University of BRISTOL

WACV 2026 TUCSON, AZ 3/6 – 3/10

CAMP-VQA:

Caption-Embedded Multimodal Perception for No-Reference Quality Assessment of Compressed Video

Xinyi Wang, Angeliki Katsenou, Junxiao Shen, David Bull

School of Computer Science, University of Bristol, Bristol, United Kingdom

March, 2026

Motivation

- Motivation:
  - User-generated content (UGC) videos exhibit large quality variations and mixed authentic distortions due to “uncontrolled” capture conditions and bandwidth constraints.
  - Metric generalization is improved with deep no-reference VQA, yet its performance depends on effective perceptual feature extraction.
  - To capture semantic factors of perceptual quality, many language-driven NR-VQA approaches rely on extensive manual labeling, which limits scalability.
- Challenge:
  - As a global label, MOS offers limited explainability and provides weak supervision for distinguishing specific distortions and semantic-level quality factors.
  - Stronger artifact awareness typically requires fine-grained annotations (e.g., blur, banding), which are costly and time-consuming to obtain.
  - A key challenge is to automatically extract semantic-aware features and integrate them with spatio-temporal features, without detailed expert annotations.

Proposed Method

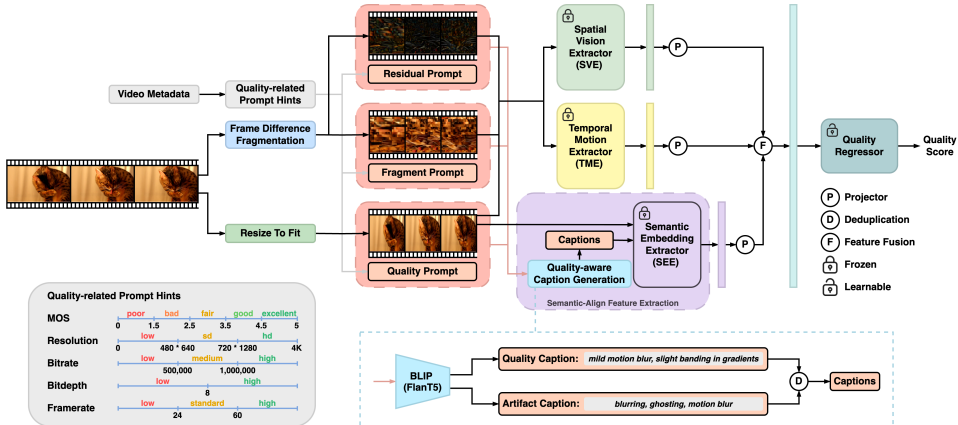

Figure 1: Overview of the proposed CAMP-VQA framework.

CAMP-VQA: Proposed Framework

- We leverage a pre-trained vision-language model (BLIP-2) to automatically generate **quality-related captions** aligned with human perception.
- Captions are produced using **quality-aware prompts** incorporating:
  - Video metadata (e.g., resolution, bitrate, frame rate)
  - Distortion-focused fragments extracted from inter-frame variations
- Video quality perception is modeled in three dimensions:
  - Semantic Embedding Extractor (SEE): artifact semantics
  - Temporal Motion Extractor (TME): temporal features
  - Spatial Vision Extractor (SVE): spatial features
- Multimodal feature fusion $\rightarrow$ MLP regressor $\rightarrow$ predicted quality score

Frame Difference Fragmentation

- Quality degradation in compression mainly occurs in regions with visual changes.
- Given consecutive frames F_t and F_{t-1}, we compute the inter-frame residual:

$$R_t(i,j) = |F_t(i,j) - F_{t-1}(i,j)|. \tag{1}$$

- We divide R_t into non-overlapping $p \times p$ patches, and compute the residual intensity for the k-th patch:

$$\Delta_k = \sum_{x=i}^{p} \sum_{y=j}^{p} \left| \mathcal{P}_t^{(k)}(x,y) - \mathcal{P}_{t-1}^{(k)}(x,y) \right|, \tag{2}$$

  where $\mathcal{P}_t^{(k)}(x,y)$ and $\mathcal{P}_{t-1}^{(k)}(x,y)$ are pixels at (x,y) in the k-th patch of F_t and F_{t-1}.
- We rank patches by Δ_k and select the top-K patches to capture regions with notable spatio-temporal variation.

