HAVADA: Heterogeneous Audio and Vision Adaptation for Video Saliency Prediction

Qi Bi[1], Remco C. Veltkamp[1], and Albert Ali Salah[1]

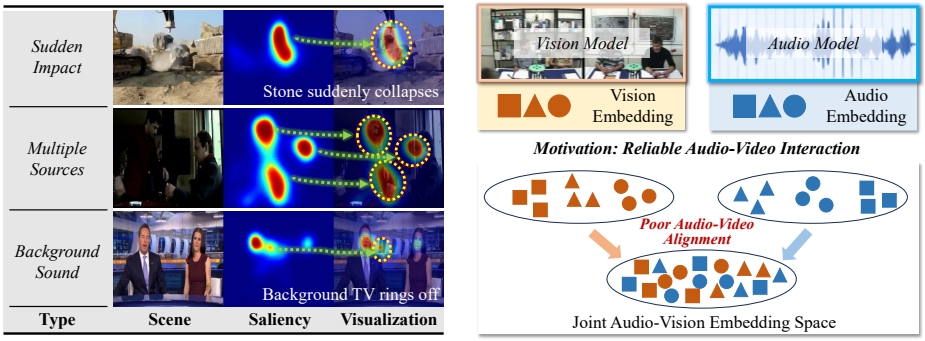
Utrecht University, Utrecht 3584CC, The Netherlands
{q.bi, R.C.Veltkamp, a.a.salah}@uu.nl

Abstract. Audio-visual video saliency prediction aims at estimating human attention in dynamic scenes by combining complementary cues from vision and sound. However, the audio-video interaction is highly complicated, as the audio can be weakly related to or even mislead the gaze due to sudden impact, off-screen narration. This work aims to jointly leverage both audio and vision foundation models for this task. The first component we propose is feature space adaptation (FSA), aiming to address the challenge of detail preserving conditioned by the audio. The second component we propose is gradient space adaptation (GSA), which aims to handle the bottleneck of task-specific capabilities from general foundation model representations. The components are complementary, and form the novel Heterogeneous Audio and Vision Adaptation (HAVADA) framework. Experiments on six audio-visual saliency benchmarks across multiple foundation models demonstrate strong performance and parameter efficiency, with state-of-the-art results on multiple benchmarks.

Keywords: Audio-Video Learning · Video Saliency Prediction · Parameter-Efficient Fine-Tuning

1 Introduction

Video saliency prediction aims to estimate the visual attention of a viewer over time by producing a dense saliency map for each video frame. It serves as a fundamental building block for downstream tasks such as video understanding [46], video captioning [17], and human-centric analytics [27]. While visual motion and semantics are primary cues, human attention in natural videos is often driven by auditory events, which can redirect gaze even when the visual appearance is ambiguous or cluttered. While video-only based saliency detection [11,13,14,41] has achieved remarkable performance, incorporating audio may provide complementary evidence for the saliency models to better capture event-driven attention shifts and audio-induced anticipation. However, this is not a trivial task, as audio can also redirect the gaze due to factors such as sudden impact, multiple sources, and off-focus sound (exemplified in Fig. 1a). Most existing audio-visual saliency models address this by feature fusion on top of conventional vision backbones pre-trained on visual datasets such as ImageNet or Kinetics [7, 18, 29, 39, 41, 43], or by treating saliency prediction as a generative modeling task [44, 48, 53]. *To*



(a) factors that redirect gaze (b) spurious correlations degrade representation

Fig. 1: (a) Audio-video saliency is challenging because auditory cues may come from sudden events, multiple sources, or background sounds. (b) The spurious audio-video correlation can lead to an unsatisfactory audio-video representation alignment in the joint embedding space, pointing to a need to learn reliable audio-video interactions.

the best of our knowledge, vision and audio foundation models, which have much stronger representation abilities, have not been combined for this task so far.

Fine-tuning audio and vision foundation models under complex audio-visual interaction scenarios is particularly difficult (illustrated in Fig. 1b). Under audio-video inconsistencies, and when audio is weakly related to visual attention, a simple audio-visual feature fusion scheme can amplify spurious correlations and degrade the representation across different audio styles and recording conditions. On the other hand, using a naive parameter-efficient fine-tuning (PEFT) [16, 28, 36] that independently adapts each modality would not explicitly model cross-modality interaction, limiting the task-specific transfer capability from the general foundation model. As a result, a joint audio-video PEFT method, which can maintain both the cross-modal context and provide durable task adaptation even under spurious audio-video correlations, is highly motivated.

To push this frontier, we propose Heterogeneous Audio and Vision Adaptation (HAVADA), a light-weight and robust framework (as demonstrated in Fig. 2). On top of a vision foundation model (e.g., CLIP [34]) and an audio foundation model (e.g., CLAP [42]) as strong and general-purpose encoders, HAVADA consists of two complementary components, namely, Feature Space Adaptation (FSA) and Gradient Space Adaptation (GSA), respectively.

Concretely, FSA aims to preserve as much spatial details as possible, which is essential for video saliency prediction, when conditioned by the audio. We devise a Multi-level Visual Token Fusion approach to jointly modulate the shallow and deep features from the vision foundation model, capturing both local details and global semantics. Plus, an Audio-Conditioned Feature Modulation is proposed to inject the audio context into the visual map. This design allows FSA to capture a more expressive spatial feature map. On the other hand, GSA, empowered by a Cross-Modally Gated Low-Rank Adaptation (CMG-LoRA) scheme,

introduces both an audio-to-video gate and a video-to-audio gate, making each low-rank update conditioned on the other modality.

We validate HAVADA on six widely-used audio-visual based video saliency datasets, namely, DIEM [31], AVAD [30], Coutrot1 [8], Coutrot2 [9], ETMD [23], and SumMe [12], following the standard protocols and metrics. Across multiple vision foundation models (i.e., CLIP [34], SigLIP [52], and DINOv2 [32]) and audio foundation models (i.e., CLAP [42], Wav2Vec [35], and HuBERT [15]), our approach achieves strong performance with high efficiency.

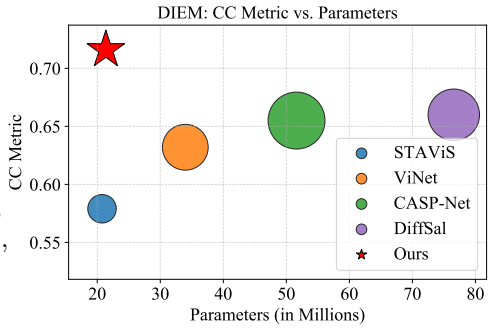


Fig. 2: Performance vs. efficiency comparison between the proposed method and existing methods. Y-axis: CC metric (higher is better); X-axis: Trainable parameter number (in millions); Bubble size corresponds to GFLOPs. Numerical results are reported in Table 1.

In a nutshell, our contributions can be summarized as follows.

