SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Weakly Supervised Visual-Auditory Fixation Prediction with Multigranularity Perception

Guotao WANG[1], Chenglizhao CHEN[2,3,4*], Deng-Ping FAN[5,6,7], Aimin HAO[1] & Qinping ZHAO[1]

[1]*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China*
[2]*College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China*
[3]*Qingdao Institute of Software, China University of Petroleum (East China), Qingdao 266580, China*
[4]*Shandong Provincial Key Laboratory of Intelligent Oil and Gas Industrial Software, Qingdao 266580, China*
[5]*Nankai International Advanced Research Institute (Shenzhen Futian), Shenzhen 518045, China*
[6]*Tianjin Visual Computing and Intelligent Perception Laboratory, Nankai University, Tianjin 300071, China*
[7]*College of Computer Science, Nankai University, Tianjin 300071, China*

This supplementary material is organized into six parts:

- **Appendix A:** Extending SCAM+ beyond video category tags.
- **Appendix B:** Conventional attention compared with our approach.
- **Appendix C:** Human-eye fixation experiments on the effect of multi-granularity information.
- **Appendix D:** Discussion on the rationale of multi-granularity information.
- **Appendix E:** Qualitative comparisons.
- **Appendix F:** Supplementary experiments and ablation studies.

For clarity, we present the supplementary content in the above order.

Appendix A Extending SCAM+ Beyond Video Category Tags

While SCAM+ utilizes video category tags during the training phase, the final STANet+ model, trained through knowledge distillation, does not require category tags for fixation prediction at inference time. This ensures that our method remains applicable even when category tags are unavailable in real-world deployment.

However, the availability of category tags during training could still limit broader applicability, particularly in scenarios where such annotations are incomplete or absent. To mitigate this issue, we propose an extension of our method that enables training without relying on video category tags.

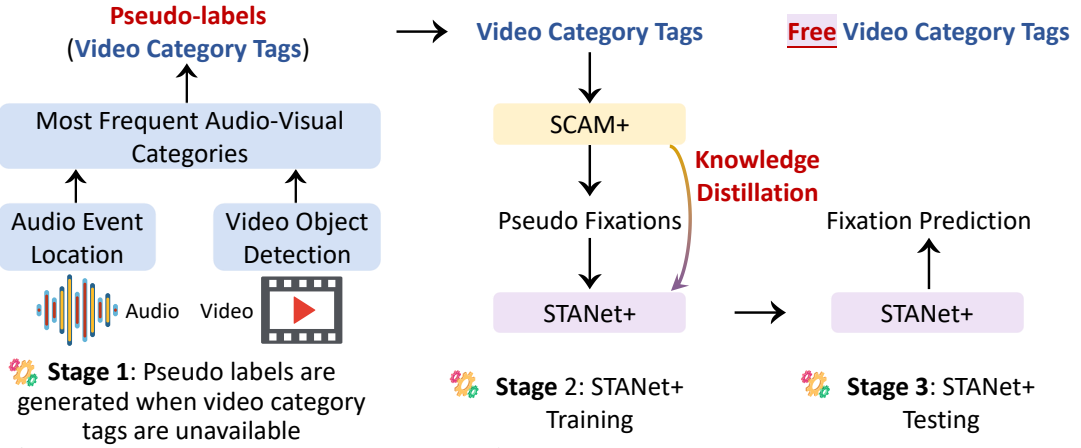

Figure A1 Extending our method to handle the absence of video category tags. When video category tags are unavailable during training, our method can be extended into three stages: *Stage 1* generates pseudo labels using existing methods [1, 2], *Stage 2* employs SCAM+ for weakly supervised pseudo fixation generation to train STANet+, and *Stage 3* enables fixation prediction without relying on category tags. This extension enhances generalization and applicability in real-world scenarios.

* Corresponding author (email: cclz123@163.com)

As illustrated in Figure A1, our method can be adapted into three stages:

Stage 1: Pseudo-Label Generation

When category tags are unavailable, pseudo labels can be generated using alternative weak supervision techniques, such as object detection-based saliency estimation [1] or audio-visual event localization [2]. These pseudo labels serve as substitutes for category tags, providing a weak supervisory signal.

Stage 2: Weakly Supervised Fixation Generation

SCAM+ leverages the pseudo labels to generate weakly supervised pseudo-fixations, mimicking the role of category tags in standard training. This allows us to train STANet+ effectively, even in the absence of manually annotated category labels.

Stage 3: Category-Tag-Free Fixation Prediction

Once trained, STANet+ functions independently of category tags. During inference, it directly predicts fixations based on the given video input, ensuring broader generalization to real-world settings where category annotations are unavailable.

Appendix B Conventional attention vs. our approach

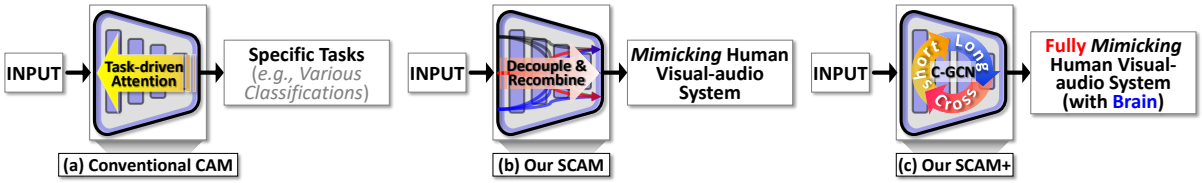

Figure B1 Illustrations of the differences between conventional attention and our novel approaches (SCAM and SCAM+).

Our approach differs from typical attention methods, including spatial attention [3–5] and spatially adaptive/dynamic models [6,7]. We have provided an illustration in Figure B1. Here we shall respectively detail each of these 'attentions'.

