MEGA: Multimodal Alignment Aggregation and Distillation For Cinematic Video Segmentation
Supplementary Material

Najmeh Sadoughi[1], Xinyu Li[1], Avijit Vajpayee[1], David Fan[1], Bing Shuai[2], Hector Santos-Villalobos[1], Vimal Bhat[1], Rohith MV[1]
[1]Amazon Prime Video, [2]AWS AI Labs
{nnnourab,xxnl,avivaj,fandavi,bshuai,hsantosv,vimalb,kurohith}@amazon.com

1. Architecture

The details of our model architecture are shown in Tab. 1a. The model is composed of unimodal encoders, fusion encoders and output layer. The unimodal encoders are repeated for all M modalities and consist of a Shot Encoder, Bottleneck layer, and Transformer Encoder repeated N_u times. The Fusion starts with concatenating the fusion tokens and further includes the Transformer Encoder repeated N_f times for each modality, and final concatenation of the latent tokens for different modalities ($C_{\text{fused}} = C \times M$) per shot. The hyperparameter values used to derive the architecture per task (Scene/Act Segmentation) are provided in Tab. 1b.

2. Feature Extraction

$\text{CLIP}_{\text{movie}}$ model is the the original CLIP [9] model with ViT-B/32 backbone fine-tuned on IMDB-image dataset. The IMDB-image dataset includes 1.6M images from 31.3K unique movies/TV series paired with 762 unique textual labels. This model is trained with contrastive loss similar to CLIP [9]. The differences with [9] are: a) the textual labels are from a limited set, b) the positive and negative keys for a query sample are identified by their labels, c) the number of positive keys per image can be more than one in a batch, and d) not all other samples in a batch are considered negative keys for a query sample, only the ones with different sets of labels are considered negative keys.

3. Details of Feature Extraction Per Shot

Bassl features are used for scene segmentation and are extracted from 3 key frames per shot, as Movient only releases 3 key frames per shot. Appr, place, clip, action features are extracted every 1 second. The input for extracting appr, place, and clip features is 1 frame, and for extracting the action features is a sequence of 16 frames. Audio features are extracted every 1 second, and the audio model's input is a window of 10 seconds. Text features are extracted for each subtitle timed text and each sentence of the synopsis. For each shot, we assign the features whose input has overlap with the shot time interval. E.g., audio is split into 10s (seconds) windows with an overlap of 9s (stride =1s). Then, we get the features for windows which overlap with each shot, or for subtitles, we get the features for the subtitle timed segments whose time interval overlaps with each shot time interval.

4. Expectation Step

For synchronization between synopsis sentences and movie shots, we use Eq. 1 as the objective. This objective is solved in an alternative manner, where we estimate the target variable W via fixed parameters in the first step (E-step), and update the parameters while the target variable is known (M-step).

$$\max_{W,\theta} \sum_{i,j} w_{ij} F(.;\theta) - \lambda \sum_{i,j} |w_{i,j}| \quad s.t.\ 0 \leq w_{ij} \leq 1 \tag{1}$$

Let M_{sim} with dimension $L_{sh} \times L_{syn}$ be the similarity matrix with m_{ij} entries representing the similarity between the i-th shot and j-th synopsis sentence for one sample. Assuming $F(.;\theta) = M_{sim}$, there is a closed form solution for Eq. 1 in the expectation step. We also use [4] to reduce the search space during optimization to only the pairs which are inside a diagonal boundary, i.e., all m_{ij} outside the diagonal boundary are ignored. Eq. 2 shows the closed form solution for the expectation step. Following [4], ξ is set to 0.3.

$$w_{ij} = \begin{cases} 1 & \text{if } m_{ij} \geq \lambda \,\&\, j < i\frac{L_{syn}}{L_{sh}} + \xi L_{syn} \,\&\, i < j\frac{L_{sh}}{L_{syn}} + \xi L_{sh} \\ 0 & \text{if } m_{ij} < \lambda \,||\, j \geq i\frac{L_{syn}}{L_{sh}} + \xi L_{syn} \,||\, i \geq j\frac{L_{sh}}{L_{syn}} + \xi L_{sh} \end{cases} \tag{2}$$