- We propose Heterogeneous Audio and Vision Adaptation (HAVADA), a parameter-efficient framework that adapts frozen audio and vision foundation models for audio-vision based video saliency prediction.
- We propose Feature Space Adaptation (FSA) that modulates vision and audio features respectively to model the cross-modal context, and Gradient Space Adaptation (GSA) that imposes a cross-modally gated low-rank adaptation to enable cross-modal information flow.
- Extensive experiments on six benchmarks across multiple foundation models demonstrate strong performance and computational efficiency, with state-of-the-art results on multiple benchmarks.

2 Related Work

2.1 Video Saliency Prediction

Early deep video saliency methods primarily relied on convolutional neural network (CNN) [11,41] or recurrent neural network (RNN) [54] architectures to integrate spatial and temporal cues. Subsequent works explored the spatio-temporal encoder-decoder paradigm. For example, TASED-Net aggregated temporal context in a sliding-window manner with an S3D encoder [29]. UNISAL further unified image and video saliency modeling through shared representations and training strategies [10]. More recent works tend to use Transformer [50] and Mamba [13, 14] as the vision encoder. Video saliency is also predicted by using parametric images [50] and depth maps [25, 26].

2.2 Audio-Visual Video Saliency Prediction

Audio-visual saliency aims to model attention under joint auditory and visual stimuli. Representative works include ViNet [18] and its variations [39] that fuse visual spatiotemporal features with audio features, as well as consistency-aware designs such as CASP-Net [43] that emphasize audio-visual consistency in perception. Recent studies continue to improve fusion strategies under temporal inconsistencies [43]. Another recent research line starts to use generative modeling to map the audio feature distribution to the video feature distribution [44]. While both vision and audio foundation models have seen rapid development in the past few years, no work has systematically leveraged foundation models from both modalities for this task so far.

2.3 Foundation Models for Vision and Audio

Foundation models trained with large-scale weak supervision yield transferable representations. In vision, CLIP learns general visual semantics via image-text contrastive learning, and has shown strong transfer across tasks [34]. Some subsequent foundation models include Self-DIstillation with NO Labels (DINO) [6], Segment Anything (SAM) [20, 21] and their variations [5, 32]. In audio, CLAP provides robust audio embeddings through language-audio contrastive pretraining on large-scale audio-text pairs [42]. Meanwhile, some video, audio and joint audio-video foundation models have also been proposed for generative tasks [1, 40,45]. However, *to the best of our knowledge*, the direct adaptation of these foundation models to video saliency prediction has been underexplored. To constrain the task complexity, in this work, we mainly focus on CLIP [34] and CLAP [42] to validate the feasibility of heterogeneous adaptation.

2.4 Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) adapts large pretrained models by training a small parameter subset, which can avoid the high computational cost and possible model collapse by full tuning. In general, such methods can be categorized into three major categories, namely, prompt tuning [2, 55], side adapter [3, 47, 49, 51] and low-rank adaptation [16, 28, 36–38]. Generally speaking, prompt tuning based methods optimize continuous prompts while freezing base parameters. Adapters insert lightweight bottleneck modules between layers, enabling efficient task-specific tuning. LoRA and its variations inject low-rank updates into linear layers and has become a widely-adopted PEFT approach due to its simplicity and effectiveness. However, these methods are usually implemented on a single representation when fine-tuning. In contrast, this work intends to conduct audio-video representation interaction when applying the low-rank adaptation, which holds the particular challenge of audio-vision inconsistency. Regrettably, no systematic analysis has been covered on this aspect in the PEFT family.

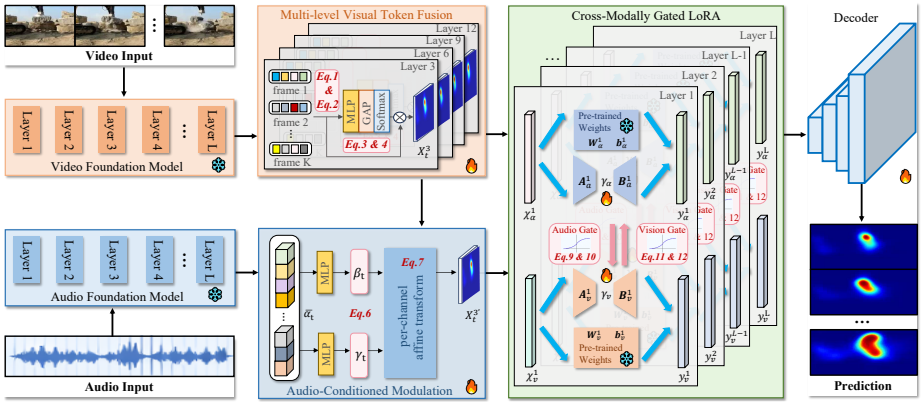


Fig. 3: Framework overview of the proposed Heterogeneous Audio and Vision Adaptation (HAVADA) approach for audio-video based video saliency prediction.

3 Method

Given a video clip centered at time t, let $\mathcal{I}_t = \{\boldsymbol{I}_{t-\Delta}, \ldots, \boldsymbol{I}_t, \ldots, \boldsymbol{I}_{t+\Delta}\}$ denote a small temporal window of frames, where $\boldsymbol{I} \in \mathbb{R}^{3\times H\times W}$. Let $\boldsymbol{w}_t$ denote an audio waveform segment associated with the same clip. Audio-visual saliency prediction aims to estimate a saliency map $\boldsymbol{S}_t \in \mathbb{R}^{H\times W}$ that approximates the ground-truth human attention density $\boldsymbol{G}_t$, typically derived from eye-tracking.

The proposed Heterogeneous Audio and Vision Adaptation (HAVADA) approach uses both a vision foundation model Φ_v (e.g., CLIP) and an audio foundation model Φ_a (e.g., CLAP) as encoders. As shown in Fig. 3, HAVADA consists of both Feature Space Adaptation (FSA) and Gradient Space Adaptation (GSA). FSA firstly uses a Multi-level Visual Token Fusion (MVTF) to obtain a spatial feature map $\boldsymbol{X}_t \in \mathbb{R}^{C\times h\times w}$ from selected layers of Φ_v; and then uses Audio-Conditioned Feature Modulation (AFM) to inject the normalized audio embedding $\bar{\boldsymbol{a}}_t \in \mathbb{R}^{d_a}$ into the spatial feature map $\boldsymbol{X}_t$, obtaining the refined spatial feature map $\boldsymbol{X}'_t$. GSA is realized by a cross-modally gated low-rank adaptation (CMG-LoRA) mechanism, which is inserted into the layers of both Φ_v and Φ_a. Finally, a lightweight UNet-style decoder Ψ maps $\boldsymbol{X}'_t$ to logits $\hat{\boldsymbol{S}}_t \in \mathbb{R}^{H\times W}$ and saliency values $\sigma(\hat{\boldsymbol{S}}_t)$.

3.1 Feature Space Adaptation

Vision foundation models are usually pre-trained to carry the global semantics, but the saliency prediction task requires a dense map that is spatially precise [19, 24]. Using tokens from deeper layers of a vision model often sacrifices precise localization, while using the tokens from layers close to the output lacks high-level semantics. On the other hand, audio cues are informative only for some samples and can exhibit weak audio-visual correlation. The proposed

Feature Space Adaptation (FSA) aims to estimate the cross-modal context by modulating both features in a divide-and-conquer manner.