1) Conventional CAM (Figure B1(a)): This simple computation logic localizes image regions contributing most to the given classification task using attention operations.

2) Our SCAM (Figure B1(b)): The attention operation used in our SCAM differs from the conventional one. We decouple multi-modality information into S, SA, and ST in advance. The CAMs from each modality are selectively combined to generate the SCAM, compressing less discriminative regions in the process. Our SCAM can better localize the most discriminative image regions, aligning with the human visual-audio system.

3) Our SCAM+ (Figure B1(c)): Our SCAM+ fully mimics the real human visual-audio system and considers passive activities in the human brain. Although the basic attention operations in SCAM+ are generally the same as in SCAM, there is one critical difference. SCAM+ incorporates multi-granularity information, resulting in a more complex input with three types of input (specifically short-/long-/cross-term info). To achieve mutual compression between these inputs and localize the most discriminative image regions like SCAM, SCAM+ performs C-GCN-based semantic reasoning.

In summary, the basic computation logic of our SCAM and SCAM+ is similar to conventional attention, but their inference procedures and detailed usages differ due to their distinct objectives.

Appendix C Experiments to show how multigranularity information affects human-eye fixations

C.1 Instruments, Participants, and Settings

We use **Tobii Eye-tracker** to collect binocular fixations. The display size is 23" with a resolution of 1920 × 1080, where the screen length and width are 50.92cm × 28.64cm. To conduct this experiment, we employed 6 participants (4 males and 2 females), whose average age is 26, and, to ensure an objective process, all participants were complete noobs regarding fixation collection, and they didn't know the purpose of our experiment in advance. All participants' heads were fixed with a pillow to prevent any head shaking, and, following a typical setting [8], the distance between human-eye and screen was set to 68cm. To prevent visual fatigue, the experiment time of each participant is strictly limited to 60 minutes.

Besides, we have followed the product manual to conduct eye tracker calibration, and this process requires participants to randomly look at a total of 9 calibration positions on the screen. The calibration program will score the quality of fixation, namely as high-quality or low-quality. We do not carry out relevant experiments until the calibration procedure has reached high quality, and the observation time for each picture is about 3 seconds.

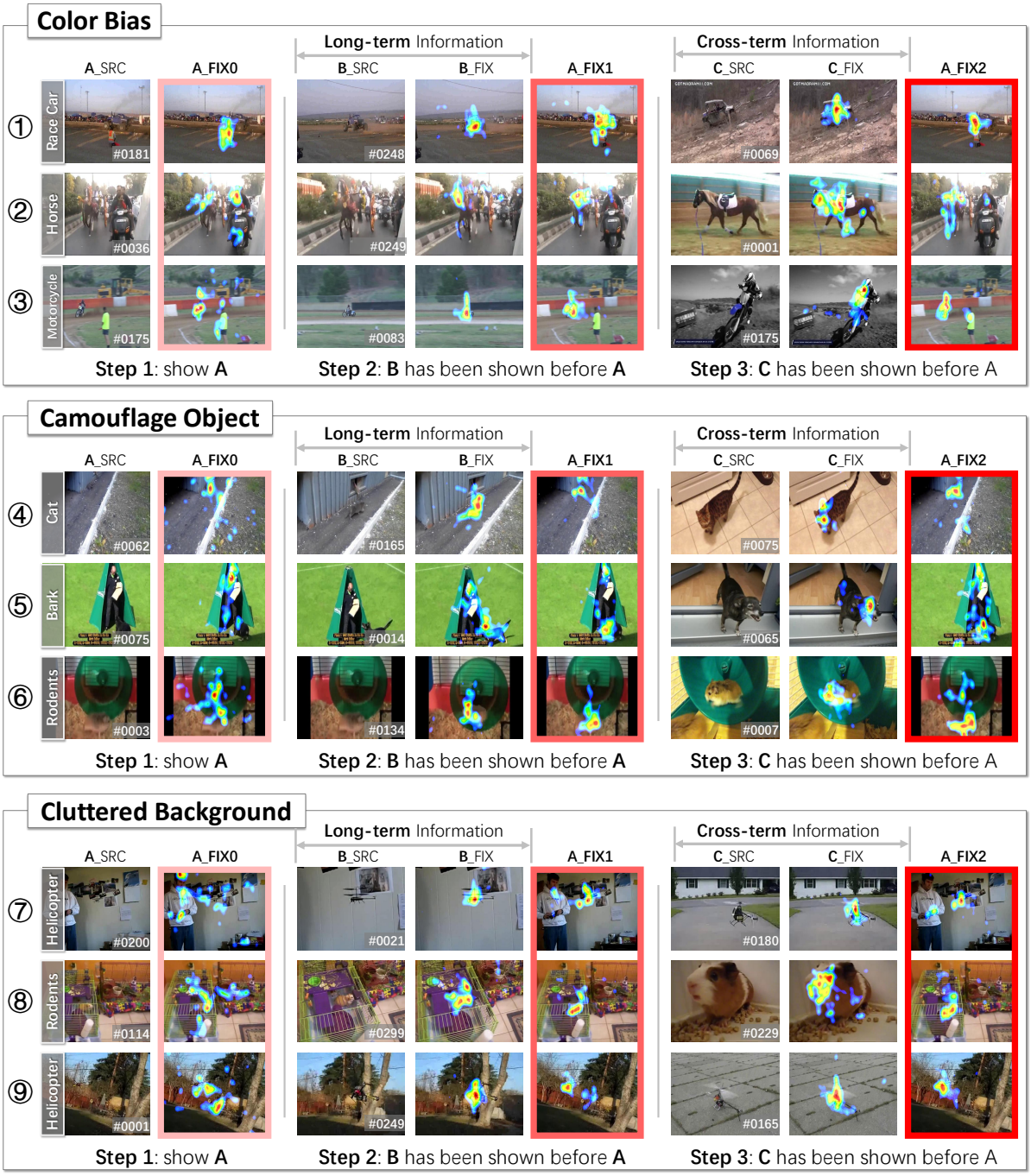

Figure C1 Some representative 'fixation shifting' cases, where, in step 2 and step 3 (Appendix C.2), additional multigranularity information (namely long/cross-term information) has been shown before collecting fixations in A_SRC. Clearly, by comparing A_FIX0, A_FIX1, and A_FIX2, we can easily notice that the multigranularity information could draw human attention to the most meaningful objects and make the fixations to be more focused.

