Dolby

Attentive AV-FusionNet: Audio-Visual Quality Prediction with Hybrid Attention

INA SALAJ, ARIJIT BISWAS

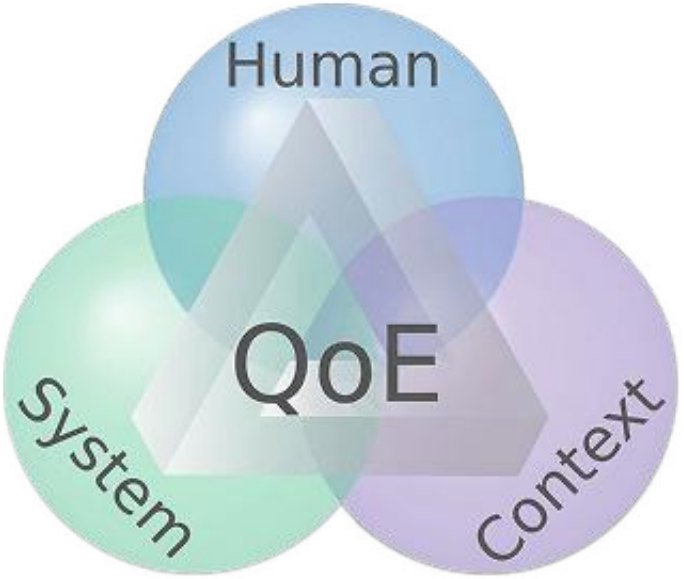
51ST ICASSP,

BARCELONA, MAY 4-8, 2026

© 2026 DOLBY | ICASSP

Motivation: Human Perception is Multimodal

- QoE is shaped by human, system, and context factors
- In audio-visual media, perceived quality is not purely a signal-level property
- The relative importance of audio and video can vary across content



Adapted from Reiter et al., *Quality of Experience: Advanced Concepts, Applications and Methods*, 2014.

© 2026 DOLBY | ICASSP

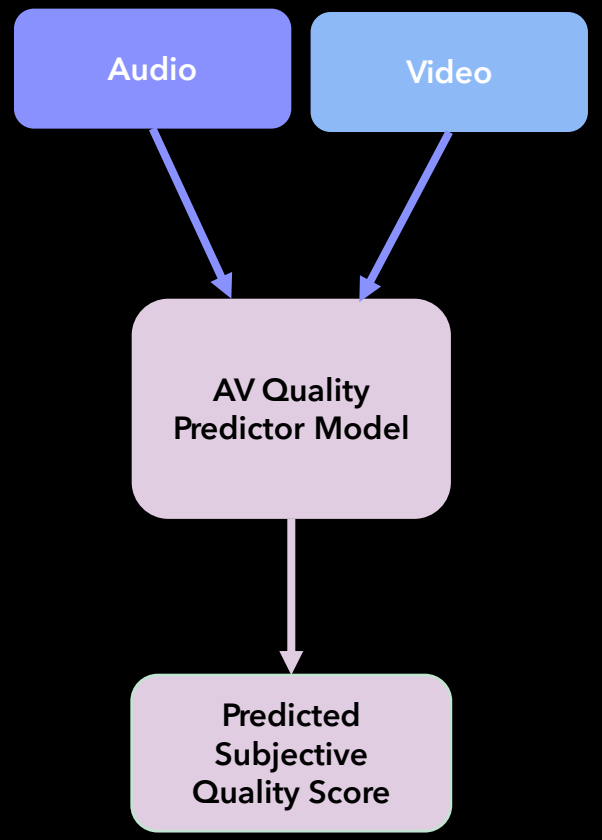
What is the general AV quality prediction setup?

The aim is to map audio and video inputs to a single estimate of overall perceived quality.

In a typical AVQ setup, both modalities are processed to extract quality-relevant information and fused into a joint prediction.