Multi-level Visual Token Fusion (MVTF). To retain spatial localization while incorporating high-level semantics, MVTF is proposed to aggregate shallow and deep visual tokens. Let Φ_v denote the vision foundation model that produces hidden states $\{\boldsymbol{H}^{(\ell)}\}$ from the selected Transformer layers $\ell \in \mathcal{L}$. Each hidden state $\boldsymbol{H}^{(\ell)} \in \mathbb{R}^{T\times C}$ contains one `[CLS]` token and $T-1$ patch tokens. We remove the `[CLS]` token and reshape the patch tokens into a spatial map $\boldsymbol{M}^{(\ell)}$, given by:

$$\tilde{\boldsymbol{H}}^{(\ell)} = \boldsymbol{H}^{(\ell)}_{\text{patch}} \in \mathbb{R}^{(T-1)\times C}, \quad \boldsymbol{M}^{(\ell)} = \text{reshape}\big(\tilde{\boldsymbol{H}}^{(\ell)}\big) \in \mathbb{R}^{C\times h\times w}, \tag{1}$$

where C is the embedding dimension of the token, h is the height of the feature map, and w is the width of the feature map. Using a temporal window $\boldsymbol{I}_t$ with K frames where $K = |\boldsymbol{I}_t|$, the per-frame maps are averaged and computed as:

$$\bar{\boldsymbol{M}}^{(\ell)}_t = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{M}^{(\ell)}_{t,k}. \tag{2}$$

Then, we align channel statistics via 1×1 projections $\pi_\ell(\cdot)$ and predict per-sample fusion weights. Let $\boldsymbol{Z}^{(\ell)}_t = \pi_\ell(\bar{\boldsymbol{M}}^{(\ell)}_t) \in \mathbb{R}^{C\times h\times w}$. We compute a pooled descriptor $\boldsymbol{p}_t \in \mathbb{R}^C$ by averaging across layers and spatial positions, given by:

$$\boldsymbol{p}_t = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} \text{GAP}\big(\boldsymbol{Z}^{(\ell)}_t\big), \tag{3}$$

and predict fusion logits $\boldsymbol{u}_t \in \mathbb{R}^{|\mathcal{L}|}$ with an MLP f_θ via $\boldsymbol{u}_t = f_\theta(\boldsymbol{p}_t)$. Finally, the fused feature map is computed as:

$$\boldsymbol{X}_t = \sum_{\ell\in\mathcal{L}} \boldsymbol{\alpha}_t[\ell]\cdot \boldsymbol{Z}^{(\ell)}_t \in \mathbb{R}^{C\times h\times w}. \tag{4}$$

Here $\boldsymbol{\alpha}_t[\ell]$ refers to the weight for layer ℓ and $\boldsymbol{\alpha}_t = \text{softmax}(\boldsymbol{u}_t)$. This component converts the tokens from a vision foundation model into the dense spatial map, and fuses the multi-level representations with the learned weights, producing a localization-aware yet semantically rich feature map $\boldsymbol{X}_t$ for the subsequent audio-video representation interaction.

Audio-Conditioned Feature Modulation (AFM). AFM uses FiLM-style affine modulation [33] to inject global audio context while preserving the spatial layout of visual features. Let $\mathcal{P}_a$ denote the pre-processing operator (e.g., resampling, padding) that maps a raw waveform $\boldsymbol{w}_t$ to the input of the audio foundation model Φ_a. The audio embedding $\boldsymbol{a}_t$ is computed as:

$$\boldsymbol{a}_t = \Phi_a\big(\mathcal{P}_a(\boldsymbol{w}_t)\big) \in \mathbb{R}^{d_a}, \tag{5}$$

where d_a refers to the dimension of the audio embedding. ℓ_2 normalization is applied for stable conditioning, given by $\bar{\boldsymbol{a}}_t = \boldsymbol{a}_t/(\|\boldsymbol{a}_t\|_2 + \epsilon)$, where ϵ is a small scalar constant.

Given a fused visual feature map $\boldsymbol{X}_t \in \mathbb{R}^{C\times h\times w}$ and normalized audio embedding $\bar{\boldsymbol{a}}_t \in \mathbb{R}^{d_a}$, we first estimate the per-channel affine parameters via:

$$\boldsymbol{\gamma}_t = g_\gamma(\bar{\boldsymbol{a}}_t) \in \mathbb{R}^C, \qquad \boldsymbol{\beta}_t = g_\beta(\bar{\boldsymbol{a}}_t) \in \mathbb{R}^C, \tag{6}$$

where $g_\gamma(\cdot)$ and $g_\beta(\cdot)$ are small MLPs.

We then modulate the vision features $\boldsymbol{X}_t$ with the aid of the above two affine parameters, given by:

$$\boldsymbol{X}'_t = \boldsymbol{X}_t \odot \big(1 + \boldsymbol{\gamma}_t\big) + \boldsymbol{\beta}_t, \tag{7}$$

where $\odot$ denotes channel-wise multiplication with broadcasting over spatial dimensions, and 1 is a scalar. This provides a lightweight audio-to-vision feature space interaction without introducing heavy cross attention, yielding audio-aware spatial features $\boldsymbol{X}'_t$ for dense decoding.

3.2 Gradient Space Adaptation

Reliability is a key issue in joint audio-video interaction. Sometimes audio cues are strongly related to the visual attention, while in other cases they may be weakly related and therefore should be suppressed. A direct implementation of standard low-rank adaptation (LoRA) results in an independent adaptation of each foundation model, and cannot modulate the strength of joint audio-video adaptation. We use a scalar gate at each layer to control the overall cross-modal adaptation strength with negligible additional overhead.

For a layer at index ℓ in modality $m \in \{v, a\}$ with frozen base weight $\boldsymbol{W}_m^{(\ell)} \in \mathbb{R}^{d_o\times d_i}$, frozen bias $\boldsymbol{b}_m^{(\ell)} \in \mathbb{R}^{d_o}$, layer input $\boldsymbol{x}_m^{(\ell)} \in \mathbb{R}^{d_i}$, and layer output $\boldsymbol{y}_m^{(\ell)} \in \mathbb{R}^{d_o}$, the standard LoRA adds a low-rank update, given by:

$$\boldsymbol{y}_m^{(\ell)} = \boldsymbol{W}_m^{(\ell)}\boldsymbol{x}_m^{(\ell)} + \boldsymbol{b}_m^{(\ell)} + \Delta\boldsymbol{y}_m^{(\ell)}, \qquad \Delta\boldsymbol{y}_m^{(\ell)} = \boldsymbol{B}_m^{(\ell)}\big(\boldsymbol{A}_m^{(\ell)}\boldsymbol{x}_m^{(\ell)}\big), \tag{8}$$

where $\boldsymbol{A}_m^{(\ell)} \in \mathbb{R}^{r\times d_i}$ and $\boldsymbol{B}_m^{(\ell)} \in \mathbb{R}^{d_o\times r}$ are trainable matrices, and r is the rank. In the context of the proposed HAVADA, we have $\boldsymbol{H}_t^{(\ell)} := \boldsymbol{y}_v^{(\ell)}$ (in Eq. 1).

