This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Eitan Shaar[1*] Ariel Shaulov[2*] Gal Chechik[1,3] Lior Wolf[2]

[1]Bar-Ilan University [2]Tel-Aviv University [3]NVIDIA

Abstract

*In the domain of audio-visual event perception, which focuses on the temporal localization and classification of events across distinct modalities (audio and visual), existing approaches are constrained by the vocabulary available in their training data. This limitation significantly impedes their capacity to generalize to novel, unseen event categories. Furthermore, the annotation process for this task is labor-intensive, requiring extensive manual labeling across modalities and temporal segments, limiting the scalability of current methods. Current state-of-the-art models ignore the shifts in event distributions over time, reducing their ability to adjust to changing video dynamics. Additionally, previous methods rely on late fusion to combine audio and visual information. While straightforward, this approach results in a significant loss of multimodal interactions. To address these challenges, we propose **A**udio-**V**isual **A**daptive **V**ideo **A**nalysis (AV^2A), a model-agnostic approach that requires no further training and integrates a score-level fusion technique to retain richer multimodal interactions. AV^2A also includes a within-video label shift algorithm, leveraging input video data and predictions from prior frames to dynamically adjust event distributions for subsequent frames. Moreover, we present the first training-free, open-vocabulary baseline for audio-visual event perception, demonstrating that AV^2A achieves substantial improvements over naive training-free baselines. We demonstrate the effectiveness of AV^2A on both zero-shot and weakly-supervised state-of-the-art methods, achieving notable improvements in performance metrics over existing approaches. Our code is available on Github.*

1. Introduction

Video event recognition, which involves the temporal localization and classification of events using both audio and visual cues (see Figure 1), is essential for a wide range of applications, including video content analysis, autonomous driving, multimedia indexing, and personalized content recommendation systems [5, 27, 30]. Despite the recognized importance of this task, current approaches face substantial limitations that hinder their effectiveness and adaptability in open-world scenarios.

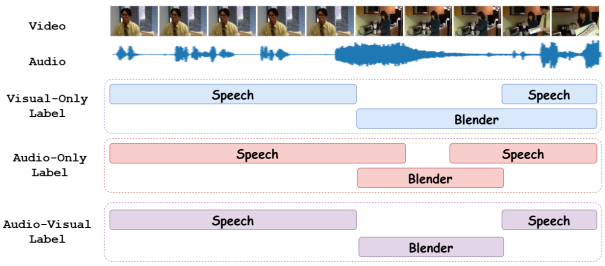

Figure 1. Overview of the AVVP task. Audio-visual event perception focuses on predicting the temporal boundaries of events within a video that are exclusively visible (shown in blue), exclusively audible (shown in red), or both audible and visible (shown in purple).

A primary limitation in previous state-of-the-art methods [2, 5, 10, 12, 18, 19, 21] is their inability to dynamically adjust to evolving events in the video, leading to an oversight of changes in event distribution over time. This issue stems from the reliance on fixed thresholds for event prediction, which restricts the model's capacity to adapt to evolving patterns and contextual shifts within the video sequence. Moreover, the current state-of-the-art methods often employ a late fusion strategy, where audio and visual information are processed independently and then merged by intersecting the predicted labels from both modalities to obtain audio-visual predictions. While this approach simplifies model design, it limits the potential to capture the rich multimodal interactions needed for complex event detection [5, 18].

Another major limitation of existing methodologies [2, 5, 10, 12, 18, 19, 21] is their reliance on fixed-vocabulary datasets. This reliance significantly restricts these mod-

[*]Equal Contribution. Order determined by a coin flip.
Corresponding authors: shaarei@biu.ac.il,arielshaulov@mail.tau.ac.il

els' ability to generalize to novel or evolving event categories—a capability that is essential for open-world environments where the range of possible event types is inherently diverse and continuously evolving [2, 30]. Addressing this challenge requires developing models that can move beyond pre-defined vocabularies and adapt to new events without extensive re-training. Lastly, the annotation process is labor-intensive, time-consuming, and costly, making it difficult to scale existing approaches to larger datasets or to adapt them for open-world deployment [12].