Frame Difference Fragmentation (FDF)

- For each sampled frame, we construct two fragment types:
  - Residual fragments $\mathrm{F_t}^{\mathrm{res}}$ from the selected residual patches
  - Frame fragments $\mathrm{F_t}^{\mathrm{frag}}$ cropped from F_t at the same pixel positions

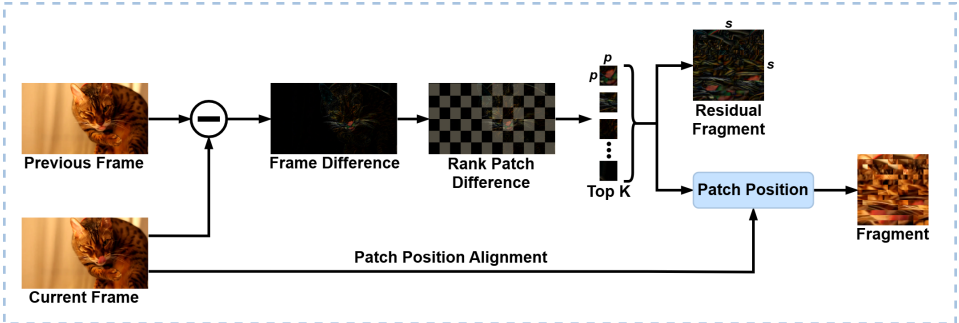

Figure 2: Frame difference fragmentation (FDF) module.

Semantic-aligned Feature Extraction: Core Idea

- **Aim**: reduce reliance on costly, fine-grained perceptual annotations, while improving awareness of artifacts and video quality factors.
- We employ a **quality-aware prompting** mechanism, incorporating video metadata with key fragments extracted from inter-frame variations to guide captioning.
- We derive **quality-related prompt hints** from multi-dimensional video metadata, and use them to dynamically adjust the input prompt.
- A pre-trained BLIP-2 model generates **quality-aware captions**, which are then mapped into a unified vision-language embedding space.

Quality-aware Caption Generation

- It automatically generates quality and artifact captions from video frames as fine-grained pseudo-annotations.
- Specifically, BLIP-2 (Flan-T5 decoder) generates three captions:

$$
\begin{aligned}
c_{\text{qlt}} &= \mathcal{G}(\boldsymbol{F}_t, \mathcal{P}_{\text{qlt}}), \\
c_{\text{res}} &= \mathcal{G}(\mathbf{F}_{\mathbf{t}}^{\text{res}}, \mathcal{P}_{\text{res}}), \\
c_{\text{frag}} &= \mathcal{G}(\mathbf{F}_{\mathbf{t}}^{\text{frag}}, \mathcal{P}_{\text{frag}}),
\end{aligned}
\tag{3}
$$

where $\mathcal{G}(\cdot, \mathcal{P})$ denotes the vision-language generation process guided by prompt $\mathcal{P}$.

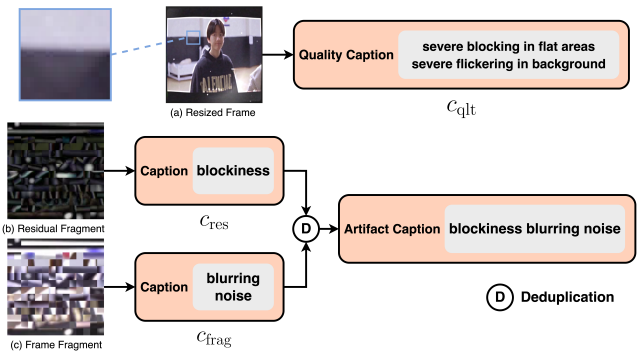

Figure 3: Example of quality-aware captioning from a sampled video frame and fragments.