Audio Gate for Vision. For vision encoder layers, we scale the LoRA update by an audio-conditioned gate:

$$g_t^{v,(\ell)} = \sigma\big(\bar{\boldsymbol{a}}_t^\top \boldsymbol{q}_v^{(\ell)} + c_v^{(\ell)}\big) \in (0, 1), \tag{9}$$

where $\boldsymbol{q}_v^{(\ell)} \in \mathbb{R}^{d_a}$ and $c_v^{(\ell)} \in \mathbb{R}$ are trainable gate parameters, and $\sigma(\cdot)$ is the sigmoid function. The gated output becomes:

$$\boldsymbol{y}_v^{(\ell)} = \boldsymbol{W}_v^{(\ell)}\boldsymbol{x}_v^{(\ell)} + \boldsymbol{b}_v^{(\ell)} + g_t^{v,(\ell)} \cdot \Delta\boldsymbol{y}_v^{(\ell)}. \tag{10}$$

Vision Gate for Audio. Symmetrically, for audio encoder layers we construct a visual summary vector $\boldsymbol{v}_t$ via global average pooling, given by $\boldsymbol{v}_t = \mathrm{GAP}(\boldsymbol{X}_t) \in \mathbb{R}^C$, where $\mathrm{GAP}(\cdot)$ provides the global visual descriptor required by the scalar audio-side gate, while dense spatial features remain available for saliency decoding. Then, we devise an audio gate, defined as:

$$g_t^{a,(\ell)} = \sigma\big(\boldsymbol{v}_t^\top \boldsymbol{q}_a^{(\ell)} + c_a^{(\ell)}\big) \in (0,1), \tag{11}$$

where $\boldsymbol{q}_a^{(\ell)} \in \mathbb{R}^C$ and $c_a^{(\ell)} \in \mathbb{R}$ are trainable gate parameters.

Finally, the audio layer output is computed as:

$$\boldsymbol{y}_a^{(\ell)} = \boldsymbol{W}_a^{(\ell)} \boldsymbol{x}_a^{(\ell)} + \boldsymbol{b}_a^{(\ell)} + g_t^{a,(\ell)} \cdot \Delta \boldsymbol{y}_a^{(\ell)}. \tag{12}$$

Concisely, both gates scale the LoRA residual in the forward pass, which does not require iterative or recurrent cross-modal inference. During training, they also modulate the effective gradients on $\boldsymbol{A}_m^{(\ell)}$ and $\boldsymbol{B}_m^{(\ell)}$, yielding reliability-aware updates.

3.3 Implementation Details

By default, CLIP-ViT-B/16 [34] is used as the vision encoder. We extract hidden states from layers $\mathcal{L} = \{3, 6, 9, 12\}$, and reshape patch tokens into 14×14 maps at input resolution 224×224. CLAP [42] is used as the default audio encoder to extract normalized audio embeddings. The proposed HAVADA is also usable with other types of vision foundation models (e.g., SigLIP [52], DINOv2 [32]) and audio foundation models (e.g., Wav2Vec [35] and HuBERT [15]).

We use $K = 3$ frames as a lightweight local context and a clip-level audio embedding, while explicit temporal-order and onset modeling are beyond the present scope. The decoder maps the 14×14 CLIP feature map with $C = 768$ channels through channel widths $512 \to 256 \to 128 \to 64 \to 32$, using bilinear upsampling, 3×3 convolutions, LeakyReLU, and residual blocks. The final 224×224 stage concatenates a 1×1-projected skip from the original latent and outputs a one-channel saliency map by a 1×1 head. Unless otherwise specified, $r_v = r_a = 8$. AdamW uses a learning rate of 2×10^{-4}, weight decay of 10^{-4}, and gradient clipping of 1.0. The total loss is a sum of BCE-with-logits, KL divergence between normalized densities, correlation-coefficient loss, similarity loss, and Normalized Scanpath Saliency (NSS) loss.

4 Experiments

4.1 Datasets and Evaluation Protocols

Datasets. Six audio-visual datasets for video saliency prediction are employed for evaluation, which are listed below.

- **AVAD** [30] includes 45 short audio-visual clips depicting a wide range of activities such as piano performance, basketball playing, and interview scenes. Eye-tracking recordings are provided for 16 viewers.

- **Coutrot1** [8] and **Coutrot2** [9] are two subsets of the Coutrot collection. Coutrot1 offers 60 videos grouped into four visual scene types (e.g., single moving object, multiple moving objects, landscapes, and faces) and comes with gaze data from 72 subjects. Coutrot2 contains 15 meeting-style clips featuring four people, annotated with eye movements from 40 subjects.
- **DIEM** [31] is composed of 84 videos (e.g., trailers, music clips, commercials) viewed by 42 participants. A key property is that the soundtrack and imagery are not always naturally aligned.
- **ETMD** [23] provides 12 excerpts taken from Hollywood films, together with eye-tracking labels collected from 10 participants.
- **SumMe** [12] consists of 25 consumer-style videos spanning multiple everyday themes (e.g., sports, cooking, and tourism).

Evaluation Measures. Following previous video saliency prediction tasks [39,43], four standard video saliency prediction measures are computed between the predicted map and the ground-truth density [4], namely, Pearson correlation coefficient (CC), Normalized Scanpath Saliency (NSS), Structural Similarity Index (SIM), and AUC-Judd (AUCJ from here on).

4.2 Comparison with State-of-the-Art

The proposed method is compared with the state-of-the-art audio-video based video saliency prediction methods, namely, ACLNet [41] TPAMI'2019, STAViS [39] CVPR'2020, ViNet [18] IROS'2021, CASP-Net [43] CVPR'2023, DiffSal [44] CVPR'2024, and TAVDif [48] ArXiv'2025. For fair comparison, we follow the official evaluation protocols when available and use the same resolution and metrics. Table 1 reports their performance. The proposed method shows a state-of-the-art performance on at least three metrics on DIEM, Coutrot1, Coutrot2 and AVAD, respectively. Particularly, on DIEM, it achieves 73.4% CC, which is 9.55% higher than the second-best approach. On ETMD, it achieves the best AUCJ (0.957) and a competitive CC of 0.651, close to the best 0.652. On SumMe, it achieves 59.0% CC, which is 3.15% higher than the runner-up. In addition, it achieves the best AUCJ on ETMD (0.957) and maintains a competitive CC (0.651, very close to the best 0.652), while on SumMe it ranks best on both CC (0.590) and AUCJ (0.938), and is consistently within the top-two on NSS and SIM. Notably, it uses only 20.58M trainable parameters and 52.66 GFLOPs, substantially below DiffSal's 76.57M parameters and 187.31 GFLOPs.