To address these challenges, we propose a model-agnostic approach requiring no additional training, termed **A**udio-**V**isual **A**daptive **V**ideo **A**nalysis (AV^2A). Our method integrates a score-level fusion mechanism that combines audio and visual cues at an initial processing stage, thereby preserving richer multimodal interactions essential for accurate and context-aware event recognition. This approach enhances the alignment of audio and visual information, improving the model's capacity to recognize complex event patterns. Additionally, AV^2A incorporates a novel within-video label-shift mechanism designed to operate within video contexts, leveraging both input video data and predictions from prior frames to dynamically adjust event distributions for subsequent frames, achieved via dynamic thresholds that adapt over time. By aligning model expectations with recurring or evolving events, this mechanism enables AV^2A to capture temporal dependencies. In scenarios where events frequently recur or evolve, the dynamic thresholds capability ensures that the model remains adaptable and responsive to changing contexts. Furthermore, by leveraging foundation models [4, 22, 37] we introduce the first training-free, open-vocabulary baseline through AV^2A. This approach circumvents the limitations of fixed vocabularies and reduces the need for additional data annotations. Our primary contributions are as follows:

- We present AV^2A, a model-agnostic approach that incorporates a score-level fusion strategy to combine audio and visual cues, preserving critical multimodal interactions. Additionally, our approach applies a within video label-shift mechanism that dynamically adjusts the predicted event distribution based on prior context, enhancing temporal adaptability.
- We introduce the first training-free approach specifically tailored for open-world audio-visual event perception, capable of recognizing and adapting to unseen events without the constraints of a fixed vocabulary.
- Through comprehensive evaluations on the AVE [27] and LLP [28] datasets, we demonstrate that our approach achieves state-of-the-art results in both zero-shot and training-free settings, showcasing its robustness and versatility.
- We illustrate the efficacy of our label-shift mechanism through experimental applications on existing state-of-the-art weakly supervised methods, demonstrating consistent improvements across all metrics without additional training. These findings underscore the potential of leveraging temporal patterns directly within existing models.

2. Related Work

Weakly-supervised Audio-Visual Event Perception. To fully exploit the potential of both audio and visual modalities for weakly-supervised video analysis, [27] introduce the audio-visual event (AVE) localization task to leverage both modalities for weakly-supervised video understanding. In AVE, the model detects when an event is both audible and visible, pinpointing its temporal boundaries. Approaches often use attention strategies [11, 31, 32, 34] or cross-modal interactions [14, 23–25], while others incorporate regularization like background suppression [33] and positive sample propagation [36]. Unlike AVE, the audio-visual video parsing (AVVP) task [28] accounts for misalignment between audio and visual signals. To address this, multimodal multiple instance learning (MMIL [28]) frameworks have been employed [2, 5, 10, 12, 18, 19, 21], enabling the model to identify relevant instances across modalities without the need for strict temporal synchronization. These methods typically include techniques such as audio-visual track swapping [30] and cross-modality enhancement [12]. Currently, most methods target either AVE or AVVP separately due to the different properties between the two tasks. Furthermore, all of the above methods are restricted to a closed vocabulary, requiring extensive training to recognize events, which limits their ability to generalize to new, unseen events.

Audio & Visual Zero-Shot Methods. Audio-visual zero-shot learning was first introduced in [20] with the Coordinated Joint Multimodal Embedding (CJME) model, which aligns video, audio, and text in a shared feature space using triplet loss. Building on this, [15] developed the Audio-Visual Generalized Zero-shot Learning Network (AVGZSLNet), which employs a cross-modal decoder and composite triplet loss to enhance the alignment of audio and video embeddings with text. The Audio-Visual Cross-Attention (AVCA) framework [17] further facilitates information exchange between modalities, achieving state-of-the-art performance in audio-visual zero-shot classification and introducing three novel benchmarks for AV-ZSL. Additionally, the Temporal Cross-Attention Framework (TCAF) [16] extends AVCA by leveraging temporal information. While [3] shows promising signs for zero-shot audio-visual event perception, the field remains largely underexplored, with significant opportunities for deeper exploration and advancement.