C.2 Experiment Steps

To verify the effectiveness of multigranularity information, we divide our experiment into 3 sequential steps, and each participant will undergo all these steps. The details of these steps will be given as follows, and some qualitative demonstrations can be seen in Figure C1:

Step 1:

1) Showing frame A_SRC to each participant for 3 seconds and collecting the corresponding fixations.

2) The obtained raw fixations (as coordinates) will be post-processed (Appendix C.3) to obtain visual gaze result A_FIX0.

3) Each participant will take a break (3 minutes) to prevent eye fatigue.

Step 2:

1) Showing frame B_SRC, the long-term information, for 3 seconds, and collecting the corresponding fixations, then formulating the visual gaze result B_FIX.

2) Showing frame A_SRC immediately for 3 seconds, and obtaining the corresponding visual gaze A_FIX1.

3) Taking a break for 3 minutes.

Step 3:

This step is quite similar to step 2, where the major difference is that frame C_SRC, the cross-term information, has replaced frame B_SRC. Thus, we can obtain a completely new fixation map (A_FIX2) after observing A_SRC.

During the entire experiment, we will immediately stop the experiment and wait for the participants to rest whenever the participant reaches visual fatigue.

C.3 Post-processing

To convert the discrete coordinates to a continuous fixation map, we resort to the GaussBlur function which ensures the obtained fixation map to be spatially smoothing. The detailed operation process of the GaussBlur function is as follows.

$$\mathbf{OUT} = \mathcal{Z}\Big(Conv\Big(Conv\big(\mathbf{IN}, GK\left(\mathrm{Kernel}_w\right)^\top\big), GK\left(\mathrm{Kernel}_h\right)\Big)\Big), \tag{1}$$

where **IN** is a matrix containing coordinates (x, y); $\top$ is the matrix transpose operation; $GK(\cdot)$ represents a Gaussian Kernel generation function, which generates a $1\times w$ or $1\times h$ matrix according to the value of Kernel_w or Kernel_h, and we set w=10, h=10; $Conv(\cdot)$ represents a smoothing operation using kernel generated by $GK(\cdot)$; $\mathcal{Z}(\cdot)$ represents the min-max normalization function, and **OUT** is the smoothed fixation map.

Appendix D Discussion regarding the rationale of multigranularity information

To demonstrate the effectiveness of multigranularity information (as long/cross-term info.) in effecting human-eye fixations, we provide 3 most representative examples, where human-eye fixations are more likely to get shifted.

D.1 Color Bias Cases

The real human-eye can be easily attracted by regions with large color contrast to nearby surroundings. As shown in Figure C1-row ①, ②, and ③, fixations in A_FIX0 are mainly localized in regions with warm colors. However, the color bias prior could become invalid if multigranularity information has been shown in advance.

As shown in step 1, specifically A_SRC and A_FIX0 in ①, we can see that the squatting person attracts the most fixations, and the monster truck is completely non-salient. In step 2, since the long-term information containing a salient monster truck (B_SRC) has been shown in advance, the fixations of A_SRC are changed from A_FIX0 to A_FIX1, where the original non-salient monster truck becomes the most eye attractor. This case clearly demonstrates the effectiveness of long-term information in shifting human-eye fixations.

A similar tendency can be found in step 3, where the cross-term information could also cause fixation shifting, and the semantic coherency — human-eye tend to focus on objects with similar semantic information to the precedent ones over time, is the main reason that explains this phenomenon.

D.2 Camouflage Object Cases

In some cases, objects may have similar colors, patterns, and textures to surroundings, making human-eye omit them. As shown in Figure C1-row ④, ⑤, and ⑥, fixations in A_FIX0 are either localized in clear objects or scattered over backgrounds.

As expected, by comparing A_FIX0 with A_FIX1 and A_FIX2 in row ④, we can easily notice that fixations are more focused on the gray cat since both long-term and cross-term information adopted in step 2 and 3 have included salient cats which push human-eye to be more infested to regions with identical semantic information.

Similar phenomenons can be seen in ⑤ and ⑥, where human-eye fixations are scattered over multiple regions. After introducing long-/cross-term information, human-eye fixations become more focused, and even the camouflaged dog in ⑤-A_SRC can receive intense fixations.

D.3 Cluttered Background

In general, human eyes tend to sweep over the entire frame when it contains cluttered backgrounds, resulting in fixations scattered across multiple regions. Even non-salient background areas may receive intensive fixations, for example, A_FIX0 in row ⑦, ⑧, and ⑨. Actually, this phenomenon is quite similar to that occurred in camouflage object cases, while the behind rationale is different.

In the camouflage object cases, human-eye might omit the most meaningful object completely, and the multigranularity information could dramatically draw our attention back. However, in the cluttered background cases, the main reason for causing human-eye fixations scattering is that the given scene contains large content, which requires human-eye to spend more effort to locate those salient regions where the most meaningful object can receive intensive fixations eventually. Therefore, the main function of multigranularity information in such cases is further to focus human attention on those most meaningful objects (see A_FIX0, A_FIX1, and A_FIX2 in row ⑦).

