ECCV 2026 | MALMÖ | SEPT 8–12

HAVADA: Heterogeneous Audio and Vision Adaptation for Video Saliency Prediction

Qi Bi · Remco C. Veltkamp · Albert Ali Salah

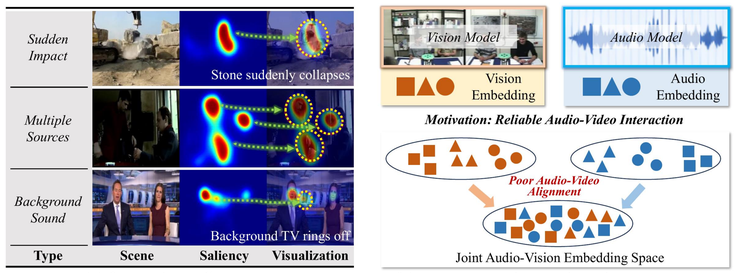
Utrecht University, The Netherlands

Utrecht University

ECCV 2026 Workshop Paper
Malmö, Sweden

Why fusion alone is not enough

Audio may guide gaze or mislead it. The model must preserve visual detail while suppressing unreliable cross-modal cues.



(a) factors that redirect gaze

(b) spurious correlations degrade representation

Sudden impact, multiple sources, and background sound create spurious audio–vision alignment.

- Late feature fusion can amplify noisy audio–vision correlations.
- Deep foundation tokens lose spatial detail needed for dense saliency.
- Independent LoRA cannot encode cross-modal reliability.

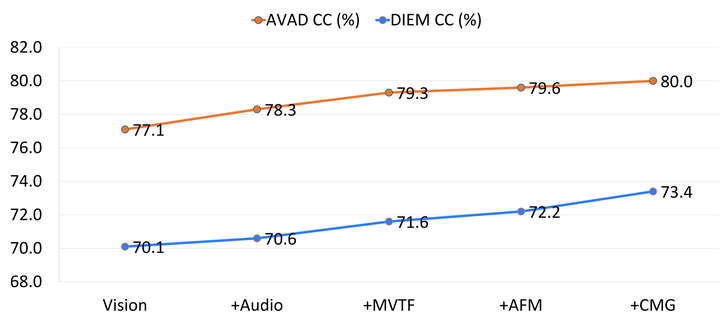
Three contributions

01 HAVADA
Parameter-efficient joint adaptation of frozen vision and audio foundation models.

02 Feature Space Adaptation
MVTF preserves spatial detail; AFM injects audio context by channel-wise modulation.

03 Gradient Space Adaptation
CMG-LoRA gates each modality's low-rank update using the other modality.

Each adaptation stage adds value



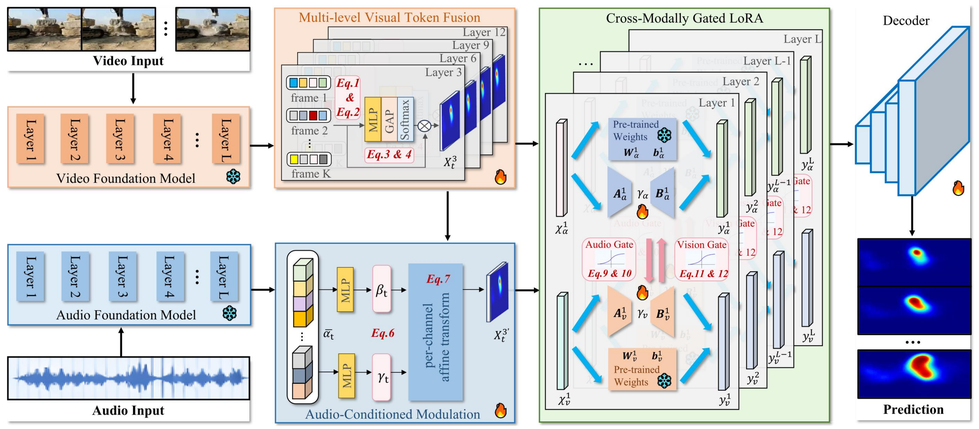
Evaluation

6 benchmarks **DIEM · AVAD · Coutrot1 · Coutrot2 · ETMD · SumMe**

Vision: CLIP / SigLIP / DINOv2
Audio: CLAP / Wav2Vec / HuBERT
Metrics: CC · NSS · AUC-Judd · SIM

HAVADA adapts features and updates

Feature space tells the decoder where to attend; gradient space controls how strongly each backbone should adapt.



Frozen foundation weights are marked by snowflakes; trainable adaptation modules are marked by flames.