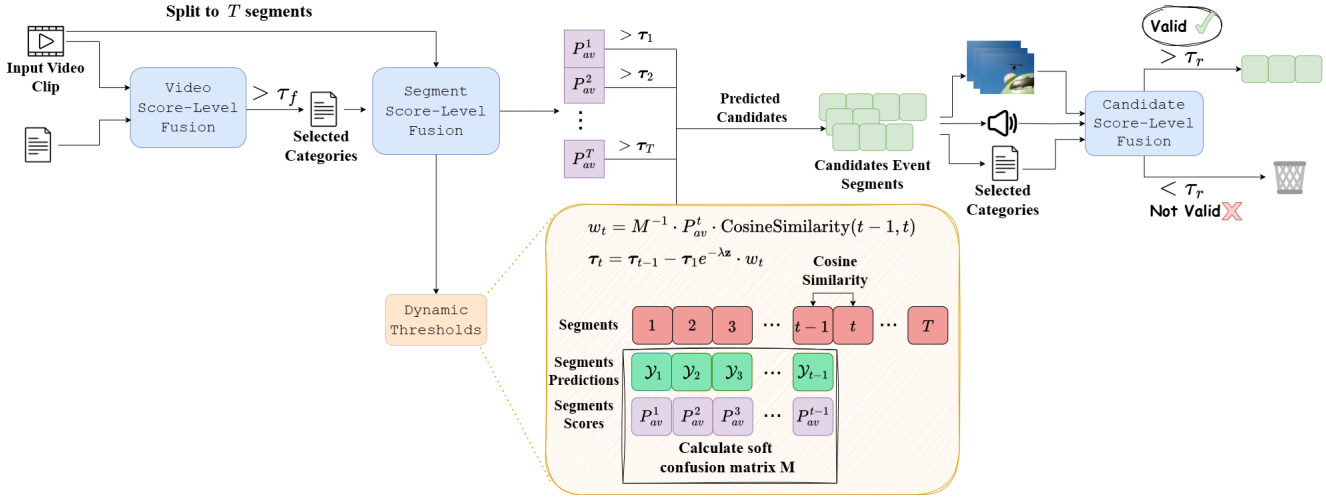

Figure 2. Overview of AV^{2}A. The process begins with category selection, where the input video clip passes through a video-level score-level fusion module (blue) to select relevant categories based on a threshold τ_f. These categories guide segment-level score-level fusion, where a dynamic threshold module (orange) updates thresholds $\boldsymbol{\tau}_t$ via our label-shift technique, using the soft confusion matrix M from prior predictions $\mathcal{Y}_1, \ldots, \mathcal{Y}_{t-1}$ and segment scores $P_{av}^1, \ldots, P_{av}^{t-1}$, cosine similarity between segments and P_{av}^t. Finally, predicted candidates are validated against a confidence threshold τ_r, retaining only those above it. This figure illustrates the process for audio-visual events; audio and visual events are handled similarly.

Label Shift Estimation. Label shift estimation refers to the task of inferring the label distribution in a target domain using labeled source data and unlabeled samples from the target domain [6]. Earlier methods [1, 26, 35] often rely on explicit modeling of the conditional probability $P(X = x \mid Y = y)$, a requirement that becomes impractical for high-dimensional data, such as images. More recently [13] proposed BBSE, a method that leverages a black box predictor $\hat{f}$ to estimate label shift between source and target distributions. Using predictions from $\hat{f}$, BBSE constructs a confusion matrix $M = p_s(\hat{y}, y) \in \mathbb{R}^{k \times k}$ on source data with k classes, where y is the true label and $\hat{y}$ is the predicted label, which may be a hard label ($\hat{y} = \arg\max \hat{f}(x)$) or a soft probability vector, $\hat{f}(x) = [p(\hat{y} = 1|x), \ldots, p(\hat{y} = k|x)]$, which is the distribution over classes predicted by $\hat{f}$ for a given input x. The label shift estimate $\hat{w}$ is then calculated as $\hat{w} = M^{-1}\hat{\mu}$, where M is the estimated confusion matrix and $\hat{\mu}$ represents the distribution of predicted labels in the target data, derived by applying $\hat{f}$ to the target samples. This correction factor $\hat{w}$ aligns the target label distribution with the source distribution. In contrast to prior approaches, we explore the integration of input data as a component within our label shift estimation method. To the best of our knowledge, this work is the first to adapt label shift estimation within video contexts.