D.4 Brief Summary

Through the above representative 'fixation shifting' cases caused by long/cross-term information, we can conclude that the multigranularity information has a strong auxiliary effect on human attention. Therefore, it is reasonable to introduce multigranularity information to facilitate the formulation of pseudo-fixations.

Appendix E Qualitative Comparisons

To further demonstrate the superiority of our proposed method, Figure D1 depicts the visual results of our method and the other 8 most representative weak/unsupervised saliency prediction models, ITTI [10], GBVS [12], SCLI [11], PS [14], AWS-D [9], WSS [16], MWS [15], WSSA [13].

We have selected multiple scenarios with various challenging properties. However, in complex cases, spatial, temporal, and audio cues should be considered jointly. Consequently, the results of SOTA methods may deviate from real fixations, since these methods do not incorporate the audio modality. In some scenarios, audio can be the dominant saliency cue, for example, in audio-dominant cases.

It is important to emphasize that, compared to other weakly supervised methods, our SCAM+ demonstrates better alignment with real human fixation, even in scenes with large objects, occlusions, and low contrast. This is due to the multi-modality and multi-granularity framework, which is further enhanced by a semantic reasoning mechanism. As shown in Fig D1, qualitative results reveal that SCAM+ generates more human-like fixation.

Appendix F Supplementary Experiment

F.1 The effectiveness of the divide-and-conquer scheme adopted by the proposed SCAM

To further verify the effectiveness of the divide-and-conquer scheme adopted by the proposed selective fusion, we compare the SCAM with the CAM result directly obtained from a three-stream spatial-temporal-audio classification net. We use '$\mathrm{CAM_{sta}}$' to represent the CAM result obtained from this three-stream network, where we simply adopt a three-stream network adopting an identical fusion unit to the proposed STA fixation prediction network, while the decoder part has been replaced by a multiclass classifier that is identical to other classification nets. The quantitative comparison result can be seen in Table F1, where, for both coarse and fine stages, the proposed SCAM can outperform $\mathrm{CAM_{sta}}$.

Table F1 Quantitative evidence towards the effectiveness of the proposed SCAM. This experiment is conducted on the AVAD set [17]. CO.: Coarse stage; FI.: Fine stage.

	Modules	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑		Modules	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
Coarse	CAMsta	0.793	0.227	0.551	0.290	1.273	**Fine**	CAMsta	0.856	0.296	0.579	0.420	1.803
	SCAM	**0.801**	**0.256**	**0.554**	**0.350**	**1.364**		**SCAM**	**0.873**	**0.334**	**0.580**	**0.440**	**2.018**

F.2 Thresholds Used in FRL

There exist 2 predefined thresholds in the proposed FRL, specifically the $\mathcal{T}_d$ and $\mathcal{T}_r$, which could directly influence the overall performance, and we will conduct an ablation study on the exact choices. We retrain the STA+ net multiple times using different $\{\mathcal{T}_d, \mathcal{T}_r\}$ combinations, and the corresponding results can be found in Table F2, where the 1st row denotes the result without using the FRL.

Since 2 thresholds should be determined, we shall adopt the one-fixed-another-solved strategy, in which we empirically assign $\mathcal{T}_d = 0.8$ and $\mathcal{T}_r = 0.6$ as our initial choice, and we have tested several other choices. Coincidentally, the initial choice turns out to be the best, and the main reason could be the fact that

Figure D1 Qualitative results of our method and eight representative saliency models: AWS-D [9], ITTI [10], SCLI [11], GBVS [12], WSSA [13], PS [14], MWS [15], WSS [16]. It can be observed that our method is able to handle various challenging scenes well and produces more accurate results than other competitors.

Table F2 Quantitative evidence towards the effectiveness of the FRL (frame refine layer), the $\mathcal{T}_d$ is the threshold for choosing the refine operation or not, the $\mathcal{T}_r$ is a suppressor for refining. The detailed description of $\mathcal{T}_d$ and $\mathcal{T}_r$ can be found in Appendix B

FRL		AVAD			DIEM			SumMe			ETMD			Coutrot 1			Coutrot 2		
$\mathcal{T}_d$	$\mathcal{T}_r$	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑
w/o FRL		0.856	0.436	1.857	0.835	0.482	1.799	0.844	0.330	1.598	0.877	0.402	1.891	0.815	0.301	1.302	0.832	0.343	1.988
0.8	0.4	0.860	0.426	1.671	0.859	0.504	1.832	0.849	0.335	1.602	0.882	0.409	1.915	0.812	0.303	1.309	0.829	0.329	1.990
0.8	0.8	0.843	0.393	1.802	0.844	0.511	1.815	0.853	0.341	1.609	0.881	0.411	1.923	0.824	0.308	1.311	0.826	0.340	1.984
0.9	0.6	0.868	0.421	1.790	0.865	0.498	1.856	0.848	0.332	1.611	0.875	0.399	1.921	0.829	0.309	1.309	0.834	0.337	1.988
0.7	0.6	0.849	0.425	1.699	0.853	0.487	1.862	0.841	0.340	1.600	0.901	0.420	1.952	0.825	0.312	1.315	0.840	0.343	1.992
0.8	**0.6**	**0.886**	**0.467**	**2.156**	**0.877**	**0.518**	**1.955**	**0.860**	**0.367**	**1.633**	**0.903**	**0.423**	**2.056**	**0.833**	**0.330**	**1.326**	**0.858**	**0.371**	**2.105**

the technical details of other major components are all implemented based on this initial choice, which

makes it passively outperform all other choices.

Table F3 Quantitative comparison of STATransformer, STAGRU, STARNN and STANet+, this experiment is conducted on the AVAD set [17].