Semantic Embedding Extractor (SEE)

- These textual inputs and sampled frames are passed to SEE, built on CLIP's image encoder $\phi_i(\cdot)$ and text encoder $\phi_t(\cdot)$, to extract three semantic embeddings:

$$\mathrm{e}^{\mathrm{img}} = \phi_{\mathrm{i}}(F_t),\ \mathrm{e}^{\mathrm{qlt}} = \phi_{\mathrm{t}}(\mathrm{c}_{\mathrm{qlt}}),\ \mathrm{e}^{\mathrm{art}} = \phi_{\mathrm{t}}(\mathrm{c}_{\mathrm{res}} + \mathrm{c}_{\mathrm{frag}}). \tag{4}$$

- We apply global average pooling (GAP) along the temporal dimension to obtain a video-level representation:

$$\bar{\mathrm{e}}^{(m)} = \mathrm{GAP}(\{\mathrm{e}_t^{(m)}\}_{t=1}^{T}) \in \mathbb{R}^{d_m},\ m \in \{\mathrm{img}, \mathrm{qlt}, \mathrm{art}\},\ d_m \in \{d_i, d_t, d_t\}. \tag{5}$$

- The three feature types capture image semantics, quality perception, and artifact perception, and serve as input to multimodal feature fusion:

$$\mathrm{z}_{\mathrm{semantic}} = [\bar{\mathrm{e}}^{\mathrm{img}} \,\|\, \bar{\mathrm{e}}^{\mathrm{qlt}} \,\|\, \bar{\mathrm{e}}^{\mathrm{art}}] \in \mathbb{R}^{d_i + 2\,d_t}. \tag{6}$$

Spatio-temporal Feature Extraction

- Spatial Vision Extractor (SVE) employs the SwinT encoder ϕ_{swin} to extract spatial features and capture long-range dependencies in images:

$$z_{\text{swint}} = \text{GAP}(\phi_{\text{swin}}(\text{X}_{\text{frame}})) \in \mathbb{R}^{N \times d_s}, \tag{7}$$

  where X_{frame} denotes N frames of a video clip in tensor form.

- Temporal Motion Extractor (TME) models temporal dynamics using a dual-pathway design (feature extractors ϕ_{s} and ϕ_{f}) based on the SlowFast backbone:

$$\text{z}_{\text{slowfast}} = [\text{GAP}(\phi_{\text{s}}(\text{X}_{\text{slow}})); \text{GAP}(\phi_{\text{f}}(\text{X}_{\text{fast}}))] \in \mathbb{R}^{B \times (d_s + d_f)}, \tag{8}$$

  where $X_{\text{slow}} = \text{Sample}(\text{X}_{\text{sequence}}, r)$ uses $r = \frac{1}{4}$, and X_{fast} keeps the original frame rate.

Quality Prediction

- Multimodal Video Feature Fusion:
  - Subjective video quality is jointly affected by **semantic perception**, **temporal variation**, and **spatial structure**. We therefore fuse features from three feature extractors.
  - We perform GAP within each video segment and then average over the M segments to obtain a single video-level feature per modality:

$$\mathrm{f}_c = \frac{1}{M} \sum_{i=0}^{M-1} \mathrm{f}_c^{(i)}, \quad c \in \{\mathrm{SE}, \mathrm{TM}, \mathrm{SV}\} \tag{9}$$

  - Finally, we concatenate the three modality features into a unified multimodal representation for quality regression.
- Quality Regressor:
  - We employ an MLP regression head, composed of three fully connected layers.
  - We use a composite loss function with MAE and Rank Loss to optimize learning.

Performance Comparison

- We evaluated CAMP-VQA on six mainstream UGC benchmark datasets.
  - We trained and tested on each target dataset (*intra-dataset* experiments).
  - We pre-trained on LSVQ and fine-tuned on the target datasets (*w/ fine-tune*).