3. Method

We present a framework for audio-visual event perception that is entirely training-free. Central to our approach is a dynamic thresholds mechanism, which adjusts threshold values over time-based on the input video data and previous segment predictions, see Figure 2. The method enables precise detection of audio, visual, and audio-visual events.

Problem Setup And Notations. Given an input video divided into T non-overlapping segments, our objective is to temporally localize and identify audio, visual, and audio-visual events present for each segment $1 \leq t \leq T$. Each segment t can contain zero or more events, which makes this a multi-label problem. The events to be identified are drawn from a set of C possible categories, defined for each benchmark. The Given video consists of audio A and video frames V:

$$A = \{a_1, a_2 \ldots, a_T\}, \quad V = \{v_1, v_2, \ldots, v_T\}$$
$$a_t \in \mathbb{R}^{1 \times F}, \quad v_t \in \mathbb{R}^{3 \times H \times W}$$

where a_t and v_t are the corresponding audio and video visual frame at segment t, H and W are the height and width for each frame, and F is number of frequency bins for each audio snippet a_t.

We define two types of vector scores: video-level vector scores and segment-level vector scores. Both types are calculated by a foundation model (FM), combined with a sigmoid function σ to address the multi-label classification problem. Our method operates in an open-vocabulary setting, where the label set C is defined per video at inference time, even ones not seen during hyper-parameter tuning, unlike the baselines which are restricted to labels from

a predefined set. Instead, C consists of user-defined textual descriptions, such as "an image of a military helicopter" or "the sound of helicopter rotor blades".

Video-level scores are represented by P_a and P_v for audio and visual data, from the entire video's audio A and visual V.

$$P_a = \sigma(FM(C, A)), \quad P_v = \sigma(FM(C, V)) \tag{1}$$
$$P_a, P_v \in [0,1]^{|C|}$$

Segment-level vector scores capture more localized information. These are denoted as P_a^t and P_v^t for audio and visual scores, respectively, in each segment t based on segment-specific audio a_t and video frame v_t.

$$P_a^t = \sigma(FM(C, a_t)), \quad P_v^t = \sigma(FM(C, v_t)) \tag{2}$$
$$P_a^t, P_v^t \in [0,1]^{|C|}$$

Each element in both types of score vectors represents the score that a specific category c occurs.

Effective audio-visual event recognition is based on the ability to leverage complementary information from both audio and visual modalities. Our score-level fusion is straightforward, we integrate audio and video scores using a weighted sum to create an audio-visual representation.

$$P_{av} = \alpha P_a + (1-\alpha) P_v \tag{3}$$

$$P_{av}^t = \alpha P_a^t + (1-\alpha) P_v^t \tag{4}$$
$$P_{av}, P_{av}^t \in [0,1]^{|C|}$$

where P_{av}, P_{av}^t are the audio-visual vector scores at the video-level and at segment-level t respectively, and $\alpha \in [0,1]$ serves as a scalar weight that determines the contribution of the audio score P_a, P_a^t relative to the visual score P_v, P_v^t.

For generality, we will focus on identifying and localizing audio-visual events from this point forward. It is important to note that the processes for localizing and identifying audio and visual events are conducted in the same manner, using $\alpha = 1$ for audio events and $\alpha = 0$ for visual events in Equations 3 and 4.

In our framework, we first want to ensure that only the relevant event categories are retained. We apply a selection step for the video-level vector P_{av}, discarding categories that do not exceed a predefined threshold τ_f. This ensures us that only the most relevant event categories $\hat{\mathcal{C}}_{av}$ are preserved for further analysis.

$$\hat{\mathcal{C}}_{av} = \{c \mid P_{av}[c] > \tau_f, \quad \forall c \in C\}, \tag{5}$$

where $P_{av}[c]$ represents the audio-visual classification score for category c is being presented in the video.

