Supplementary Material: Adapting to the Unknown: Training-Free Audio-Visual Event Perception with Dynamic Thresholds

A. Score-Level Fusion

We conducted an experiment to evaluate the impact of a straightforward score-level fusion approach on existing baselines. The score-level fusion technique involves integrating audio and visual vectors through a weighted sum, controlled by a parameter α. To ensure minimal modifications, we exclusively searched for the optimal α value through hyperparameter tuning on the LLP validation set, without retraining the models or adjusting any other parameters. As shown in Table 6, this simple score-level fusion (Eq. 4) method improved the performance of audio-visual results, demonstrating its effectiveness in enhancing baseline models without additional computational overhead. Notably, this analysis focuses exclusively on audio-visual results, as the score-level fusion technique does not alter the outcomes of individual audio or visual modalities.

Method	Audio Visual	
	segment	event
MGN-MA [18]	50.6	44.4
+ score-level fusion	51.8	45.7
JoMoLD [2]	57.2	49.6
+ score-level fusion	57.7	51.0
CMPAE [5]	59.5	52.4
+ score-level fusion	**59.8**	**52.6**
LanguageBind with AV^{2}A	57.1	50.6
+ score-level fusion	59.1	52.3
CLIP+CLAP with AV^{2}A	51.8	45.7
+ score-level fusion	55.1	48.2

Table 6. The effect of score-level fusion instead of late fusion for the current state-of-the-art weakly supervised methods and for our training-free method.

B. Comparison to Weakly-Supervised Methods

We compare our training-free method performance against weakly supervised baselines across segment-based and event-level metrics (see Table 8), analyzing its strengths in multimodal fusion despite the absence of training. The results are reported for 1124 out of 1200 test videos, as only these videos are accessible on the internet. For the audio-visual segment-based score, our training-free method is on par with CMPAE [5] and better than JoMoLD [2] and MGN-MA [18]. For the event score, it is slightly better than CMPAE and much better than the rest. In separate audio and visual metrics, our method is inferior to the weakly supervised methods, emphasizing its multimodal nature.

C. Linear Classification with CLIP/CLAP

To further analyze the effectiveness of the features extracted from CLIP and CLAP, we trained a linear classifier to predict the category of events per second. This evaluation serves as a measure of the linear separability of these features and provides information on their suitability for event classification in a weakly supervised setting.

Experimental Setup. We trained a fully supervised linear classifier using the features from CLIP and CLAP. The classifier was optimized to predict the event category for each second of the video, thereby assessing the discriminative power of these features at a fine temporal granularity. The LLP dataset was used for evaluation, and we report both segment-level and event-level performance.

The results shown in Table 7 indicate that CLIP features exhibit stronger linear separability than CLAP features for event classification in this setting. This suggests that vision-language models such as CLIP encode more discriminative information that can be leveraged for event recognition with a simple linear probe. The lower performance of CLAP features may be attributed to the nature of audio embeddings, which might require more complex modeling techniques beyond linear classification to effectively capture event distinctions.

Method	segment	event
CLIP [22]	33.9	32.1
CLAP [4]	27.2	25.9

Table 7. Linear classifier performance on CLIP and CLAP features for event classification on the LLP dataset.

	Audio		Visual		Audio-Visual		Type@AV		Event@AV	
	segment	event	segment	event	segment	event	segment	event	segment	event
Weakly Supervised Methods										
MGN-MA [18]	60.3	50.5	55.3	52.2	50.1	44.0	55.3	48.9	56.9	48.8
JoMoLD [2]	61.1	53.7	63.5	59.8	56.8	49.3	60.5	54.2	59.7	52.3
CMPAE [5]	**64.1**	**57.2**	**66.1**	**63.3**	**59.1**	52.2	**63.3**	**57.7**	**62.9**	**56.3**
Training-Free Methods										
LanguageBind+AV^{2}A	40.9 ± .09	35.9 ± .12	57.4 ± .04	54.4 ± .09	**59.1** ± .1	**52.3** ± .13	52.4 ± .04	47.5 ± .04	43.4 ± .04	38.9 ± .06

Table 8. Performance of state-of-the-art weakly supervised methods in comparison to our training-free method on the LLP test set.