Proof: Due to non-negative constraint $0 \leq w_{ij} \leq 1$, the

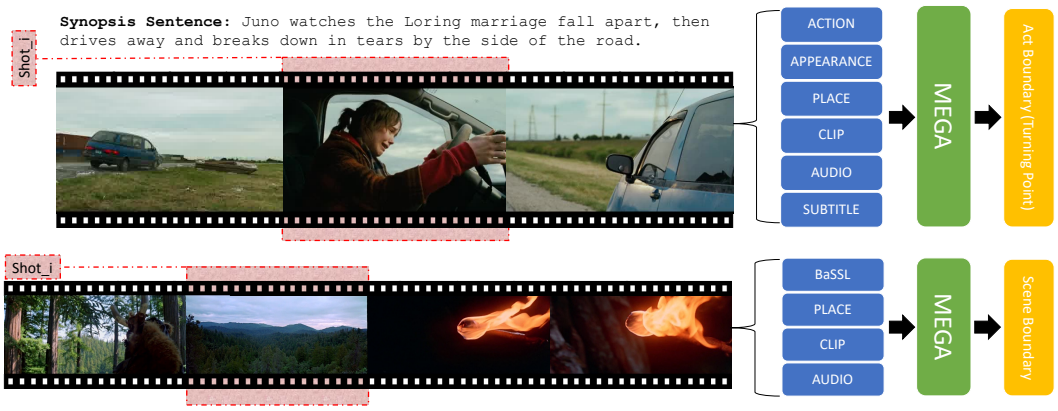

Figure 1: This figure demonstrates the two applications used to test our model for video segmentation: 1) act boundary segmentation (tops) and 2) scene boundary segmentation (bottom). The model utilizes several features extracted from pretrained models to predict a decision for each shot.

objective function in Eq. 1 is reduced to Eq. 3.

$$\max_{W} \sum_{i,j} w_{ij}(m_{ij} - \lambda) \quad s.t.\ 0 \leq w_{ij} \leq 1 \tag{3}$$

This objective has a closed form solution, shown in Eq. 4. Combining this solution with the diagonal constraint from [4] the final solution boils down to Eq. 2.

$$w^*_{ij} = \begin{cases} 1 & \text{if } m_{ij} \geq \lambda \\ 0 & \text{if } m_{ij} < \lambda \end{cases} \tag{4}$$

5. Knowledge Transfer

Knowledge distillation is used to transfer the knowledge available for the training samples on synopsis. More concretely, the soft similarity scores for each shot are normalized using Eq. 5, and the shot level probability scores are derived using Eq. 6 for i-th shot and n-th TP. Let Y be the predicted logits from the shot model ($\in \mathbb{R}^{L_{sh} \times N_{tp}}$, with entries y_{in} for the i-th shot and n-th TP) , where a softmax function derives the probability predictions per shot and TP (Eq. 7). Given predicted probabilities from the model on the shot level O ($\in \mathbb{R}^{L_{sh} \times N_{tp}}$, with entries o_{in} for the i-th shot and n-th TP), Kullback–Leibler divergence loss between $O.n$ and $P.n$ for each of the turning points (i.e., each n) is minimized during training (Eq. 8).

$$a_{ij} = \frac{exp(u_i v_j/\tau)}{\sum_k^{L_{syn}} exp(u_i v_k/\tau)} \tag{5}$$

$$p_{in} = \frac{exp\left(\sum_{k=1}^{L_{syn}} a_{ik} q_{kn}\right)}{\sum_{j=1}^{L_{sh}} exp\left(\sum_{k=1}^{L_{syn}} a_{jk} q_{kn}\right)} \tag{6}$$

$$o_{in} = \frac{exp(y_{in})}{\sum_{k=1}^{L_{sh}} exp(y_{kn})} \tag{7}$$

$$\mathscr{L}_{kd} = \sum_{n=1}^{N_{tp}} \mathrm{KL}\left(O_{.n} || P_{.n}\right) \tag{8}$$

6. Experiments

6.1. Additional Results

Unless otherwise specified, this section includes further statistics and metrics for the same experiments which are provided in the paper.

Tab. 2 includes the $F1$ score for Scene Segmentation model in all the ablation experiments. The ablation experiments show a similar trend across AP and $F1$ score. Tab. 2d shows an extra experiment with ablation of bassl features (i.e., -bassl(50)). In this experiment, we further inspect this ablation by continuing the training for this model for 50 epochs, which still shows a significant difference with the model with all the modalities included, demonstrating the effectiveness of multimodal fusion in MEGA.

