DTFSal: Audio-Visual Dynamic Token Fusion for Video Saliency Prediction

Kiana Hooshanfar, Alireza Hosseini, Mona Ahmadian,
Ahmad Kalhor, Babak Nadjar Araabi
k.hooshanfar@ut.ac.ir

BMVC 2025

1. Motivation

Why Audio and Video Matter for Saliency:

- Human attention in dynamic scenes is driven by both sight and sound; relying on visuals alone misses salient events cued by audio.

Limits of Current Audio-Visual Models:

- Many AV saliency methods fuse modalities with heavy, complex attention or diffusion pipelines, yet still struggle with robust spatio-temporal reasoning and cross-modal alignment.

Computational Bottlenecks Block Deployment:

- State-of-the-art AV models are often large and computationally demanding (e.g., CASP-Net; DiffSal), constraining real-time or resource-limited use.

Need for a Balanced Solution:

- There is a clear gap for an AV saliency framework that (1) **captures long-range dependencies**, (2) **aligns audio–visual signals adaptively**, and (3) **delivers strong accuracy under tight compute budgets**.

3. Experiments

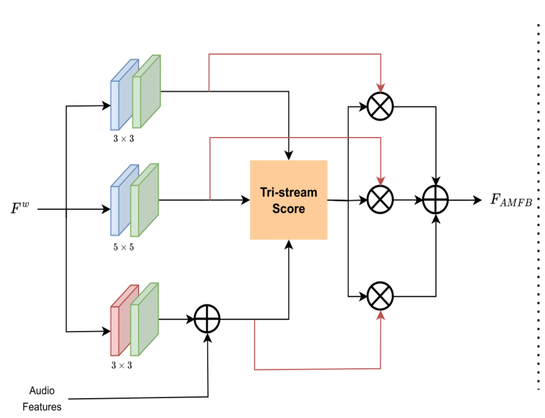
Datasets: 6 AV (DIEM, Coutrot1/2, AVAD, ETMD, SumMe) + 2 visual (DHF1K, UCF-Sports);
Results: new SOTA - 24/24 AV wins; DTFSal-V #1 vs visual SOTAs;
Efficiency: lighter-weight than DiffSal yet better.

Table 1. Comparison with previous methods on six audio visual saliency datasets.

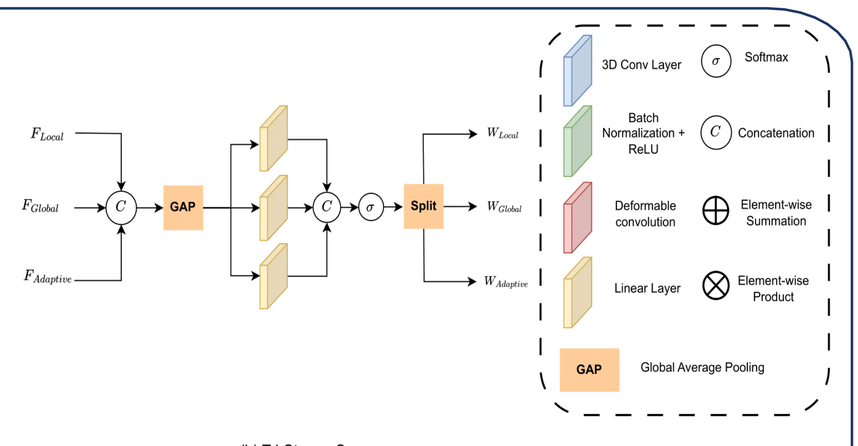
Method	#Parameters	DIEM				Coutrot1				Coutrot2			
		CC ↑	NSS ↑	AUC-J ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	SIM ↑
STAViS	20.76*M*	0.579	2.26	0.883	0.482	0.472	2.11	0.868	0.393	0.734	5.28	0.958	0.511
ViNet	33.97*M*	0.632	2.53	0.899	0.498	0.560	2.73	0.889	0.425	0.754	5.95	0.951	0.493
TSFP-Net	-	0.651	2.62	0.906	0.527	0.571	2.73	0.895	0.447	0.743	5.31	0.959	0.528
CASP-Net	51.62*M*	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	76.57*M*	0.660	2.65	0.907	0.543	0.638	3.20	0.901	0.515	0.835	6.61	0.964	0.625
MSPI	105.25*M*	0.671	2.71	0.905	0.545	0.605	2.98	0.899	0.473	0.799	6.40	0.964	0.5961
DAVS	-	0.580	2.28	0.884	0.483	0.481	2.19	0.869	0.400	0.734	4.984	0.960	0.512
DTFSal(V)	40.73*M*	0.677	2.76	0.910	0.548	0.623	3.21	0.904	0.505	0.842	6.42	0.965	0.618
DTFSal	49.08*M*	$0.682_{+1.64\%}$	$2.98_{+9.23\%}$	$0.921_{+1.54\%}$	$0.562_{+3.12\%}$	$0.661_{+3.61\%}$	$3.31_{+3.44\%}$	$0.911_{+1.11\%}$	$0.523_{+1.55\%}$	$0.845_{+1.20\%}$	$6.72_{+1.66\%}$	$0.971_{+0.73\%}$	$0.649_{+3.84\%}$
Method	**#Parameters**	**AVAD**				**ETMD**				**SumMe**			
		CC ↑	NSS ↑	AUC-J ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	SIM ↑
STAViS	20.76*M*	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	33.97*M*	0.674	3.77	0.927	0.491	0.571	3.08	0.928	0.406	0.463	2.41	0.897	0.343
TSFP-Net	-	0.704	3.77	0.932	0.521	0.576	3.07	0.932	0.428	0.464	2.30	0.894	0.360
CASP-Net	51.62*M*	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	76.57*M*	0.738	4.22	0.935	0.571	0.652	3.66	0.943	0.506	0.572	3.14	0.921	0.447
MSPI	105.25*M*	0.702	3.92	0.936	0.533	0.623	3.35	0.940	0.471	0.502	2.61	0.909	0.380
DAVS	-	0.610	3.19	0.919	0.458	0.600	2.96	0.932	0.425	0.422	2.28	0.888	0.338
DTFSal(V)	40.73*M*	0.742	4.21	0.938	0.559	0.653	3.57	0.939	0.499	0.561	2.91	0.911	0.441
DTFSal	49.08*M*	$0.746_{+1.09\%}$	$4.36_{+3.32\%}$	$0.939_{+0.32\%}$	$0.595_{+4.20\%}$	$0.667_{+2.30\%}$	$3.82_{+4.37\%}$	$0.957_{+1.48\%}$	$0.531_{+4.94\%}$	$0.573_{+0.175\%}$	$3.20_{+1.91\%}$	$0.927_{+0.65\%}$	$0.449_{+0.45\%}$