FSA — preserve detail, add context

MVTF Fuses layers 3 / 6 / 9 / 12 into a localization-aware spatial map.

AFM Uses normalized audio embeddings to scale and shift visual channels.

audio-aware features without heavy cross-attention.

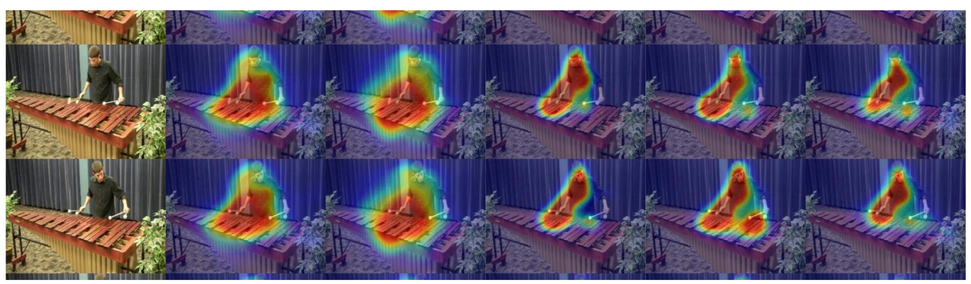
GSA — couple low-rank updates

A → V Audio-conditioned gates scale LoRA residuals in vision layers.

V → A A pooled visual descriptor gates LoRA residuals in audio layers.

sample-adaptive, reliability-aware PEFT.

Qualitative: modality-aware attention



HAVADA combines coarse auditory localization with precise visual boundaries. Representative frames shown for poster readability.

Portable across foundation models

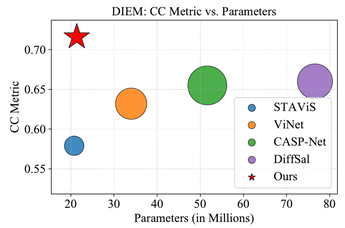
HAVADA remains competitive with CLIP, SigLIP, or DINOv2 for vision and CLAP, Wav2Vec, or HuBERT for audio.

Best AVAD CC: 0.808
CLIP + HuBERT

Lowest tested cost: 30.63 GFLOPs
CLIP + Wav2Vec / HuBERT

Strong accuracy, lightweight adaptation

20.58M	52.66G	5 / 6
trainable parameters	FLOPs	best CC scores



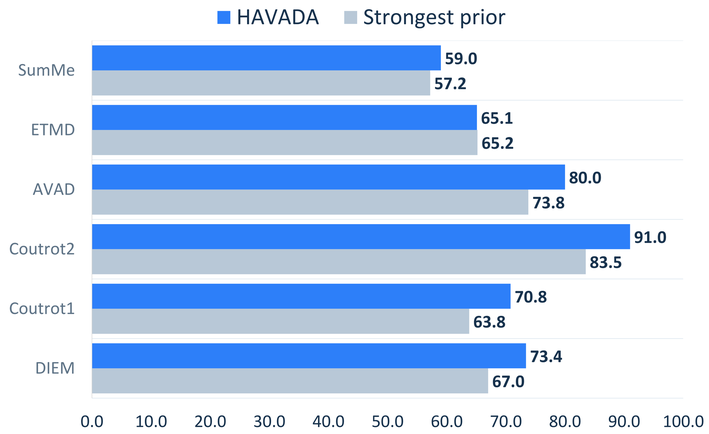
Compared with DiffSal

3.72× fewer trainable parameters

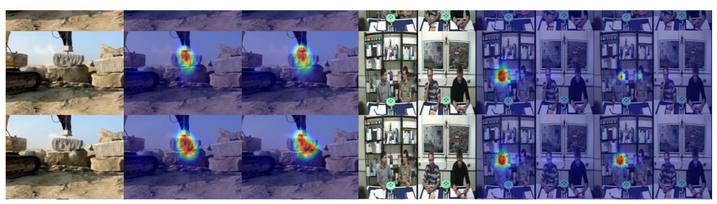
3.56× fewer FLOPs

+9.55% relative DIEM CC vs. the second-best method

CC (%) across six benchmarks



Tighter attention than DiffSal



HAVADA follows the sound-producing contact region and reduces false alarms around inactive speakers.

Take-home

Jointly condition both the features and the updates.

HAVADA turns frozen audio and vision foundation models into an efficient dense predictor. It reaches state-of-the-art performance on multiple benchmarks without full fine-tuning.

Acknowledgement: NWA-ORC HAICu (NWA 1518.22.105)

Contact: q.bi@uu.nl