Additionally, given that all experiments for Act Segmentation with MEGA are repeated 8 times, and the performance metrics are averaged, the standard deviations (STD) of all such experiments for comparison with previous SoTA and in ablation experiments are provided in parentheses in Tabs. 3 and 2, respectively. Considering the STD values, results still demonstrate that MEGA outperforms previous SoTA on act segmentation, and the ablations of various components in MEGA worsen the performance, showing the importance of each of those components. Additionally, Tab. 2a includes an extra experiment (i.e., w/o align. PE(50)), where we further investigate the ablation of this module for act segmentation by training the model without the normalized positional encoding for longer time (50 epochs). By further training, the performance gap reduces but still remains to be significant, which indicates that the proposed normalized positional encoding not only helps the model converge faster but also is a

	stage	layer	output size
Unimodal	Shot Encoder (Sentence Encoder)	Linear: $D^m \times C$	$L \times C$
		Pos. Enc.: $L \times C$	$L \times C$
		Align.PE: $L_n \times C$	$L \times C$
		LayerNorm: C	$L \times C$
		Dropout (p)	$L \times C$
	Bottleneck	$L_n \times C$	$L_n \times C$
		Align.PE: $L_n \times C$	$L_n \times C$
	Concat		$(L_n + L) \times C$
	Trans.Enc. $\times N_u$	Self Attn: $3 \times C$	$(L_n + L) \times C$
		LayerNorm: C	$(L_n + L) \times C$
		Linear: $C \times C_k$	$(L_n + L) \times C_k$
		Activation: GeLU	$(L_n + L) \times C_k$
		Linear: $C_k \times C$	$(L_n + L) \times C$
		LayerNorm: C	$(L_n + L) \times C$
		Dropout (p)	$(L_n + L) \times C$
Fusion	Concat		$(M \times L_n + L) \times C$
	Trans.Enc. $\times N_f$	Self Attn: $3 \times C$	$(M \times L_n + L) \times C$
		LayerNorm: C	$(M \times L_n + L) \times C$
		Linear: $C \times C_k$	$(M \times L_n + L) \times C_k$
		Activation: GeLU	$(M \times L_n + L) \times C_k$
		Linear: $C_k \times C$	$(M \times L_n + L) \times C$
		LayerNorm: C	$(M \times L_n + L) \times C$
		Dropout (p)	$(M \times L_n + L) \times C$
	Concat		$L \times C_{\text{fused}}$
	Output	Linear ($C_{\text{fused}} \times N_c$)	$L \times N_c$

(a) Architecture

Name	Scene Segmentation	Act Segmentation	
		Shot Model	Syn. Model
L	17	3000	60
L_n	2	100	20
C	768	128	$128 \times M$
C_k	3072	128	$128 \times M$
N_u	2	1	1
N_f	1	1	-
p	0.1	0.5	0.1
N_c	2	5	5

(b) Hyperparameters

Table 1: Architecture details.

necessary component. Furthermore, Tab. 2e shows an extra experiment (-clip(20)), where we further trained the model for 20 epochs. The results are improved but there is still a significant gap with the model which uses clip, showing that the importance of multimodal fusion in MEGA.

6.2. Hyperparameter Search in Align. PE. for L_n

We demonstrate the performance of act segmentation model with different values for L_n (i.e., length of Align. PE.) in Fig. 2. The experiment results reveal that while MEGA performs best at $L_n = 100$ for act segmentation, it performs robustly across a range of L_n values (specifically, within the range of 50 to 150). Align. PE. is designed to provide video-level coarse progress information as a complementary signal next to regular PE. If L_n is very small, the

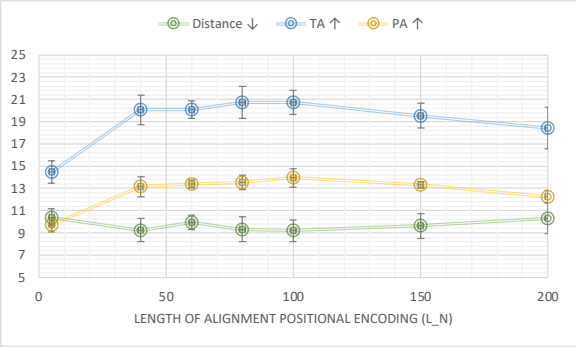

Figure 2: Evaluation of act segmentation in terms of Distance, TA, and PA metrics by changing L_n.

granularity of the align. PE. becomes too coarse, whereas if L_n is too large, it becomes overly detailed as regular PE; which explain the performance decline we observe in Fig. 2 at small and big L_n values. Aside from better performance, using $L_n \ll L$, where L is the input length, results in better efficiency. During multimodal fusion, the number of fusion tokens is equal to L_n, hence the memory consumption during multimodal fusion attention calculation is $(L_n \times m + L)^2$, where m is the number of modalities. Our approach enables aligned multimodal alignment during fusion that scales efficiently with respect to the number of modalities in terms of memory consumption.