Modules	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	Modules	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
STATransformer	0.844	0.319	0.548	0.411	1.972	STARNN	0.857	0.315	0.557	0.442	1.979
STATransformer (w FRL)	0.883	0.337	0.572	0.457	2.066	STAGRU	0.868	0.307	0.555	0.437	1.965
STANet+	**0.887**	**0.361**	**0.595**	**0.490**	**2.318**	**STANet+**	**0.887**	**0.361**	**0.595**	**0.490**	**2.318**

As shown in Table F3, since our C-GCN-based variant incorporates the newly devised **F**ixation **R**efine **L**ayer (FRL), we also evaluate a Transformer-based counterpart equipped with FRL for a fair comparison (see Figure F2, bottom). Table F3 shows that our C-GCN-based variant consistently outperforms the Transformer-based variants, particularly STATransformer without FRL.

We have also tested both RNN-based and GRU-based versions with FRL, and the quantitative results have been given in Table F3. Since RNN-based and GRU-based versions cannot handle the adopted multi-granularity information well. Thus, these versions have demonstrated poor numeric results.

F.3 Variants whose C-GCN have been replaced by RNNs, GCN, and Transformers

We tested multiple variants whose C-GCN has been replaced by RNNs, GCN, and Transformers. As reported in Table F4, the one with the best performance among the Transformer-based variants places at the same performance level as our C-GCN-based version, and the GCN- and RNNs-based version is clearly inferior to our C-GCN-based version. Since we have tested multiple Transformer-based variants, we detail their implementations/architectures in Figure F1. In our implementation, we have respectively tested both singlehead and multihead Transformers, each of which has two different realizations, specifically {**(a)**, **(c)**} *v.s.* {**(b)**, **(d)**}, where **(c)** is the MultiHead version of **(a)**, and **(d)** is the MultiHead version of **(b)**. An ablation study for verifying the number of Encoder Blocks (Figure F1-**(e)**) is introduced. As reported in Table F4, the performance of the singlehead Transformer-based variant with 3 encoder blocks is comparable to our C-GCN-based model with 3-step reasoning. The Transformer might require additional learning parameters, specifically almost 52 times the model size and 5.6 times FLOPs than C-GCN.

Table F4 Ablation study on semantic reasoning stage number m and MG node number n, on the AVAD set [17]. *w/o* Rea.: without Reasoning. We tested multiple variants whose C-GCN has been replaced by RNNs, GCN, and Transformers.

Modules and Variants		AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	Modules and Variants		AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
w/o Reasoning		0.825	0.253	0.556	0.338	1.519	w/o Reasoning		0.825	0.253	0.556	0.338	1.519
Reasoning Stage=m	m=2	0.873	0.316	0.575	0.442	1.98	TA=1, MG=n	n =1	0.871	0.319	0.58	0.432	1.965
	m=3	**0.886**	**0.336**	**0.596**	**0.467**	**2.156**		n =2	**0.886**	**0.336**	**0.596**	**0.467**	**2.156**
	m=4	0.872	0.324	0.58	0.451	1.981		n =3	0.867	0.306	0.578	0.424	1.946
	m=5	0.87	0.309	0.576	0.439	1.967		n =4	0.852	0.289	0.565	0.415	1.863
A-SingleHead Realization		0.848	0.306	0.548	0.425	1.951	Version-B with 2 Encoder		0.875	0.335	0.581	0.46	2.155
B-SingleHead Realization		**0.882**	**0.339**	**0.599**	**0.464**	**2.157**	Version-B with 3 Encoder		**0.882**	**0.339**	**0.599**	**0.464**	**2.157**
C-MultiHead Realization		0.872	0.328	0.573	0.44	1.975	Version-B with 4 Encoder		0.876	0.33	0.584	0.453	2.164
D-MultiHead Realization		0.861	0.32	0.559	0.438	1.973	Version-B with 5 Encoder		0.874	0.327	0.576	0.439	1.996

Modules and Variants		AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	Modules and Variants	FLOPs↓	Params↓
	GCN	0.874	0.312	0.573	0.446	1.991	**RNNs**	347504640.0	113120.0
	RNNs	0.875	0.319	0.562	0.441	1.983	**Version-B with 3 Encoder**	1071673344.0	1043376.0
	C-GCN	**0.886**	**0.336**	**0.596**	**0.467**	**2.156**	**C-GCN (TA=1, MGNodes=2)**	189923328.0	19768.0

F.4 Ablation Study of Reasoning Steps and Node Number

To combine the multi-granularity information with the proposed SCAM, we have devised the reasoning unit, where there exist 2 key parameters that could influence the overall performance, specifically the exact reasoning steps (m) and the **m**ulti**g**ranularity (MG) node number (n).

Similar to the previous experimental settings, we test various versions of STA+ trained on data using SCAM+ as the pseudo GTs, where we set the ranges of m and n as $\{2, 3, 4, 5\}$ and $\{1, 2, 3, 4\}$, respectively. The corresponding quantitative results are in Table F4.

As seen in this table, we can find that '*w/o* Reasoning' indicates combining both the target node and other MG nodes simply via feature concatenation. Since these concatenated features are basically not aligned well, this plain implementation has demonstrated the worst result. Then, by performing the reasoning 3 times, our approach has reached its best performance, and if we continue the reasoning steps more than 3 times, the overall performance could degenerate significantly. The main reason is that the

