fr/sec
Video duration	20 seconds
Experiment	Crowdsourcing (AMT)
Rating scale	Continuous rating 1-5
Number of subjects	>8,000
Number of ratings	170,159 (123 votes/video)

Team	Method	# Params. [M]	Runtime [ms]	MACs [G]
FudanVIP	COVER [9]	61.02	79.37	NA
TVQE	TVQE	8254	705.30	1127.35
Q-Align	Q-Align [40]	8198	1707.06	991.17
SJTU MMLab	SimpleVQA+ [27]	207.7	245.512	140.175
Baseline	NDNet [31]	6.95	209.43	479.06
Baseline	MobNet	2.22	347.51	1585.32

Table 3. **High-Resolution Efficiency study** using as input a clip of 30 frames of 4K resolution 3840×2160. We report the runtime and MACs for the complete clip of 30 frames.

Method	SROCC	KROCC	PLCC	RMSE
BRISQUE [17]	0.4398	0.2934	0.4525	0.5608
GM-LOG [41]	0.3501	0.2336	0.3424	0.5904
VIDEVAL [28]	0.7946	0.5959	0.7691	0.4024
RAPIQUE [29]	0.7483	0.5556	0.7482	0.4177
FAVER [45]	0.7897	0.5832	0.7898	0.3861
NIQE [18]	0.2479	0.1689	0.3146	0.5976
HIGRADE [13]	0.7639	0.5524	0.7507	0.4156
FRIQUEE [5]	0.7182	0.5268	0.7091	0.4439
CORNIA [42]	0.5988	0.4113	0.5905	0.5064
TLVQM [12]	0.6690	0.4833	0.6412	0.4831
CLIPIQA+ [32]	0.5374	0.3734	0.5801	0.5128
FasterVQA [38]	0.5345	0.3716	0.5438	0.5284
FASTVQA [37]	0.6493	0.4676	0.6792	0.4621
DOVER [39]	0.7359	0.5391	0.7653	0.4053
FasterVQA*	0.6937	0.4965	0.6909	0.4552
FASTVQA*	0.8617	0.6716	0.8669	0.3139
DOVER*	0.8761	0.6865	0.8753	0.3144
FasterVQA* (Sec. 4.6)	0.8170	0.6380	0.7510	-
AVT (Sec. 4.5)	0.8775	0.6909	0.8785	-
SimpleVQA+ [27]	0.9060	0.7280	0.9110	-
Q-Align [40]	0.9080	0.7340	0.9120	-
TVQE (Sec. 4.2)	0.9150	0.7410	0.9182	-
COVER [9]	0.9143	0.7413	0.9122	0.2519

Table 2. Extended comparison with classical and previous *state-of-the-art* methods. We report some numbers from [9]. "*" indicates models were fine-tuned using the AIS Challenge dataset.

Ablation Study

Table 7. Ablation studies of the component designs of COVER on the YouTube-UGC database [38].

No.	Branch			SCGB	YouTube-UGC [38] Validation			
	Technical	Aesthetics	Semantic		SROCC ↑	KROCC ↑	PLCC ↑	RMSE ↓
1	✔				0.8659	0.6759	0.8650	0.3159
2		✔			0.8234	0.6295	0.8439	0.3378
3			✔		0.8005	0.6096	0.8311	0.3502
4	✔	✔			0.8960	0.7180	0.8928	0.2916
5	✔		✔		0.8824	0.6997	0.8890	0.2883
6		✔	✔		0.8347	0.6455	0.8582	0.3232
7	✔	✔	✔		0.9006	0.7260	0.9052	0.2731
8	✔	✔	✔	✔	**0.9143**	**0.7413**	**0.9165**	**0.2519**

- **No.1-3**: The **technical** branch has the **best** performance among the three branches
- **No.4-6**: **Combining** either the **aesthetic** branch **or** the **semantic** branch **with** the **technical** branch can **lead to** significant performance **improvements**
- **No.7-8**: **Adding** the **SCGB** feature fusion block can further **push** the **performance** limit by approximately **1.5%** in SROCC

Thank you!

COVER

🤗 Space

🌟Code