6.3. Visualization

6.3.1 Feature Importance

To further look into how MEGA is integrating different modalities to make a prediction, we calculated the Grad-CAM [11] for the output features from the fusion module. More concretely, the derivative of the outputs (the maximum prediction logit of the two dimensional output for *scene segmentation* and the max predicted shot for *act segmentation*) with respect to the final FC layer parameters are calculated. These values are then multiplied with activation scores coming out of each modality fusion module, summed across the channel dimension and undergone a ReLU non-linearity function. The value scores are then normalized across all modalities, such that their summation is 1. This helps to visualize the effect of each of the modalities in the prediction from the model.

For scene segmentation, Fig. 3 demonstrates the results. The results in this figure are aligned with Tab. 2d showing the order of importance for the modalities are bassl, place and clip. For act segmentation, Fig. 4 demonstrates the results for the 5 predicted act boundaries. The demonstrations are aligned with the ablation experiments in Tab. 2e, showing the highest contributions are from the clip and subtitle modalities.

6.3.2 Expected Synchronization Matrix

Fig. 5 shows the expected value of synchronization matrix during optimization on different samples from TRIPOD test set. The results demonstrate that the synchronization matrix synchronizes the synopsis sentences and shots more along the diagonal line, which is expected.

6.3.3 Attention Scores

Figs. 6 and 7 demonstrate the attention scores for the fusion modules across the Scene Segmentation and Act Segmentation models on several test samples. For better visualization, all scores within one image are normalized by their max value within that image. These figures clearly demonstrate that the model is fusing different modalities flexibly for different time units (i.e., shots). And, while for some of the modalities the fusion patterns remains more similar across different time units, for some there is a clear change in pattern across time (e.g., see the zoomed area in the last row of Clip attention, which shows MEGA's fusion tokens have the capability to preserve temporal information).

References

[1] Philip Gorinski and Mirella Lapata. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076, 2015. 6

[2] Michael Hauge. *Storytelling Made Easy: Persuade and Transform Your Audiences, Buyers, And Clients-Simply, Quickly, and Profitably*. BookBaby, 2017. 6

[3] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004. 6

[4] Paramita Mirza, Mostafa Abouhamra, and Gerhard Weikum. Alignarr: Aligning narratives on movies. In *The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 427–433. ACL, 2021. 1, 2

[5] Jonghwan Mun, Minchul Shin, Gunsoo Han, Sangho Lee, Seongsu Ha, Joonseok Lee, and Eun-Sol Kim. Boundary-aware self-supervised learning for video scene segmentation. *arXiv preprint arXiv:2201.05277*, 2022. 5

[6] Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. Screenplay summarization using latent narrative structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1920–1933, 2020. 6

[7] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie plot analysis via turning point identification. *arXiv preprint arXiv:1908.10328*, 2019. 5, 6

[8] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie summarization via sparse graph construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13631–13639, 2021. 5, 6

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[10] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10155, 2020. 5

[11] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

	Scene Seg.		Act Segmentation		
case	*AP*	*F1*	*TA*	*PA*	*D*
w/o align. PE	57.77	54.75	5.29 (4.74)	7.37 (0.75)	31.04 (1.04)
w/o align. PE(50)	-	-	7.31 (1.13)	10.26 (1.19)	15.75 (1.17)
w. align. PE	58.59	55.29	13.93 (2.21)	20.72 (2.28)	9.19 (0.58)

(a) Effectiveness of Alignment Positional Encoding.

		Scene Seg.		Act Segmentation		
case	*modality*	*AP*	*F1*	*TA*	*PA*	*D*
w/o align.PE	V	58.31	54.83	13.60 (1.59)	20.53 (2.27)	9.47 (1.23)
w. align.PE	V	58.59	55.29	13.93 (2.21)	20.72 (2.28)	9.19 (0.58)
w/o align.PE	V + T	-	-	13.01 (1.43)	20.13 (2.33)	9.56 (0.92)
w. align.PE	V + T	-	-	14.63 (1.10)	21.78 (1.22)	8.96 (0.65)

(b) Effectiveness of Normalized Positional Encoding in bottleneck tokens.