4.3 Scalability to Different Foundation Models

To test the scalability of the proposed method on different vision foundation models and audio foundation models, we further test its performance when using the vision foundation model SigLIP [52] and DINOv2 (-Base) [32], and when using the audio foundation model Wav2Vec [35] and HuBERT [15]. The performance of all combinations is reported in Table 2. HAVADA shows a consistently

Table 1: Comparison with state-of-the-art methods on DIEM [31], Coutrot1 [8], Coutrot2 [9], AVAD [30], ETMD [23] and SumMe [12]. Higher (↑) is better. '-': Not reported in the corresponding paper. #Para.: Trainable parameter number. Top three results are highlighted as **best**, second and third, respectively. By default, the results of other methods are directly cited from DiffSal [44] and TAVDif [48].

Method	#Para.	#FLOPs	DIEM				Coutrot1				Coutrot2			
			CC↑	NSS↑	AUCJ↑	SIM↑	CC↑	NSS↑	AUCJ↑	SIM↑	CC↑	NSS↑	AUCJ↑	SIM↑
ACLNet [41]	-	-	0.522	2.02	0.869	0.427	0.425	1.92	0.850	0.361	0.448	3.16	0.926	0.322
STAViS [39]	20.76M	15.31G	0.579	2.26	0.883	0.482	0.472	2.11	0.868	0.393	0.734	5.28	0.958	0.511
ViNet [18]	33.97M	115.31G	0.632	2.53	0.899	0.498	0.560	2.73	0.889	0.425	0.754	5.95	0.951	0.493
CASP-Net [43]	51.62M	283.35G	0.655	2.61	0.906	0.543	0.561	2.65	0.889	0.456	0.788	6.34	0.963	0.585
DiffSal [44]	76.57M	187.31G	0.660	2.65	0.907	0.543	0.638	3.20	0.901	**0.515**	0.835	6.61	0.964	0.625
TAVDif [48]	-	-	0.670	2.75	0.909	0.547	0.607	2.85	0.892	0.486	0.798	6.52	0.963	0.594
HAVADA	20.58M	52.66G	**0.734**	**2.77**	**0.936**	**0.554**	**0.708**	**3.43**	**0.940**	0.498	**0.910**	**7.26**	**0.976**	**0.635**
Method	#Para.	#FLOPs	AVAD				ETMD				SumMe			
			CC↑	NSS↑	AUCJ↑	SIM↑	CC↑	NSS↑	AUCJ↑	SIM↑	CC↑	NSS↑	AUCJ↑	SIM↑
ACLNet [41]	-	-	0.580	3.17	0.905	0.446	0.477	2.36	0.915	0.329	0.379	1.79	0.868	0.296
STAViS [39]	20.76M	15.31G	0.608	3.18	0.919	0.457	0.569	2.94	0.931	0.425	0.422	2.04	0.888	0.337
ViNet [18]	33.97M	115.31G	0.674	3.77	0.927	0.491	0.571	3.08	0.928	0.406	0.463	2.41	0.897	0.343
CASP-Net [43]	51.62M	283.35G	0.691	3.81	0.933	0.528	0.620	3.34	0.940	0.478	0.499	2.60	0.907	0.387
DiffSal [44]	76.57M	187.31G	0.738	4.22	0.935	0.571	0.652	**3.66**	0.943	**0.506**	0.572	**3.14**	0.921	**0.447**
TAVDif [48]	-	-	0.729	4.29	0.949	0.550	0.613	3.15	0.937	0.446	0.483	2.44	0.902	0.372
HAVADA	20.58M	52.66G	**0.800**	**4.26**	**0.968**	**0.574**	**0.651**	3.31	**0.957**	0.448	**0.590**	2.85	**0.938**	0.398

Table 2: Scalability on different vision and audio foundation models on DIEM [31], Coutrot1 [8], Coutrot2 [9], AVAD [30], ETMD [23] and SumMe [12]. Higher (↑) is better. Top three results are highlighted as **best**, second and third, respectively. Notice that Wav2Vec and HuBERT have very similar numbers of parameters, and we do not observe clear differences on the computational cost metrics. #Para.: Trainable parameter number.

Vision	Audio	#Para.	#FLOPs	DIEM				Coutrot1				Coutrot2			
				CC↑	NSS↑	AUCJ↑	SIM↑	CC↑	NSS↑	AUCJ↑	SIM↑	CC↑	NSS↑	AUCJ↑	SIM↑
CLIP	CLAP	20.58M	52.66G	**0.734**	**2.77**	**0.936**	0.554	0.708	**3.43**	0.940	0.498	0.910	7.26	0.976	0.635
SigLIP	CLAP	33.05M	63.40G	0.701	2.60	0.935	0.519	0.667	3.06	0.939	0.462	0.872	6.30	0.976	0.574
DINOv2	CLAP	**20.14M**	36.42G	0.706	2.64	0.933	0.541	0.685	3.28	0.937	0.501	0.890	6.58	0.976	0.646
CLIP	Wav2Vec	20.97M	**30.63G**	0.723	2.72	**0.936**	0.548	0.707	3.42	0.942	0.508	**0.919**	**7.70**	**0.977**	**0.699**
CLIP	HuBERT	20.97M	**30.63G**	0.723	2.71	**0.936**	**0.556**	**0.713**	3.42	**0.944**	**0.529**	0.883	6.33	**0.977**	0.676
Vision	Audio	#Para.	#FLOPs	AVAD				ETMD				SumMe			
				CC↑	NSS↑	AUCJ↑	SIM↑	CC↑	NSS↑	AUCJ↑	SIM↑	CC↑	NSS↑	AUCJ↑	SIM↑
CLIP	CLAP	20.58M	52.66G	0.800	**4.26**	0.968	0.574	0.651	3.31	0.957	0.448	**0.590**	**2.85**	0.938	0.398
SigLIP	CLAP	33.05M	63.40G	0.763	3.86	0.967	0.545	0.625	3.10	0.956	0.409	0.561	2.62	0.938	0.363
DINOv2	CLAP	**20.14M**	36.42G	0.755	3.88	0.962	0.550	0.612	3.03	0.953	0.416	0.565	2.78	0.935	0.395
CLIP	Wav2Vec	20.97M	**30.63G**	0.803	4.23	0.969	0.607	0.654	**3.33**	0.957	0.452	0.589	2.84	0.940	0.403
CLIP	HuBERT	20.97M	**30.63G**	**0.808**	4.22	**0.971**	**0.634**	**0.655**	3.32	**0.958**	**0.456**	0.588	2.82	**0.941**	**0.408**

strong performance when used with different types of vision and audio foundation models. These results indicate that CLIP with CLAP or HuBERT provides a good trade-off between performance and computational cost.

4.4 Ablation Studies

Impact of Each Component. Apart from the video (V) and audio (A) foundation models, the two key components in HAVADA are Feature Space Adaptation

Table 3: Ablation of each component. Feature Space Adaptation (FSA) is realized by both Multi-level Visual Token Fusion (MVTF) and Audio-Conditioned Feature Modulation (AFM); Gradient Space Adaptation (GSA) is realized by Cross-Modally Gated Low-Rank Adaptation (CMG).