2. Proposed Method

- **AV saliency:** multiscale visual + audio.
- **Encoders:** MViTv2; log-Mel → VGGish + temporal Transformer.
- **Refinement:** top-down + lateral 3D convs.
- **Fusion/decoding:** AMFB (local/global/deformable) + tri-stream score; hierarchical decoders.
- **Tokens:** LTEB gates salient; DLTFB shifts/fuses for long-range.



(a) AMFB



(b) Tri-Stream Score

Figure 2. Overview of AMFB.

Figure 1. Overview of our SUM model.

Figure 3. Overview of LTEB and DLTFB.

3. Experiments - continued

Table 2. Visual-only baseline vs. visual SOTAs on DHF1K.

Model	DHF1K			
	CC ↑	NSS ↑	SIM ↑	AUC-J ↑
DeepVS	0.344	1.911	0.256	0.856
ACLNet	0.434	2.354	0.315	0.890
STRA-Net	0.458	2.558	0.355	0.895
TASED-Net	0.470	2.667	0.361	0.895
SalSAC	0.479	2.673	0.357	0.896
UNISAL	0.490	2.776	0.390	0.901
ECANet	0.500	2.814	0.385	0.903
HD2S	0.503	2.812	0.406	0.908
ViNet	0.511	2.872	0.381	0.908
TSFP-Net	0.517	2.967	0.392	0.912
GFNet	0.526	2.995	0.379	0.913
STSANet	0.529	3.010	0.383	0.913
TMFI-Net	0.546	3.146	0.407	0.9153
TinyHD-S	0.492	2.873	0.388	0.907
THTD-Net	0.547	3.138	0.406	0.9152
TempVST	0.494	3.73	0.373	0.904
ViNet-E	0.549	3.13	0.409	0.922
DiffSal(V)	0.533	3.066	0.405	0.918
HSFI-Net	0.548	3.163	0.411	0.914
DTFSal(V)	$0.561_{+2.37\%}$	$3.205_{+1.33\%}$	$0.436_{+6.08\%}$	$0.923_{+0.55\%}$

Table 3. FLOPs, input size, and model size vs prior visual & AV saliency methods.