	Scene Seg.		Act Segmentation		
MM. integ. type	*AP*	*F1*	*TA*	*PA*	*D*
LateFusion	58.24	48.27	12.57 (1.71)	19.21 (2.34)	10.00 (1.22)
Bottleneck	58.59	55.29	13.93 (2.21)	20.72 (2.28)	9.19 (0.58)

(c) Multi-modal fusion strategies.

change	*AP*	*F1*
-clip	58.09	53.95
-place	57.51	54.83
-bassl	51.88	43.04
-bassl(50)	53.80	43.67
-clip-place	57.92	50.71
-	58.59	55.29

(d) Impact from input modalities on scene segmentation.

change	*TA*	*PA*	*D*
-clip	6.09 (0.86)	10.66 (1.48)	21.81 (4.30)
-clip(20)	11.99 (1.73)	17.89 (3.04)	10.32 (1.68)
-place	13.57 (2.78)	19.87 (4.00)	9.22 (1.38)
-action	13.31 (1.98)	20.20 (2.90)	10.38 (2.03)
-appr	13.42 (2.34)	20.59 (3.40)	8.85 (0.90)
-	13.93 (2.21)	20.72 (2.28)	9.19 (0.58)
+subtitle	13.93 (2.21)	20.72 (2.28)	9.19 (0.58)
+subtitle+audio	14.19 (1.13)	22.10 (1.46)	9.68 (1.06)

(e) Impact from input modalities on act segmentation.

		Act Segmentation		
synopsis synch. by	*M for synch.*	*TA*	*PA*	*D*
[7]	T	10.51 (0.72)	14.54 (1.09)	8.98 (0.25)
MEGA	V	13.93 (2.21)	20.72 (2.28)	9.19 (0.58)
MEGA	V + T	14.63 (1.10)	21.78 (1.22)	8.96 (0.65)

(f) Impact of Synchronization with multimodal video features.

Approach	*Feature Set Pretrained on*	*AP*	*F1*	*Params*	*SPS*
BaSSL [5]	Movienet	57.4	47.02	15.77M	6244.99
LGSS [10]	M+P+I	52.93	48.75	66.16M	206.36
MEGA	M+P+I	58.59	55.30	67.57M	1736.13

(g) Impact from feature set and model size on scene seg. SPS denotes # of samples per second.

Approach	*Feature Set*	*TA*	*PA*	*D*	*Params*	*SPS*
GRAPHTP [8]	Set1 [8]	9.12	12.63	9.77	0.745M	25.40
GRAPHTP [8]	Set2	4.72	7.37	9.69	6.78M	14.36
MEGA	Set2	14.19 (1.13)	22.10 (1.46)	9.68 (1.06)	6.78M	18.24

(h) Impact from feature set and model size on act seg. Set1 includes Visual (appr), Audio (YAMNet), Textual (script-USE). Set2 has Visual (appr,clip,action,place), Audio (audio), Textual (text from subtitle).

Table 2: Ablation studies on MEGA components. Values within parentheses are standard deviations for multiple runs.

Approach	*Modality*	*Modality for synch.*	*TA [%]*	*PA [%]*	*D [%]*
Random (Even. distribution) [8]	-	T*	4.82	6.95	12.35
Theory [2, 7]	-	T*	4.41	6.32	11.03
Distribution position [8]	-	T*	5.59	7.37	10.74
Single modality input					
TEXTRANK [3]	T	T*	6.18	10.00	17.77
SCENESUM [1]	T	T*	4.41	7.89	16.86
TAM [6]	T	T*	7.94	9.47	9.42
GRAPHTP [8]	T	T*	6.76	10.00	9.62
MEGA*	V	T*	10.51 (0.72)	14.54 (1.09)	8.98 (0.25)
MEGA	V	V	13.93 (2.21)	20.72 (2.28)	9.19 (0.58)
Multi-modality input					
TEXTRANK [3]	T+A+V	T*	6.18	10.00	18.90
SCENESUM [1]	T+A+V	T*	6.76	11.05	18.93
TAM [6]	T+A+V	T*	7.36	10.00	10.01
GRAPHTP [8]	T+A+V	T*	9.12	12.63	9.77
MEGA*	T+V	T*	11.14 (1.77)	15.20 (2.33)	**8.96 (0.35)**
MEGA	T+V	T+V	**14.63 (1.10)**	21.78 (1.22)	**8.96 (0.65)**
MEGA*	T+A+V	T*	10.00 (0.98)	14.08 (1.56)	8.96 (0.39)
MEGA	T+A+V	T+A+V	14.19 (1.13)	**22.10 (1.46)**	9.68 (1.06)

Table 3: TP identification: comparison with SoTA. MEGA* denotes the MEGA using the same synchronization as in [7] for fair comparison. T*,V,T,A denote Textual-screenplay, Visual, Textual-subtitle and Acoustic features respectively. Values within parentheses are standard deviations for multiple runs.