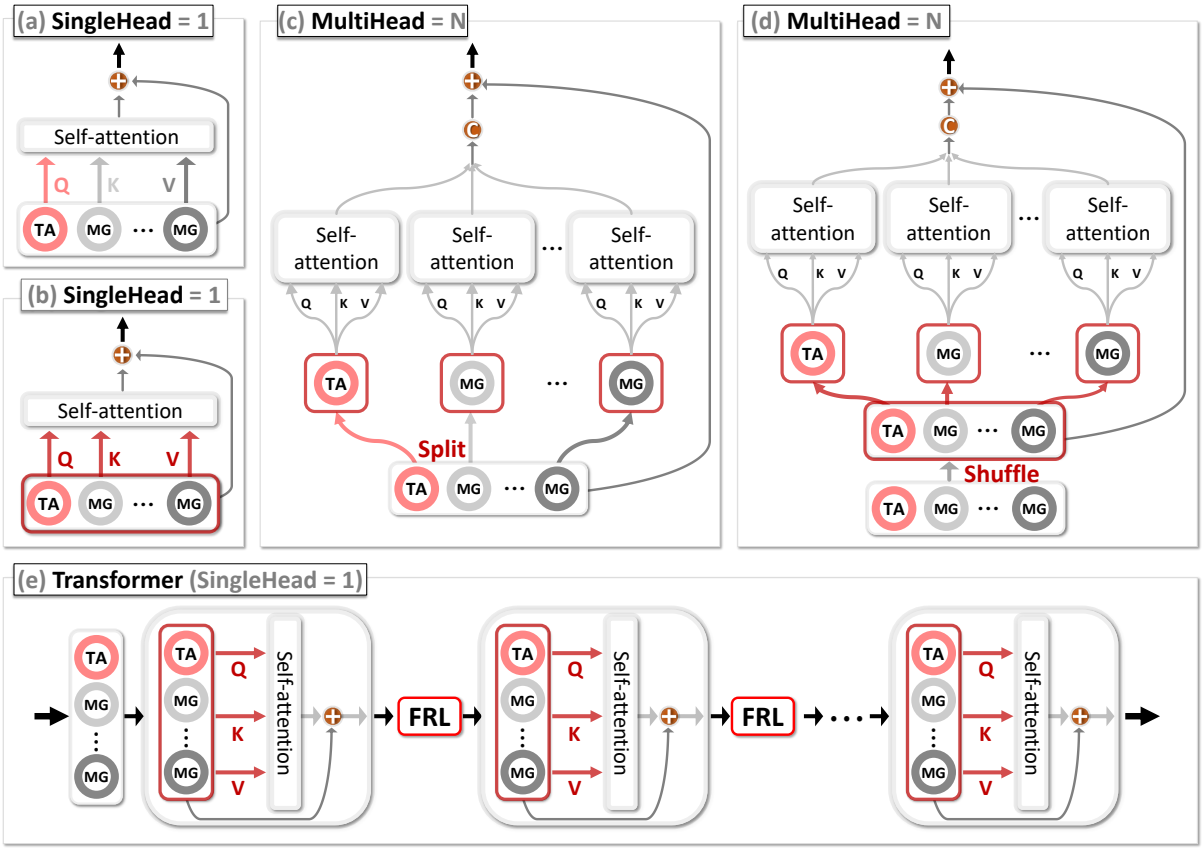

Figure F1 The detailed data flow process of replacing C-GCN with Transformer for relation modeling. For fair cooperation, we have incorporated FRL into the Transformer. The detailed experimental setup is described in Appendix F.4

reasoning step number usually determines the network's learning capacity; with a given training set with a limited size, there shall be a balance between capacity and data size, and $m = 3$ is the breaking point.

w.r.t. the exact MG node number, we can easily notice that the performance increases with more MG nodes ($1 \rightarrow 2$), while the overall performance could decrease if we use too many MG nodes ($3 \rightarrow 4$). Actually, with the increasing number of MG nodes, the feature space of the nodes' relationships could become more complex, making the learning task quite difficult. Though the MG nodes could make the process of computing SCAM+ more consistent with the real human attention mechanism, we shall stay conservative towards the exact choice of n, and thus we choose $n = 2$ as the optimal choice.

F.5 Different versions of multi-granularity information perception

We use a Transformer to jointly model all audio-visual fragments (for example, cross-, short-, and long-term information) as input, but this scheme often leads to a substantial performance degradation. The primary reason is that the Transformer-based CAM maps cannot cope with using our SCAM+. In our approach, we use multiple sub-branches to generate the multi-granularity CAM maps, each representing the target frame's feature response enhanced by the corresponding multi-granularity information (or example, short-, long-, and cross-term info.).

(1) As shown in Fig F2-top, our C-GCN-based version has adopted a total of three classification networks (as S+, SA+, and ST+ classification nets) to output 8 CAM maps. ***All these CAM maps mainly focus on the Target frame***, while the multi-granularity information is just the subordinates to enhance the target. Our approach can output high-quality fixation maps by applying the proposed SCAM+ over these CAM maps.

(2) However, when all this multi-granularity information is simultaneously fed to a Transformer (as Fig F2-middle), ***there is only one CAM map rooted from the target frame*** (as the $\{C, \phi\}^{\text{Target}}$), while others are multi-granularity centered. Thus, in this case, we ***cannot conduct the proposed SCAM+ over this solely CAM map***, resulting in significant performance degeneration.