Target Quality Dataset		**CVD2014**		**KoNViD-1k**		**LIVE-VQC**		**YouTube-UGC**		**$\text{LSVQ}_{\text{test}}$**		**LSVQ_{1080p}**		**FineVD**	
Model	Extra Data	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
TLVQM[13]	N/A	0.540	0.579	0.762	0.746	0.813	0.791	0.680	0.688	0.772	0.774	0.589	0.616	0.654	0.655
VIDEVAL[32]	N/A	0.766	0.806	0.807	0.792	0.773	0.775	0.781	0.793	0.794	0.783	0.545	0.554	0.731	0.731
VSFA[16]	None	0.870	0.868	0.773	0.775	0.773	0.795	0.724	0.743	0.801	0.796	0.675	0.704	0.773	0.793
PVQ[44]	LSVQ	N/A	N/A	0.791	0.786	0.827	0.837	N/A	N/A	0.827	0.828	0.711	0.739	N/A	N/A
BVQA[15]	None	0.872	0.869	0.834	0.836	0.834	0.842	0.818	0.826	0.852	0.855	0.771	0.782	N/A	N/A
SimpleVQA[31]	None	N/A	N/A	0.856	0.860	0.845	0.859	0.847	0.856	0.867	0.861	0.764	0.803	0.831	0.836
FAST-VQA[39]	LSVQ	0.891	0.903	0.891	0.892	0.849	0.862	0.855	0.852	0.876	0.877	0.779	0.814	0.835	0.847
Zoom-VQA[48]	LSVQ	N/A	N/A	0.877	0.875	0.814	0.833	N/A	N/A	0.886	0.879	0.799	0.819	N/A	N/A
DOVER[41]	LSVQ	0.858	0.881	0.909	0.906	0.860	0.875	0.890	0.891	0.888	0.889	0.795	0.830	0.842	0.839
SAMA[21]	LSVQ	N/A	N/A	0.892	0.892	0.860	0.878	0.881	0.880	0.883	0.884	0.782	0.822	N/A	N/A
ReLaX-VQA[34]	LSVQ	0.897	0.929	0.872	0.867	0.847	0.888	0.847	0.865	0.869	0.869	0.768	0.810	N/A	N/A
PTM-VQA[47]	None	N/A	N/A	0.857	0.872	0.811	0.820	0.858	0.857	0.864	0.855	0.782	0.736	N/A	N/A
COVER[8]	None	N/A	N/A	0.893	0.895	0.809	0.848	**0.914**	0.917	N/A	N/A	N/A	N/A	N/A	N/A
KSVQE[24]	KVQ	N/A	N/A	0.922	0.921	0.861	0.883	0.900	0.912	0.886	0.888	0.790	0.823	N/A	N/A
LMM-VQA[6]	LSVQ	N/A	N/A	**0.929**	0.930	0.891	0.903	0.901	0.897	**0.916**	**0.919**	**0.891**	**0.899**	N/A	N/A
FineVQ[3]	FineVD	N/A	N/A	0.915	0.910	0.895	0.895	0.910	0.914	0.900	0.900	0.828	0.857	0.883	0.889
CAMP-VQA	None	**0.933**	**0.944**	0.927	**0.936**	**0.922**	**0.940**	0.901	**0.920**	**0.920**	**0.933**	**0.908**	**0.920**	**0.919**	**0.923**
CAMP-VQA (*w/ fine-tune*)	LSVQ	**0.966**	**0.964**	**0.930**	**0.944**	**0.934**	**0.946**	**0.912**	**0.928**	**0.920***	**0.933***	**0.908***	**0.920***	**0.924**	**0.933**

Table 1: Performance comparison of the evaluated NR-VQA models on the six NR-VQA datasets.

Performance Comparison On Specific Use Case

- CAMP-VQA achieved leading results on LIVE-YT-Gaming, which focuses on UGC gaming content, and on KVQ, which targets short-video scenarios.

Target Quality Dataset		**LIVE-YT-Gaming**		**KVQ**	
Model	Extra Data	SRCC	PLCC	SRCC	PLCC
FastVQA[39]	LSVQ	0.869	0.880	0.832	0.834
DOVER[41]	LSVQ	0.852	0.868	0.833	0.837
KSVQE[24]	KVQ	N/A	N/A	0.867	0.869
LMM-VQA[6]	LSVQ	0.816	0.801	N/A	N/A
FineVQ[3]	FineVD	**0.912**	0.926	N/A	N/A
CAMP-VQA	None	0.903	0.922	0.956	0.958
CAMP-VQA (*cross-dataset*)	LSVQ	0.864	0.884	0.811	0.810
CAMP-VQA (*w/ fine-tune*)	LSVQ	0.905	**0.942**	**0.967**	**0.967**

Table 2: Performance comparison on UGC video datasets for gaming and short-form use cases.

Cross-dataset Evaluation

- In the cross-dataset setting (trained on LSVQ and evaluated on other datasets without fine-tuning), CAMP-VQA exhibited strong generalization.