Key challenges

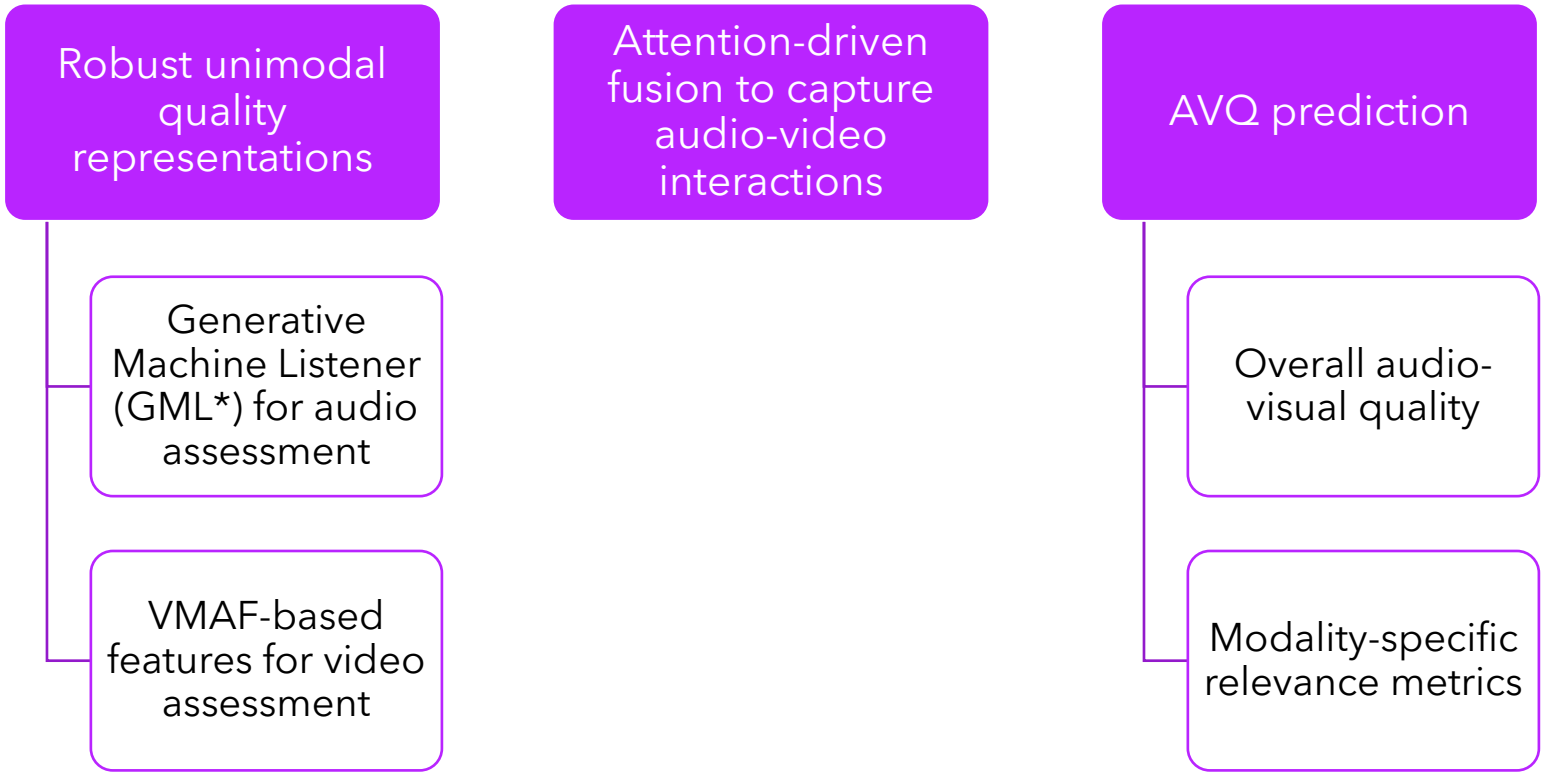
- Simple combinations may not fully capture cross-modal interaction
- Modality importance can vary with content