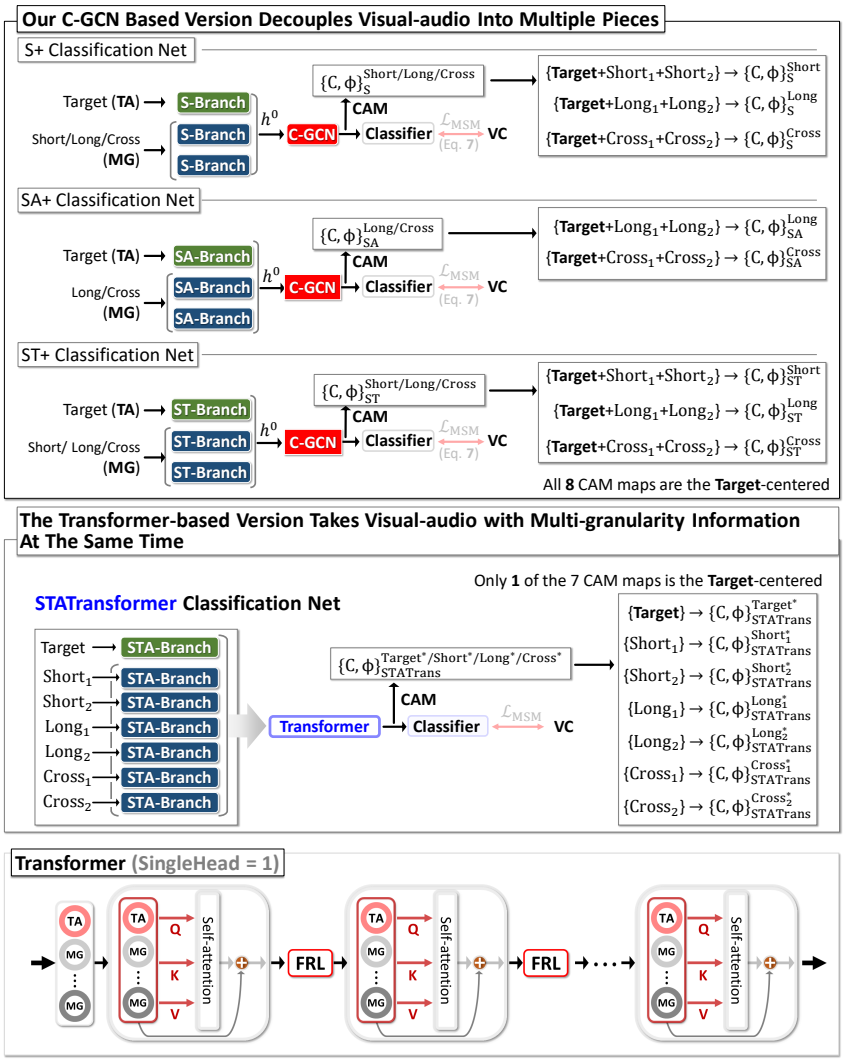

Figure F2 The network architecture differences between our C-GCN-based version and the Transformer-based version.

F.6 Fusion Schemes of the Final Fixation Prediction Nets

In the main paper, we have proposed a new fusion scheme to combine 3 fixation maps predicted by 3 different networks, namely the source wise PF_{sta}, and the multi-granularity perception-based $\mathrm{PF}_{sta+}^{\mathrm{Short}}$ and $\mathrm{PF}_{sta+}^{\mathrm{Long}}$. The rationale of the proposed fusion scheme is simply based on the complementary facts between these 3 fixation maps; thus, we verify the advantage of this fusion scheme towards conventional plain schemes from a quantitative perspective.

The comprehensive quantitative evaluation results are in Table F5. Compared with the 3 basic fixation maps (as the top-3 rows), the plain fusion schemes cannot take full advantage of their complementary attribute, and thus, the performance gains tend to shift between different sets and are truly marginal.

Specifically, since the proposed PF_{final} is based on performing an addition operation over its two parts, we have additionally tested the multiplicative-based alternative, denoted as PF_{agg}, which can be formulated as:

$$\mathrm{PF}_{agg} = \frac{1}{2} \times \mathcal{Z}\Big(\mathcal{Z}(\mathrm{PF}_{sta}) \odot \mathcal{Z}(\mathrm{PF}_{sta+}^{\mathrm{Short}}) \odot \mathcal{Z}(\mathrm{PF}_{sta+}^{\mathrm{Long}})\Big) \times \frac{1}{2} \times \mathcal{Z}\Big(\mathrm{PF}_{sta} + \mathrm{PF}_{sta+}^{\mathrm{Short}} + \mathrm{PF}_{sta+}^{\mathrm{Long}}\Big), \tag{2}$$

where $\mathcal{Z}(\cdot)$ is the typical *min-max* normalization function, and $\times$ represents the standard multiplicative operation. and $\odot$ denotes the element multiplicative operation.

As seen from Table F5, we find that the performance of PF_{agg} outperforms PF_{final} in several cases, the numeric gaps are marginal. In sharp contrast, PF_{final} can outperform PF_{agg} significantly in many cases, or example, the CC metric in the SunMe dataset, $0.359 \rightarrow 0.402$.

F.7 Quantitative evaluation of proposed SA Fuse with the existing plain ones

In addition, since the proposed audio switch is a major component of the proposed 'SA Fuse' module, we shall additionally verify whether the SA Fused module can outperform the conventional fusion scheme.

Table F5 Quantitative comparison of different fusion schemes for generating generic fixation maps on 6 testing datasets. The detailed explanation of PF-*sta* (PF_{sta}), PF-*sta* + *short* ($\mathrm{PF}_{sta+}^{\mathrm{Short}}$), PF-*sta* + *long* ($\mathrm{PF}_{sta+}^{\mathrm{Long}}$), PF-*vgg* ($\mathrm{PF}_{vgg}$), and PF-*final* ($\mathrm{PF}_{final}$) can be found in Appendix F.6.