Figure 3: GradCAM values shown in pie charts for 9 different predictions on scene segmentation on the test set, along with the probability prediction score for the middle shot being the end of a scene and its groundtruth label.

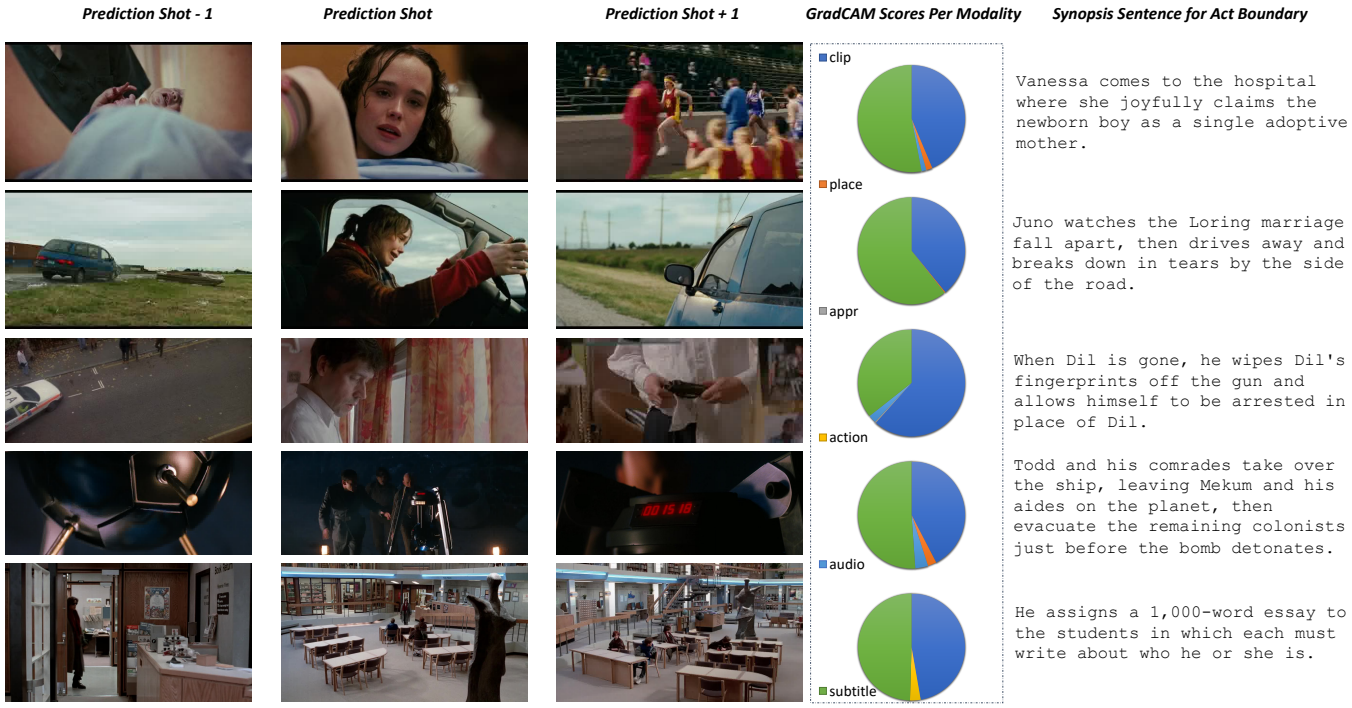

Figure 4: GradCAM for 5 different predicted turning points on the test set, along with their synopsis annotated sentence for that turning point.