Index	V	A	FSA		GSA	AVAD				DIEM			
			MVFT	AFM	CMG	CC↑	NSS↑	AUCJ↑	SIM↑	CC↑	NSS↑	AUCJ↑	SIM↑
(i)	✓					0.771	3.85	0.961	0.541	0.701	2.62	0.917	0.524
(ii)	✓	✓				0.783	3.88	0.959	0.558	0.706	2.59	0.932	0.522
(iii)	✓	✓	✓			0.793	3.99	0.962	0.567	0.716	2.66	0.936	0.545
(iv)	✓	✓	✓	✓		0.796	4.19	0.960	0.563	0.722	2.70	**0.937**	0.547
(v)	✓	✓	✓	✓	✓	**0.800**	**4.26**	**0.968**	**0.574**	**0.734**	**2.77**	0.936	**0.554**

(FSA) and Gradient Space Adaptation (GSA). FSA handles the vision and audio features by Multi-level Vision Token Fusion (MVTF) and Audio-Conditioned Feature Modulation (AFM), respectively. GSA is empowered by the proposed Cross-Modally Gated Low-Rank Adaptation (CMG). Five settings are tested to validate each component's contribution, namely, (i) Vision Baseline: removes all the proposed component and only uses visual features; (ii) Vision and Audio Baseline: uses both features with a linear fusion and removes all the proposed component, which separates the effect of the proposed modules from the benefit of using foundation-model encoders; (iii) MVTF: uses both features with the MVTF module; (iv) MVTF & AFM: uses both features with the MVTF and AFM modules; and (v) Proposed: uses both features with MVTF, AFM and CMG modules. All variants use the same decoder and training protocol. Table 3 reports the performances. Each component contributes positively to the performance on all the involved datasets. Specifically, compared with the Vision Baseline (i), introducing audio (ii) improves CC/NSS/AUCJ/SIM from 0.771/3.85/0.961/0.541 to 0.783/3.88/0.959/0.558 on AVAD. Adding MVTF (iii) further brings consistent gains on both datasets, reaching 0.793/3.99/0.962/0.567 on AVAD and 0.716/2.66/0.936/0.545 on DIEM, in terms of CC/NSS/AUCJ/SIM. With AFM enabled (iv), the model becomes notably stronger on AVAD and DIEM, achieving the best DIEM AUCJ. Finally, equipping CMG (v) yields the overall best configuration, attaining the best performance on AVAD across all four metrics and the best DIEM CC/NSS/SIM, while keeping DIEM AUCJ competitive.

Impact of Video and Audio Modality. We further conduct ablations to quantify how each modality contributes to the final performance when adapting foundation models for audio-visual saliency prediction. Specifically, we compare four settings: (i) *Video-only FM*: freezing the video foundation model and disabling any audio stream; (ii) *Audio-only FM*: freezing the audio foundation model and removing the video stream; (iii) *Independent LoRA*: enabling both modalities but adapting the video and audio backbones independently with LoRA, without cross-modal interaction; and (iv) *Proposed*: jointly leveraging both modalities with our cross-modally coupled adaptation strategy. Table 4 summarizes the results. Overall, the video-only configuration is notably stronger

Table 4: Ablation of video and audio modalities. FZ: Frozen; FT: Fine-Tuned.

Vision		Audio		AVAD				DIEM			
FZ	FT	FZ	FT	CC↑	NSS↑	AUCJ↑	SIM↑	CC↑	NSS↑	AUCJ↑	SIM↑
✓				0.771	3.85	0.961	0.524	0.701	2.62	0.917	0.524
		✓		0.605	3.04	0.929	0.441	0.448	1.58	0.867	0.357
	✓	✓		0.799	4.25	**0.968**	0.556	0.726	2.75	0.927	0.546
✓		✓		0.783	3.88	0.959	0.558	0.706	2.59	0.932	0.522
	✓		✓	**0.800**	**4.26**	**0.968**	**0.574**	**0.734**	**2.77**	**0.936**	**0.554**

Table 5: Impact of LoRA variants.

Method	AVAD				DIEM			
	CC↑	NSS↑	AUCJ↑	SIM↑	CC↑	NSS↑	AUCJ↑	SIM↑
Frozen	0.783	3.88	0.959	0.558	0.706	2.59	0.932	0.522
VeRA [22]	0.794	4.23	0.967	0.542	0.721	2.71	0.924	0.538
DoRA [28]	0.798	4.18	0.967	**0.581**	0.723	2.62	0.934	0.534
LoRA [16]	0.796	4.23	**0.968**	0.574	0.729	2.73	**0.936**	0.548
Proposed (CMG-LoRA)	**0.800**	**4.26**	**0.968**	0.574	**0.734**	**2.77**	**0.936**	**0.554**

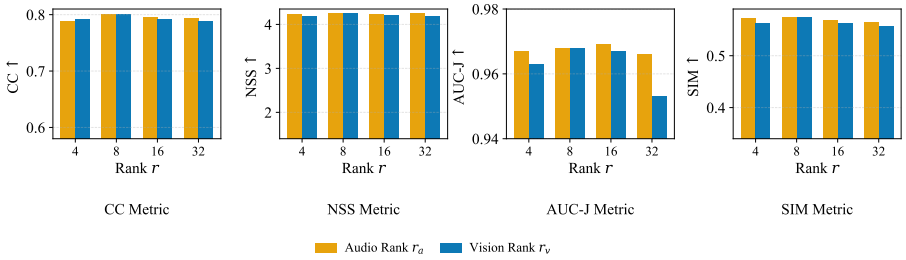


Fig. 4: Impact of r_a and r_v. All the experiments are conducted on AVAD.

than the audio-only counterpart, while combining video and audio consistently yields the best performance across datasets. Fig. 5 further provides qualitative comparisons under different modality inputs. When only using the audio features, only the approximated sound regions can be coarsely localized, resulting in unnecessary activations. By further leveraging the vision features, the sound regions can be more precisely activated.

Impact of LoRA Variants. We further study how different parameter-efficient adaptation strategies affect the overall performance. Beyond the fully frozen setting, we consider three LoRA-style variants that inject both the audio and video information, namely, LoRA [16], VeRA [22] and DoRA [28]. Table 5 reports the results. Overall, enabling adapter-based tuning consistently improves over the fully frozen model, and our cross-modally gated design yields the strongest performance across datasets, as it is modulated by the complementary modality (both from audio to video and from video to audio).