© 2026 DOLBY | ICASSP

Method: Attentive AV-FusionNet

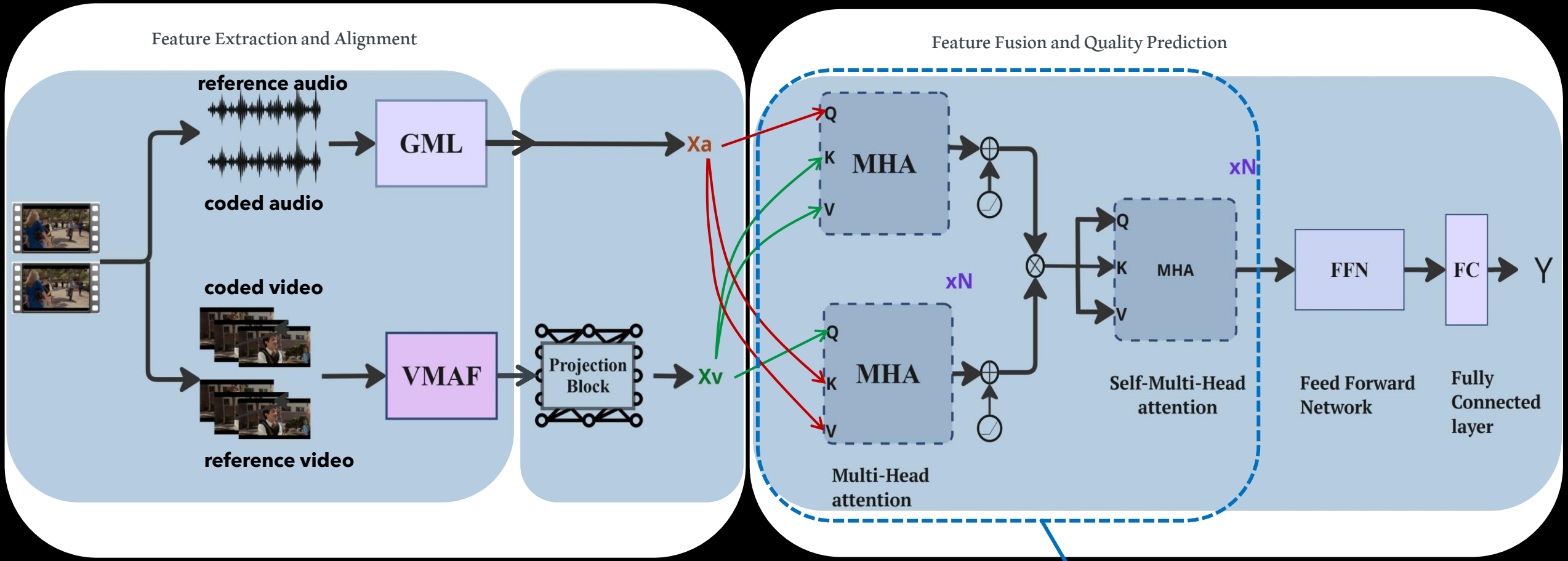
Reference-based joint prediction of audio-visual quality



*G. Jiang et al., “Generative machine listener,” in 155th AES Convention, October 2023.

© 2026 DOLBY | ICASSP

Attentive AV-FusionNet



GML: Generative Machine Listener
512-dimensional features

VMAF: Video Multimethod Assessment Fusion
4-VIF, 1-Motion2, 1-ADM

Modality Importance

Ablation sensitivity

- mask one modality
- measure prediction change

$$I_m^{\text{abl}} = \frac{\|f(x) - f(x \setminus m)\|_2}{\|f(x)\|_2}$$

Feature change norm

- compare embeddings before and after cross-attention

$$I_m^{\text{norm}} = \frac{\|\phi_m^{\text{pre}} - \phi_m^{\text{post}}\|_2}{\|\phi_m^{\text{pre}}\|_2}$$

Final importance score

- normalized weighted combination

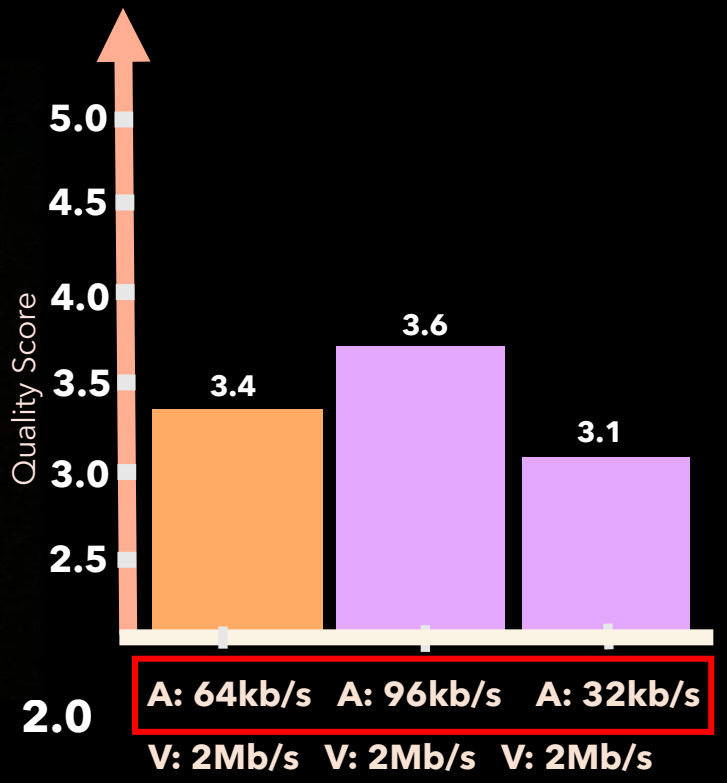
$$I_m^{\text{final}} = \alpha \cdot \text{norm}(I_m^{\text{abl}}) + \beta \cdot [1 - \text{norm}(I_m^{\text{norm}})]$$