Test on:	**KoNViD-1k**		**LIVE-VQC**		**YouTube-UGC**	
Train on: LSVQ	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
PVQ[44]	0.791	0.795	0.770	0.807	0.742	0.754
FastVQA[39]	0.859	0.855	0.823	0.844	0.730	0.747
DOVER[41]	0.884	0.883	0.832	0.855	0.777	0.792
LMM-VQA[6]	0.875	0.876	0.831	0.863	0.858	0.877
CAMP-VQA	**0.926**	**0.932**	**0.919**	**0.937**	**0.880**	**0.898**

Table 3: Cross-dataset evaluation results, the models were trained on LSVQ and tested on other datasets.

Ablation Studies on Semantic Embeddings

- We explored the impact of generated captions on video quality scoring through the semantic embeddings of sampled frames.
- $\bar{e}^{img}$, $\bar{e}^{qlt}$, and $\bar{e}^{art}$ denote the use of only image, quality, and artifact embeddings of video clips, respectively.

Dimension:	Color			Noise			Artifact			Blur			Temporal			Overall		
Metrics	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC
VIDEVAL	0.692	0.503	0.694	0.691	0.503	0.651	0.744	0.550	0.737	0.761	0.565	0.764	0.717	0.531	0.712	0.731	0.537	0.731
VSFA	0.762	0.567	0.784	0.764	0.571	0.728	0.801	0.608	0.817	0.777	0.586	0.800	0.728	0.536	0.702	0.773	0.583	0.793
FAST-VQA	0.802	0.608	0.818	0.809	0.621	0.776	0.818	0.630	0.833	0.835	0.650	0.851	0.756	0.561	0.739	0.835	0.648	0.847
DOVER	0.824	0.631	0.831	0.802	0.606	0.742	0.827	0.634	0.829	0.840	0.650	0.836	0.766	0.570	0.757	0.842	0.652	0.839
FineVQ	**0.850**	**0.667**	**0.853**	**0.844**	**0.661**	**0.799**	**0.885**	**0.711**	**0.892**	**0.871**	**0.696**	**0.883**	**0.809**	**0.617**	**0.760**	**0.883**	**0.712**	**0.889**
$\bar{e}^{img}$	0.804	0.612	0.828	0.763	0.569	0.727	0.806	0.612	0.814	0.802	0.607	0.814	0.723	0.532	0.721	0.804	0.611	0.817
$\bar{e}^{qlt}$	0.768	0.579	0.804	0.745	0.556	0.767	0.761	0.559	0.827	0.768	0.565	0.828	0.723	0.535	0.736	0.816	0.616	0.869
$\bar{e}^{art}$	0.750	0.541	0.765	0.764	0.556	0.736	0.787	0.577	0.807	0.778	0.563	0.809	0.736	0.527	0.713	0.812	0.594	0.840
CAMP-VQA	**0.893**	**0.712**	**0.884**	**0.873**	**0.680**	**0.819**	**0.902**	**0.725**	**0.904**	**0.906**	**0.730**	**0.907**	**0.851**	**0.658**	**0.831**	**0.919**	**0.749**	**0.923**

Table 4: Performance comparison is reported in terms of quality scoring across different quality dimensions on FineVD.

Ablation Studies on Semantic Embeddings

- The model performed best when the three quality-aware dimensions (image, quality, and artifact) were combined.
- In addition, we analyzed the impact of content embeddings.

			Test Set:	**KoNViD-1k**		**FineVD**	
$\bar{\mathbf{e}}^{\text{img}}$	$\bar{\mathbf{e}}^{\text{qlt}}$	$\bar{\mathbf{e}}^{\text{art}}$	$\text{content}_{\text{embs}}$	SRCC	PLCC	SRCC	PLCC
✓				0.778	0.804	0.804	0.817
	✓			0.631	0.792	0.816	0.869
		✓		0.735	0.763	0.812	0.840
			✓	0.409	0.451	0.401	0.409
✓	✓			0.830	0.871	0.899	0.911
✓	✓	✓		**0.903**	**0.922**	**0.901**	**0.919**
✓	✓	✓	✓	0.892	0.919	0.896	0.917

Table 5: Ablation study on the effect of different component semantic embeddings on KoNViD-1k and FineVD.

Ablation Studies on Quality-related Prompt Hints

- Similarly to the above, we extracted semantic embeddings that combined the three dimensions and tested them on KoNViD-1k.
- Upon incorporating quality-related hints, SRCC and PLCC improved from 0.767 to 0.900 and from 0.790 to 0.921, respectively.
- The results showed that our quality prompting mechanism helped generate captions that aligned more closely with human perception.