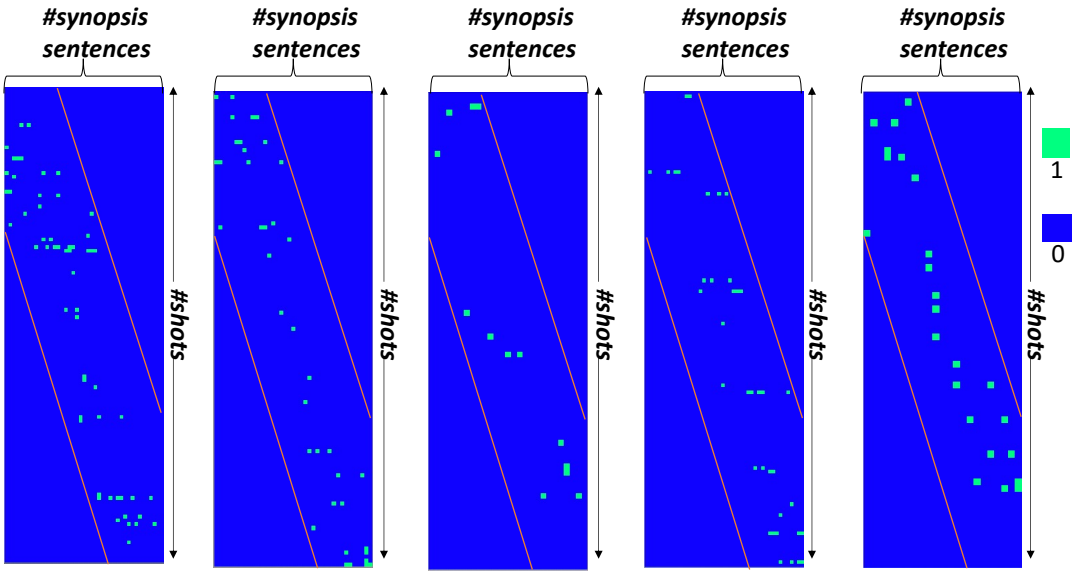

Figure 5: The expected value of synchronization matrix for 5 different samples on the test set. The actual matrices had to be resized for visualization (ratio of height/width is set to 3).

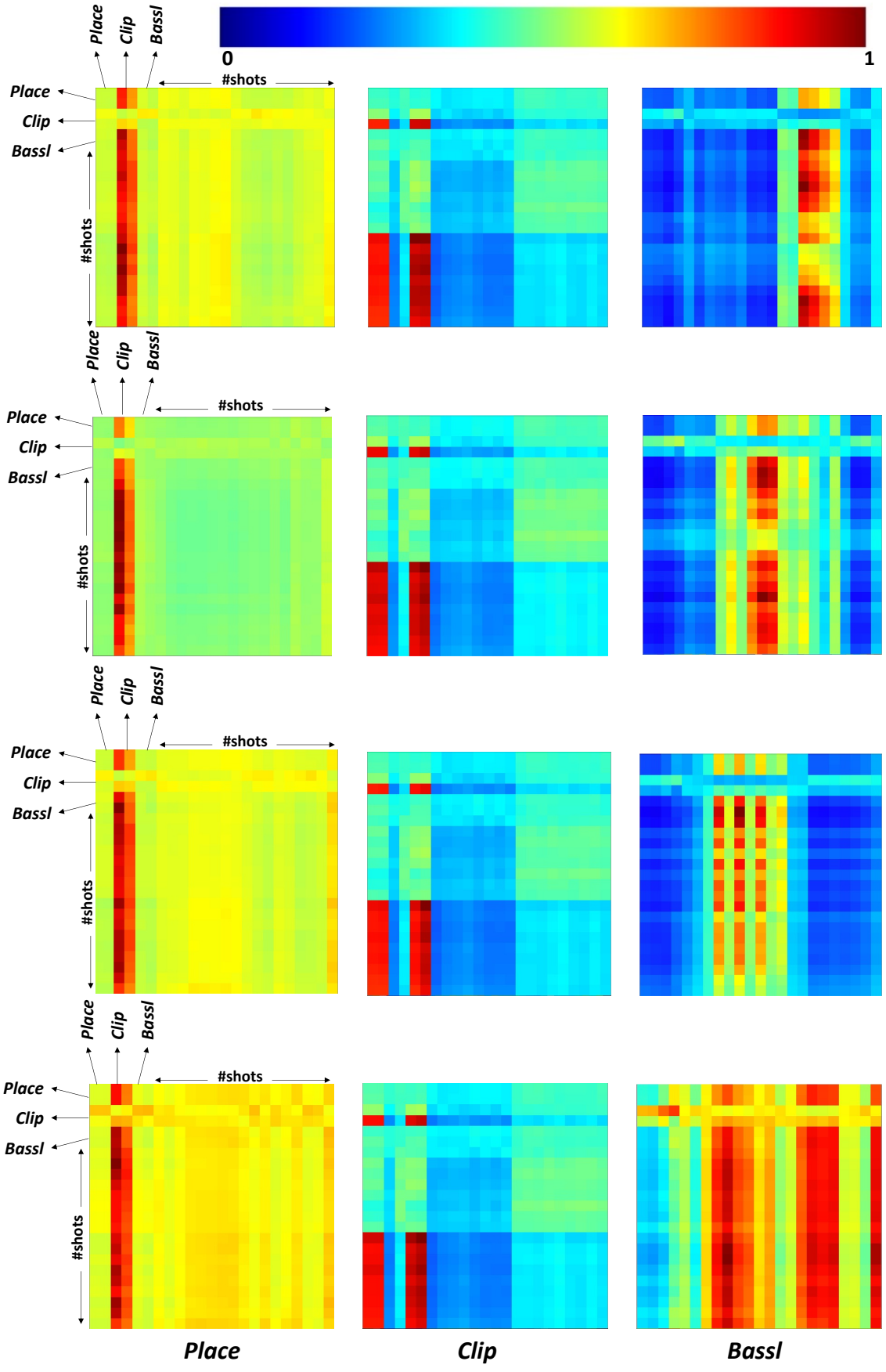

Figure 6: Attention scores derived from the fusion transformer encoder on Scene Segmentation model.