© 2026 DOLBY | ICASSP

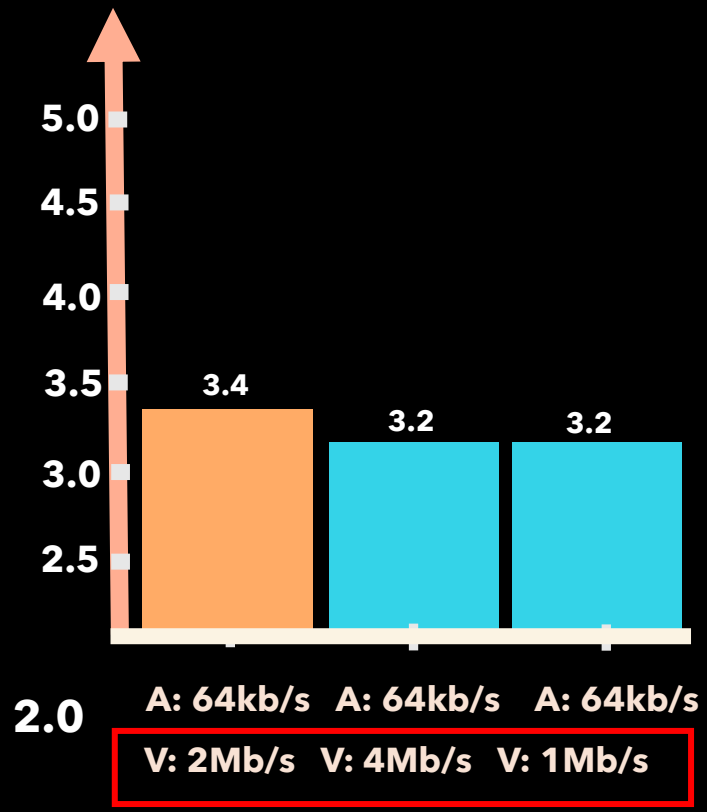
Demo – what does the model think?

Audio is more important

Graph 1: Effect of **Audio** bitrate on Perceived Quality



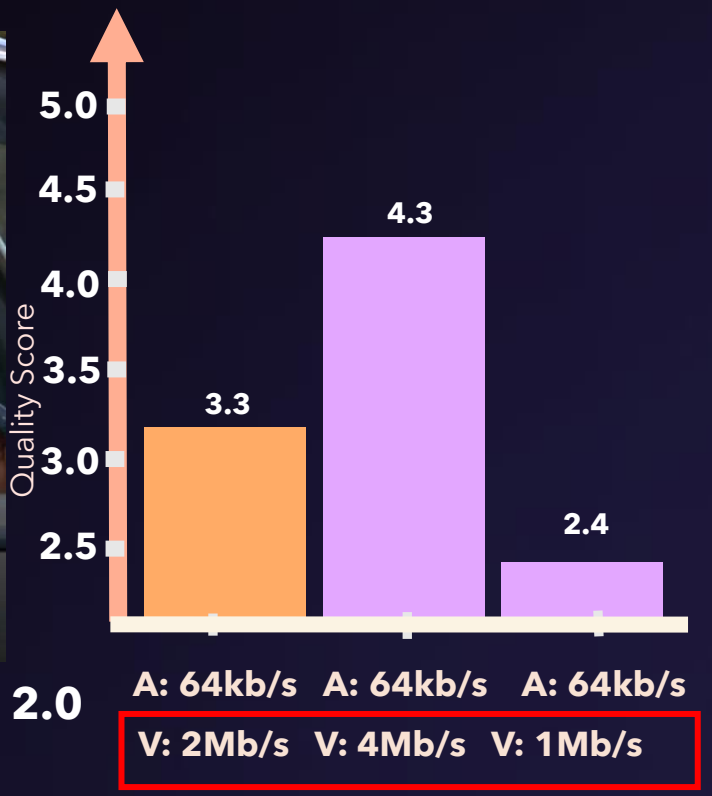
Graph 2: Effect of **Video** bitrate on Perceived Quality