Ablation Studies on Multimodal Feature Extractors

- Table 6 demonstrated the impact of different modalities on performance across six UGC datasets.
- These three features complemented one another, forming a more robust and discriminative representation of video quality.

Test Set:			CVD2014		KoNViD-1k		LIVE-VQC		YouTube-UGC		$\text{LSVQ}_{\text{test}}$		$\text{LSVQ}_{\text{1080p}}$		FineVD	
f_{SE}	f_{TM}	f_{SV}	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
✓			0.916	0.924	0.902	0.922	0.905	0.935	0.887	0.919	0.898	0.921	0.882	0.903	0.902	0.919
	✓		0.825	0.835	0.770	0.782	0.739	0.758	0.710	0.718	0.824	0.827	0.728	0.762	0.719	0.725
		✓	0.858	0.887	0.861	0.860	0.777	0.820	0.817	0.833	0.867	0.868	0.754	0.804	0.838	0.840
✓	✓		0.923	0.941	0.909	0.921	0.917	0.936	0.885	0.910	0.910	0.928	0.901	0.912	0.907	0.918
✓	✓	✓	**0.933**	**0.944**	**0.927**	**0.936**	**0.922**	**0.940**	**0.901**	**0.920**	**0.920**	**0.933**	**0.908**	**0.920**	**0.919**	**0.923**

Table 6: Ablation study on semantic, temporal, and spatial feature (f_{SE}, f_{TM}, f_{SV}) extractors.

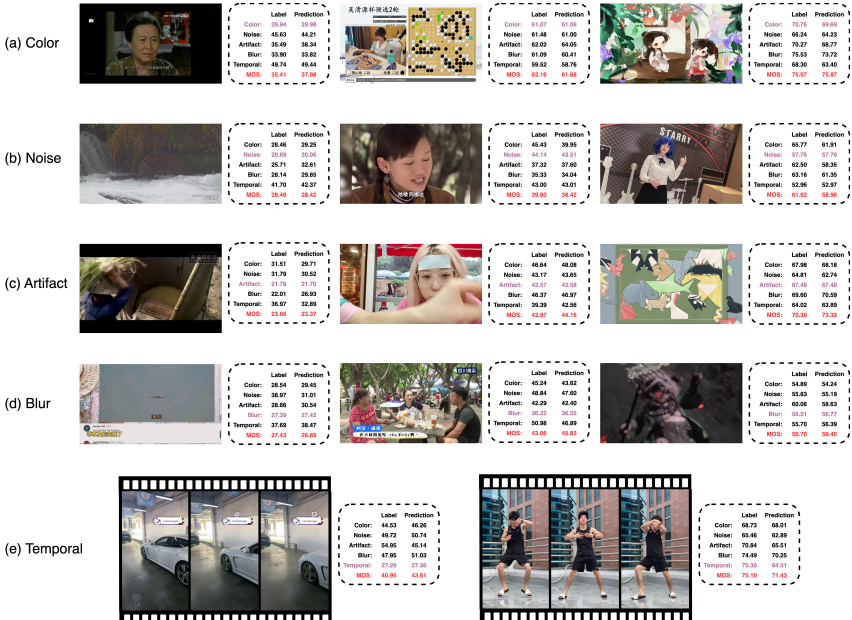

Figure 4: Predicted scores of CAMP-VQA and ground-truth MOS of FineVD across different

Conclusion and Future Work

- We propose a novel NR-VQA model that uses a VLM to obtain artifact annotations and combines those with spatio-temporal features.
- Our method introduces quality-aware caption generation, enabling automatic extraction of fine-grained, quality-related captions as substitutes for manual annotations of visual artifacts.
- The unified multimodal framework integrates SEE, TME, and SVE through feature fusion for video representations.
- CAMP-VQA achieved SOTA performance and robust generalization (average SRCC of 0.928, PLCC of 0.938).
- Future work will incorporate high-level semantic understanding, aiming to unify video question answering and video quality assessment.

Q & A

- **Thank you for listening!**
- I am happy to take any questions.
- If you find our project useful, give it a **star** on GitHub!
- **GitHub:** `github.com/xinyiW915/CAMP-VQA`
- **Contact:** xinyi.wang@bristol.ac.uk