Model	Input Size	FLOPs (G)	Model Size (MB)
Visual Models			
DeepVS	448×448	819.5	344
STRA-Net	256×224	55	324
TASED-Net	256×224	91.5	82
UNISAL	288×224	14.4	146
HD2S	192×128	22	155
ViNet	192×128	115	152
STSA-Net	384×224	493	245
TMFI-Net	384×224	358	234
THTD-Net	384×224	-	220
TM2SP-Net	384×224	1330	700
ViNet-E	384×224	-	183.84
HSFI-Net	384×224	-	253
DTFSal(V)	384×224	259.57	155.37
Audio-Visual Models			
ACLNet	224×224	164	250
TSFP-Net	352×192	316	192
STAVis	384×224	15.31	79.19
Casp-Net	384×224	283.35	196.91
DiffSal	384×224	834	351
DTFSal	384×224	297.91	187.26

For more Information:

SCAN ME

4. Visualizations

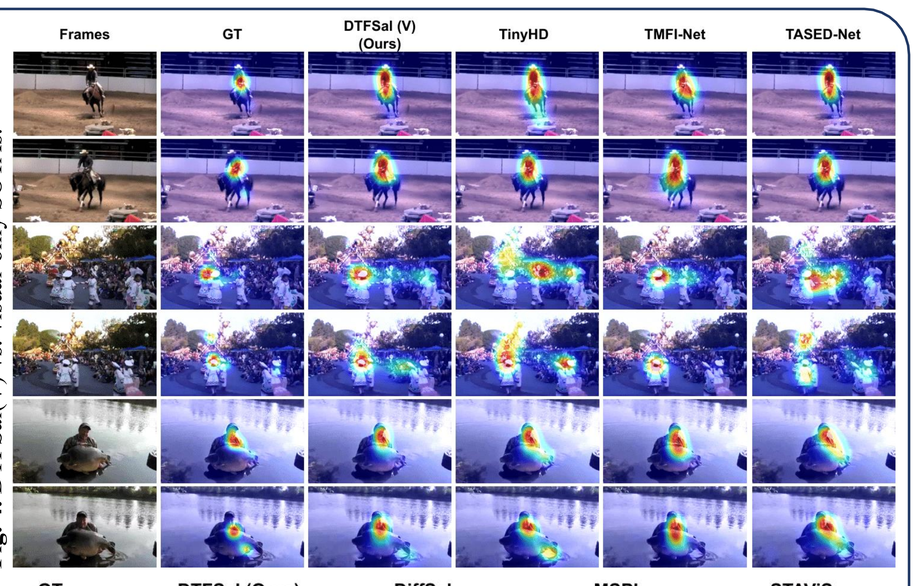


Fig. 4. DTFSal(V) vs. visual-only SOTAs.

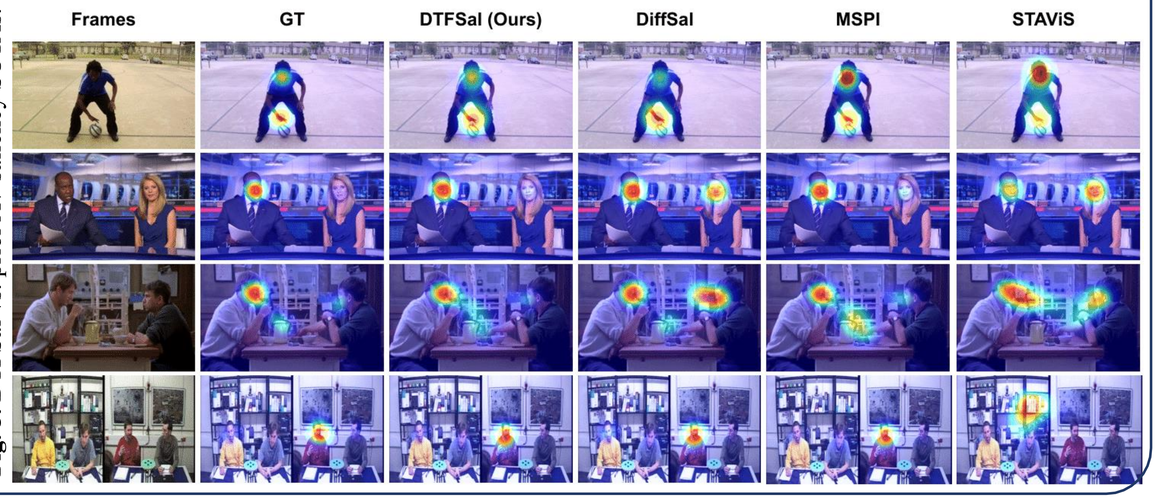


Fig. 5. DTFSal vs. prior AV-saliency SOTAs.