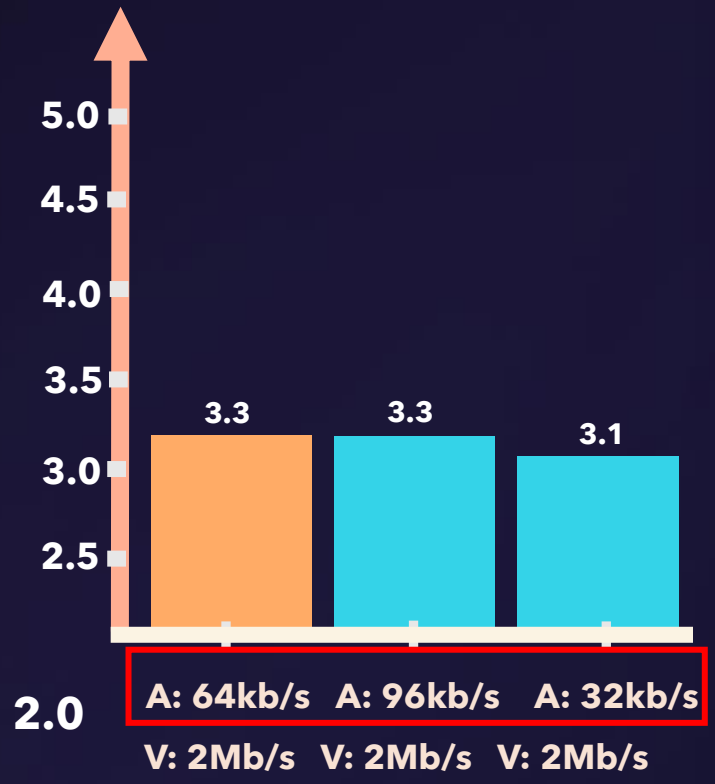
Demo – what does the model think?

Video is more important

Graph 1: Effect of **Video** bitrate on Perceived Quality



Graph 2: Effect of **Audio** bitrate on Perceived Quality



Dataset*

1625 stimuli

(65 clips × 5 video bit rates × 5 audio bit rates)

10 subjects, 5-point MOS

Audio Codec: AAC, HE-AAC v2, HE-AAC v1

Video Codec: H.264

*E. Tünçok et al., "Predicting subjective audiovisual quality of experience with oculomotor responses and perceptual video quality," 2025 17th International Conference on QoMEX

Training and Evaluation

- Training (80%) and validation (20%)
- Loss: $\boldsymbol{L = \lambda(1 - CCC) + (1 - \lambda)RMSE}$
- $\lambda = 0.6$

Evaluation Metrices

- Pearson correlation coefficient $\mathbf{R_p}$
- Spearman rank correlation coefficient $\mathbf{R_s}$
- Root mean squared error **RMSE**

Comparative Evaluation

Methods	PLCC ↑	SRCC ↑	RMSE ↓
AVMAF*	0.91	0.91	0.55
Attentive AV-FusionNet	**0.97**	**0.96**	**0.22**

Category	Model	Internal Dataset			External Dataset**		
		PLCC ↑	SRCC ↑	RMSE ↓	PLCC ↑	SRCC ↑	RMSE ↓
Baseline	Q-External	0.83	0.87	1.13	0.90	0.91	0.68
SVR	SVR-3F	0.89	0.88	0.47	0.89	0.91	0.47
Deep Learning	**Attentive AV-FusionNet**	**0.97**	**0.96**	**0.22**	**0.92**	**0.92**	**0.44**

*Min, Xiongkuo, et al. "Study of subjective and objective quality assessment of audio-visual signals." *IEEE Transactions on Image Processing* 29 (2020): 6054-6068.

**Live-SJTU

© 2026 DOLBY | ICASSP

Conclusion

- ❑ Full-reference AV quality prediction model
- ❑ Hybrid attention fusion to model cross-modal interaction
- ❑ Strong performance on internal and external datasets
- ❑ Derived modality relevance prediction

Future Work

- ❑ Support Dolby immersive formats
- ❑ Use improved GML-v2*

*V. Raj, G. K V, S. Gehlot, L. Villemoes and A. Biswas, "Enhanced Generative Machine Listener," ICASSP 2026

© 2026 DOLBY | ICASSP

—

THANK YOU

© 2026 DOLBY | ICASSP