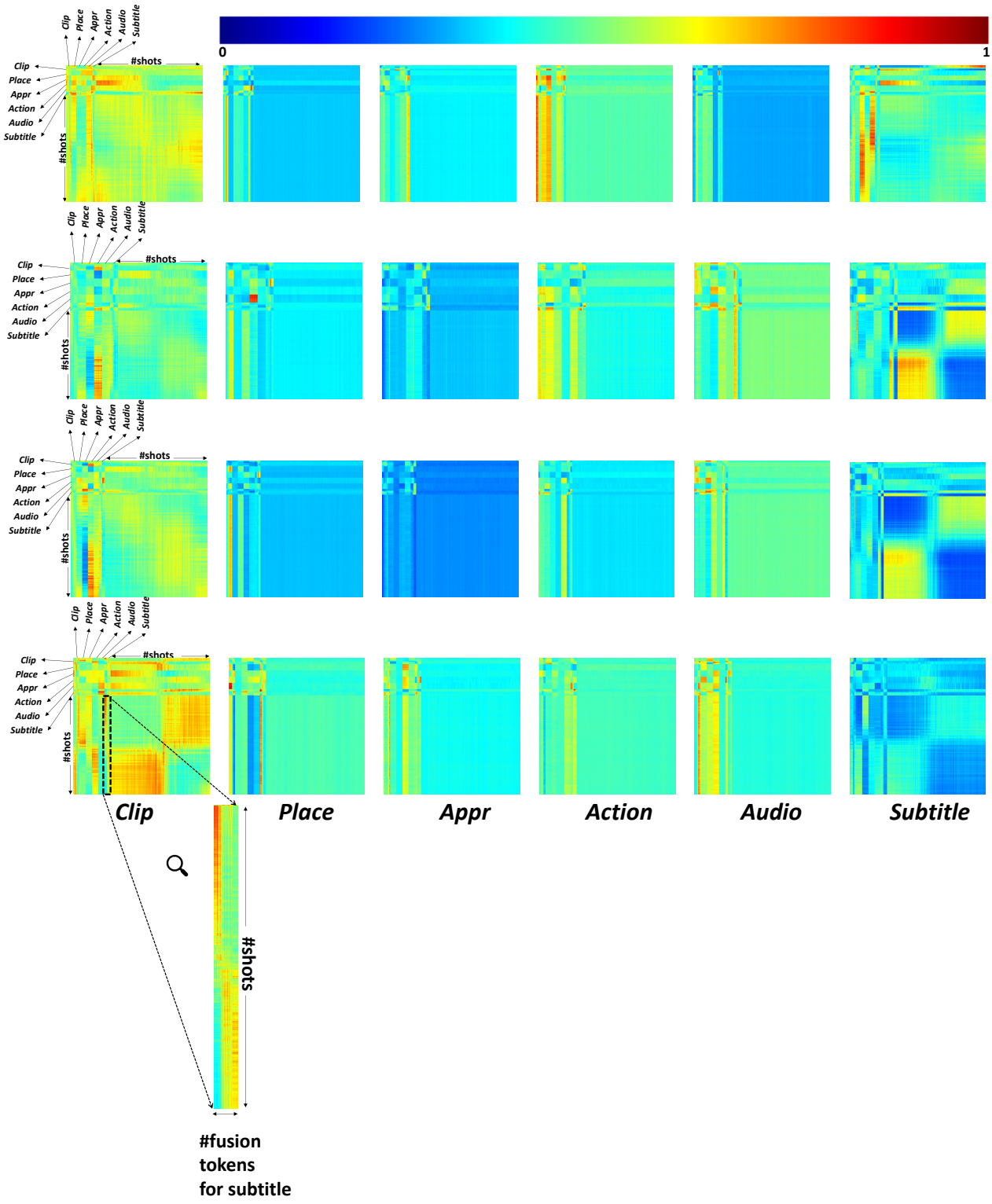

Figure 7: Attention scores derived from the fusion transformer encoder on Act Segmentation model.