Impact of Rank Number. We further study how the rank number of the audio foundation model r_a and the rank number of the vision foundation model

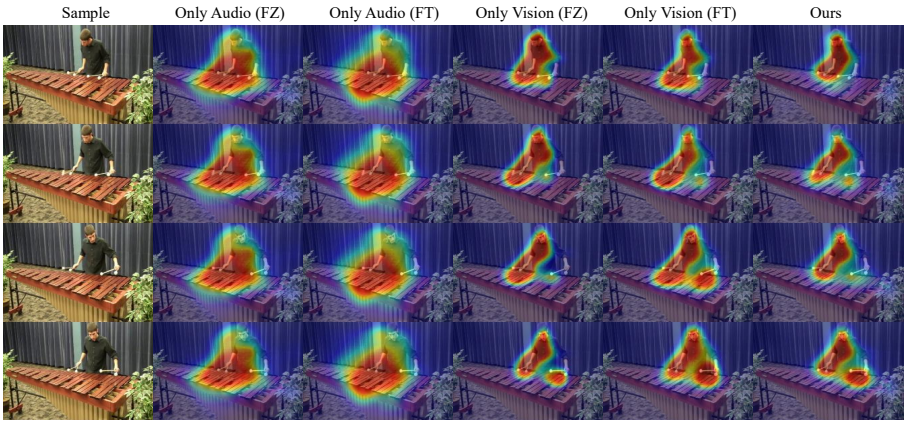


Fig. 5: Qualitative Results. Given the clips of a video, from the left to the right, we display the prediction maps when only using frozen audio foundation model, fine-tuned audio foundation model, frozen vision foundation mode, fine-tuned vision foundation model, and the proposed method. Experiments are conducted on AVAD dataset.

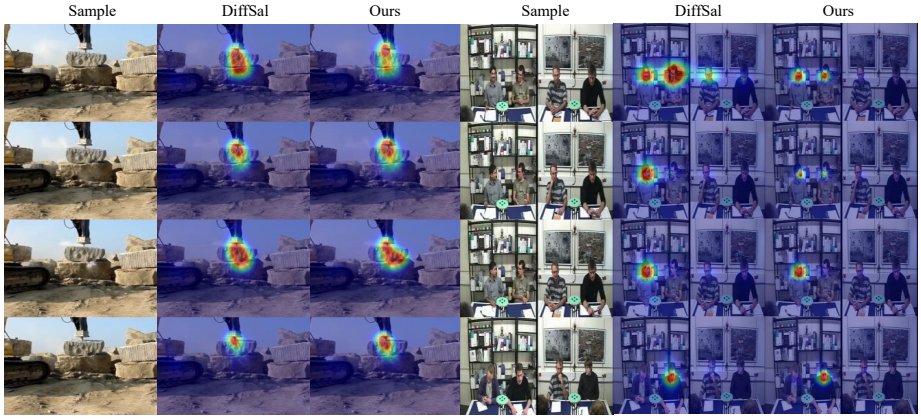


Fig. 6: Qualitative Results. Given the clips of a video from both indoor and outdoor scenes, we display the prediction maps between the state-of-the-art DiffSal [44] and the proposed method.

r_v impact the overall performance. By default both r_a and r_v are set to be 8. We further test the cases when they are set to be 4, 16 and 32, respectively. As shown in Fig. 4, the proposed HAVADA achieves the optimal performance when r_a and r_v are both set to be 8, which gives the best trade-off. Further increasing the rank number does not lead to a clear performance improvement but can instead increase the computational cost and lead to a potential performance drop.

4.5 Further Analysis

Visualization of Saliency Maps. We visualize predicted saliency maps and overlays to qualitatively assess whether our model attends to sound-producing regions more precisely. Fig. 6 compares the outcomes between the proposed method and the state-of-the-art DiffSal [44]. The proposed method shows a stronger ability to precisely perceive the salient regions in the videos. Specifically, in the first outdoor video, the proposed method is more focused on the interaction between the drill and the stone that causes the sound. In the second indoor video, the proposed method can better locate the mouth of the speaker in the first column and reduce the false alarm from the second column that is salient.

Feature Space Analysis. To understand whether the proposed HAVADA approach allows the audio and vision features to be better aligned over the baseline, we extract the audio and video embeddings after training and compute their correlation matrix. All the samples' embeddings from AVAD have been extracted. Fig. 7 compares the correlation matrix between the baseline (left) and the proposed approach (right), where a stronger red color indicates a higher correlation value (elements normalized to $[0, 1]$). More elements from the proposed method's correlation matrix have higher responses than those from the baseline, providing an illustrative view of improved audio-visual alignment.

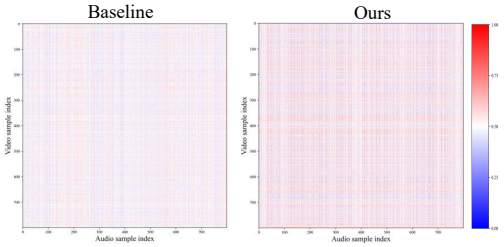


Fig. 7: The correlation matrix between the audio and video embedding. The correlation increases from blue to red.

5 Conclusion

We proposed Heterogeneous Audio and Vision Adaptation (HAVADA), a novel parameter-efficient fine-tuning framework for audio-visual video saliency prediction. The Feature Space Adaptation (FSA) enhanced the cross-modal context representation capability by both Multi-level Visual Token Fusion (MVTF) and Audio-Conditioned Feature Modulation (AFM). The Gradient Space Adaptation (GSA), empowered by a cross-modally gated low-rank adaptation (CMG-LoRA), realized strong task-specific adapting capability. Both components complement each other, allowing HAVADA to achieve the state-of-the-art performance and notable computational efficiency on six audio-visual benchmarks and across multiple types of foundation models. We hope this work opens a new direction for leveraging paired audio-vision foundation models in audio-visual saliency and other dense prediction tasks.

Acknowledgements

This work is supported by the NWA-ORC Project HAICu (NWA 1518.22.105).

References

1. Akman, A., Sun, Q., Schuller, B.W.: Audio explanation synthesis with generative foundation models. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5 (2025)
2. Bang, J., Ahn, S., Lee, J.G.: Active prompt learning in vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27004–27014 (2024)
3. Bi, Q., Yi, J., Huang, H., Zheng, H., Zhan, H., Huang, Y., Li, Y., Wu, X., Zheng, Y.: Nightadapter: Learning a frequency adapter for generalizable night-time scene segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23838–23849 (2025)
4. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(3), 740–757 (2018)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Chang, Q., Zhu, S.: Temporal-spatial feature pyramid for video saliency detection. arXiv preprint arXiv:2105.04213 (2021)
8. Coutrot, A., Guyader, N.: How saliency, faces, and sound influence gaze in dynamic social scenes. Journal of vision **14**(8), 5–5 (2014)
9. Coutrot, A., Guyader, N.: Multimodal saliency models for videos. In: From Human Attention to Computational Attention: A Multidisciplinary Approach, pp. 291–304. Springer (2016)
10. Droste, R., Jiao, J., Noble, J.A.: Unified image and video saliency modeling. In: European Conference on Computer Vision. pp. 419–435 (2020)
11. Ehteshami Bejnordi, B., Habibian, A., Porikli, F., Ghodrati, A.: Salisa: Saliency-based input sampling for efficient video object detection. In: European conference on computer vision. pp. 300–316 (2022)
12. Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: European conference on computer vision. pp. 505–520 (2014)
13. He, J., Fu, K., Liu, X., Zhao, Q.: Samba: A unified mamba-based framework for general salient object detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25314–25324 (2025)
14. Hosseini, A., Kazerouni, A., Akhavan, S., Brudno, M., Taati, B.: Sum: Saliency unification through mamba for visual attention modeling. In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1597–1607 (2025)
15. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM transactions on audio, speech, and language processing **29**, 3451–3460 (2021)