Dataset	**AVAD**			**DIEM**			**SumMe**			**ETMD**			**Coutrot 1**			**Coutrot 2**		
Module	**AUC-J↑**	**CC↑**	**NSS↑**	**AUC-J↑**	**CC↑**	**NSS↑**	**AUC-J↑**	**CC↑**	**NSS↑**	**AUC-J↑**	**CC↑**	**NSS↑**	**AUC-J↑**	**CC↑**	**NSS↑**	**AUC-J↑**	**CC↑**	**NSS↑**
PF-sta	0.873	0.438	2.018	0.861	0.469	1.716	0.854	0.368	1.647	0.908	0.448	2.176	0.829	0.339	1.376	0.850	0.273	1.475
PF-sta+short	0.881	0.465	2.151	0.872	0.486	1.843	0.855	0.331	1.466	0.895	0.381	1.850	0.829	0.305	1.232	0.847	0.272	1.883
PF-sta+long	0.886	0.469	2.159	0.874	0.501	1.896	0.860	0.383	1.650	0.904	0.437	2.101	0.832	0.327	1.355	0.854	0.332	1.920
PF-vgg	0.886	**0.507**	**2.500**	0.862	0.500	1.920	0.844	0.359	1.593	0.898	0.434	2.192	0.814	0.314	1.270	0.803	0.341	1.952
PF-final	**0.887**	0.490	2.318	**0.884**	**0.544**	**2.042**	**0.866**	**0.402**	**1.761**	**0.910**	**0.452**	**2.297**	**0.84**	**0.354**	**1.429**	**0.862**	**0.349**	**1.952**

Thus, we have conducted an additional quantitative evaluation, where, in Table F6, we have compared the proposed SA Fuse with the existing plain ones, including addition, multiplicative, concatenation, and bilinear. As expected, the proposed SA Fuse module significantly outperforms all other approaches.

Table F6 Quantitative evidence towards the effectiveness of the proposed 'SA Fuse' module, 'Addition, Multiplication, Concatenation, Bilinear' is the visual and audio info fused mode, 'SA Fuse' is our proposed module of visual-audio fusion. This experiment is conducted on the AVAD set [17]

Modules	**AUC-J↑**	**SIM↑**	**s-AUC↑**	**CC↑**	**NSS↑**	Modules	**AUC-J↑**	**SIM↑**	**s-AUC↑**	**CC↑**	**NSS↑**
Addition	0.857	0.325	0.581	0.438	1.880	Concatenation	0.873	0.331	0.585	0.459	1.944
Multiplicative	0.860	0.327	0.575	0.442	1.915	Bilinear	0.864	0.329	0.577	0.428	1.825
SA Fuse	**0.886**	**0.336**	**0.596**	**0.467**	**2.156**	**SA Fuse**	**0.886**	**0.336**	**0.596**	**0.467**	**2.156**

F.8 The relationship between dataset size and performance

To further explore the relationship between dataset size and our approach's overall performance, we have conducted an ablation study, where we tested the overall performance of our approach under different training data sizes. As shown in Table F7, the overall performance of our method could decrease with the decrease in training data size. We have also noticed that when the training data size is very small, specifically 180, our method could degenerate into a typical CAM-based method.

Table F7 The overall performance of our approach under different training data sizes, "0.04" means that the number of randomly selected video fragments from the AVE dataset accounts for 4% of the entire dataset. This experiment is conducted on the AVAD set [17].

Modules	**AUC-J↑**	**SIM↑**	**s-AUC↑**	**CC↑**	**NSS↑**	Modules	**AUC-J↑**	**SIM↑**	**s-AUC↑**	**CC↑**	**NSS↑**
Addition	0.857	0.325	0.581	0.438	1.880	Concatenation	0.873	0.331	0.585	0.459	1.944
Multiplicative	0.860	0.327	0.575	0.442	1.915	Bilinear	0.864	0.329	0.577	0.428	1.825
SA Fuse	**0.886**	**0.336**	**0.596**	**0.467**	**2.156**	**SA Fuse**	**0.886**	**0.336**	**0.596**	**0.467**	**2.156**

References

1 Carion N, Massa F, Synnaeve G, et al, End-to-end object detection with transformers. In: ECCV, 213–229, 2020.
2 Wang H, Zha Z.-J, Li L, et al, Semantic and relation modulation for audio-visual event localization. TPAMI, 2022, 7711–7725.
3 Hu J, Shen L, Albanie S, et al, Gather-excite: Exploiting feature context in convolutional neural networks. In: NIPS, 2018.
4 Wang Q, Wu B, Zhu P, et al, Ecanet: Efficient channel attention for deep convolutional neural networks. In: CVPR, 2020.
5 Woo S, Park J, Lee J.-Y, et al, Cbam: Convolutional block attention module. In: ECCV, 3–19, 2018.
6 Liu R, Deng H, Huang Y, et al, Decoupled spatial-temporal transformer for video inpainting. arXiv preprint arXiv:2104.06637, 2021.
7 Dai J, Qi H, Xiong Y, et al, Deformable convolutional networks. In: ICCV, 2017.
8 Wang W, Shen J, Xie J, et al, Revisiting video saliency prediction in the deep learning era. TPAMI, 2021, 220–237.
9 Leboran V, GarciaDiaz A, FdezVidal X, et al, Dynamic whitening saliency. TPAMI, 2016, 893–907.
10 Itti L, Koch C, and Niebur E, A model of saliency-based visual attention for rapid scene analysis. TPAMI, 1998, 1254–1259.
11 Rudoy D, Goldman D, Shechtman E, et al, Learning video saliency from human gaze using candidate selection. In: CVPR, 2013.
12 Harel J, Koch C, and Perona P, Graph-based visual saliency. In: NeurIPS, 2007.
13 Zhang J, Yu X, Li A, et al, Weakly-supervised salient object detection via scribble annotations. In: CVPR, 2020.
14 Gao S, Zhang W, Wang Y, et al, Weakly-supervised salient object detection using point supervision. In: AAAI, 2022.
15 Zeng Y, Zhuge Y, Lu H, et al, Multi-source weak supervision for saliency detection. In: CVPR, 2019.
16 Wang L, Lu H, Wang Y, et al, Learning to detect salient objects with image-level supervision. In: CVPR, 2017.
17 Min X, Zhai G, Gu K, et al, Fixation prediction through multimodal analysis. ACM TOMM, 2016.