16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
17. Iashin, V., Rahtu, E.: Multi-modal dense video captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 958–959 (2020)
18. Jain, S., Yarlagadda, P., Jyoti, S., Karthik, S., Subramanian, R., Gandhi, V.: Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3520–3527 (2021)
19. Ji, W., Li, J., Bi, Q., Liu, J., Cheng, L., et al.: Promoting saliency from depth: Deep unsupervised rgb-d saliency detection. In: International Conference on Learning Representations (2022)
20. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F., et al.: Segment anything in high quality. Advances in Neural Information Processing Systems **36**, 29914–29934 (2023)
21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
22. Kopiczko, D.J., Blankevoort, T., Asano, Y.M.: Vera: Vector-based random matrix adaptation. In: The Twelfth International Conference on Learning Representations (2024)
23. Koutras, P., Maragos, P.: A perceptually based spatio-temporal computational framework for visual saliency estimation. Signal Processing: Image Communication **38**, 15–31 (2015)
24. Li, J., Ji, W., Bi, Q., Yan, C., Zhang, M., Piao, Y., Lu, H., et al.: Joint semantic mining for weakly supervised rgb-d salient object detection. Advances in Neural Information Processing Systems **34**, 11945–11959 (2021)
25. Li, J., Ji, W., Wang, S., Li, W., et al.: Dvsod: Rgb-d video salient object detection. Advances in Neural Information Processing Systems **36**, 8774–8787 (2023)
26. Lin, J., Zhu, L., Shen, J., Fu, H., Zhang, Q., Wang, L.: Vidsod-100: A new dataset and a baseline model for rgb-d video salient object detection. International Journal of Computer Vision **132**(11), 5173–5191 (2024)
27. Lin, W., Liu, H., Liu, S., Li, Y., Xiong, H., Qi, G., Sebe, N.: Hieve: A large-scale benchmark for human-centric video analysis in complex events. International Journal of Computer Vision **131**(11), 2994–3018 (2023)
28. Liu, S.Y., Wang, C.Y., Yin, H., Molchanov, P., Wang, Y.C.F., Cheng, K.T., Chen, M.H.: Dora: Weight-decomposed low-rank adaptation. In: International Conference on Machine Learning (2024)
29. Min, K., Corso, J.J.: Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2394–2403 (2019)
30. Min, X., Zhai, G., Gu, K., Yang, X.: Fixation prediction through multimodal analysis. ACM Transactions on Multimedia Computing, Communications, and Applications **13**(1), 1–23 (2016)
31. Mital, P.K., Smith, T.J., Hill, R.L., Henderson, J.M.: Clustering of gaze during dynamic scene viewing is predicted by motion. Cognitive computation **3**(1), 5–24 (2011)
32. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal (2024)

33. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
35. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: Unsupervised pre-training for speech recognition. In: Proc. Interspeech 2019. pp. 3465–3469 (2019)
36. Shen, Y., Bi, Q., Huang, J.H., Zhu, H., Pimentel, A.D., Pathania, A.: Macp: Minimal yet mighty adaptation via hierarchical cosine projection. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 20602–20618 (2025)
37. Shen, Y., Bi, Q., Huang, J.H., Zhu, H., Pimentel, A.D., Pathania, A.: Ssh: Sparse spectrum adaptation via discrete hartley transformation. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 10400–10415 (2025)
38. Shen, Y., Bi, Q., Wang, Z., Yang, Z., Wang, C., Zhang, Z., Tiwari, P., Pimentel, A.D., Pathania, A.: Efficient multimodal spatial reasoning via dynamic and asymmetric routing. In: The Fourteenth International Conference on Learning Representations (2026)
39. Tsiami, A., Koutras, P., Maragos, P.: Stavis: Spatio-temporal audiovisual saliency network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4766–4776 (2020)
40. Wang, H., Ma, J., Pascual, S., Cartwright, R., Cai, W.: V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 15492–15501 (2024)
41. Wang, W., Shen, J., Xie, J., Cheng, M.M., Ling, H., Borji, A.: Revisiting video saliency prediction in the deep learning era. IEEE transactions on pattern analysis and machine intelligence **43**(1), 220–237 (2019)
42. Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., Dubnov, S.: Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5 (2023)
43. Xiong, J., Wang, G., Zhang, P., Huang, W., Zha, Y., Zhai, G.: Casp-net: Rethinking video saliency prediction from an audio-visual consistency perceptual perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6441–6450 (2023)
44. Xiong, J., Zhang, P., You, T., Li, C., Huang, W., Zha, Y.: Diffsal: Joint audio and video learning for diffusion saliency prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27273–27283 (2024)
45. Yang, Q., Xu, J., Liu, W., Chu, Y., Jiang, Z., Zhou, X., Leng, Y., Lv, Y., Zhao, Z., Zhou, C., et al.: Air-bench: Benchmarking large audio-language models via generative comprehension. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1979–1998 (2024)
46. Ye, Q., Shen, X., Gao, Y., Wang, Z., Bi, Q., Li, P., Yang, G.: Temporal cue guided video highlight detection with low-rank audio-visual fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7950–7959 (2021)

47. Yi, J., Bi, Q., Zheng, H., Zhan, H., Ji, W., Huang, Y., Li, Y., Zheng, Y.: Learning spectral-decomposited tokens for domain generalized semantic segmentation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8159–8168 (2024)
48. Yu, L., Sun, X., Zhou, W., Gabbouj, M.: Text-audio-visual-conditioned diffusion model for video saliency prediction. arXiv preprint arXiv:2504.14267 (2025)
49. Yu, W., Liu, Y., Hua, W., Jiang, D., Ren, B., Bai, X.: Turning a clip model into a scene text detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6978–6988 (2023)
50. Yun, H., Lee, S., Kim, G.: Panoramic vision transformer for saliency detection in 360 videos. In: European Conference on Computer Vision. pp. 422–439 (2022)
51. Zaken, E.B., Goldberg, Y., Ravfogel, S.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 1–9 (2022)
52. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
53. Zhang, N., Xiong, M., Zhu, D., Zhu, K., Zhai, G., Yang, X.: Audio-visual saliency prediction model with implicit neural representation. ACM Transactions on Multimedia Computing, Communications and Applications **21**(4), 1–23 (2025)
54. Zhao, X., Liang, H., Li, P., Sun, G., Zhao, D., Liang, R., He, X.: Motion-aware memory network for fast video salient object detection. IEEE Transactions on Image Processing **33**, 709–721 (2024)
55. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)