Method	α	τ^0	τ_r	τ_f	λ
AV^{2}A (CLIP [22] + CLAP [4])	0.45	0.75	0.75	0.5	1.0
AV^{2}A (LanguageBind [37])	0.5	0.75	0.75	0.55	2.5
AV^{2}A (MGN-MA [18])	0.3	0.55	-	-	0.5
AV^{2}A (JoMoLD [2])	0.4	0.5	-	-	0.5
AV^{2}A (CMPAE [5])	0.45	0.5	-	-	0.5

Table 1. Hyperparameters for each baseline using AV^{2}A.

To adapt to the content of the video domain, we apply thresholds for each category $c \in \hat{\mathcal{C}}_{av}$ that change over time. We initialize the threshold for each category with the same predefined value, denoted as τ^0, which serves as the initial threshold at segment $t = 1$.

$$\boldsymbol{\tau}_{t=1} = [\tau^0 \mid \forall c \in \hat{\mathcal{C}}_{av}]$$

Then, we predict events for segment t in the following way:

$$\mathcal{Y}_t = \left[y_c^t \mid \forall c \in \hat{\mathcal{C}}_{av}\right] \tag{6}$$

$$\text{Where} \quad y_c^t = \begin{cases} 1 & \text{if } P_{av}^t[c] > \boldsymbol{\tau}_t[c], \\ 0 & \text{otherwise} \end{cases}$$

where $P_{av}^t[c]$ represents the audio-visual score of category c being presented at segment t and $\boldsymbol{\tau}_t[c]$ is the category threshold for category c at segment t.

To modify our dynamic thresholds, we estimate for each segment t the ratio $\hat{w}_t$ between the predictions generated from segments $1, \ldots, t-1$ and those from segment t across each category c.

To perform this estimation, we first define a soft confusion matrix M, where each entry represents the score that a given segment t has the true category c but it's predicted category is $\hat{c}$, averaged over all samples. Let $M_p \in \mathbb{R}^{(t-1)\times|\hat{\mathcal{C}}_{av}|}$ be the matrix of prediction scores where each row is given by P_{av}^t (Eq. 4), which is the prediction score vector for segment t. We also define $M_y \in \mathbb{R}^{(t-1)\times|\hat{\mathcal{C}}_{av}|}$ as the matrix of the predicted categories where each row $\mathcal{Y}_t$ is predictions for segment t. The soft confusion matrix $M \in \mathbb{R}^{|\hat{\mathcal{C}}_{av}|\times|\hat{\mathcal{C}}_{av}|}$ is define as:

$$M = \frac{1}{t-1}(M_p)^T \cdot M_y\,,$$

i.e., each entry $M_{\hat{c},c}$ in M is given by:

$$M_{\hat{c},c} = \frac{1}{t-1}\sum_{i=1}^{t-1} P_{av}^i[\hat{c}] \cdot y_c^i\,.$$

$M_{\hat{c},c}$ captures the average score up to segment $t-1$ for instances where the model's predicted score for the category $\hat{c}$ aligns with segments where the predicted category is c. Additionally, we propose incorporating the input data as a

	Audio		Visual		Audio Visual		Type@AV		Event@AV	
	segment	event	segment	event	segment	event	segment	event	segment	event
CLIP [22] + CLAP [4]	17.6	15.1	18.8	17.8	30.0	25.6	22.1	19.5	18.3	16.2
with AV^{2}A	32.7 ± .01	28.5 ± .01	43.5 ± .01	41.5 ± .01	55.1 ± .06	48.2 ± .04	43.8 ± .04	39.4 ± .01	34.7 ± .01	31.2 ± .01
LanguageBind [37]	20.3	17.5	21.3	20.0	24.3	21.0	22.0	19.5	20.8	18.4
with 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 2. Comparison of training-free methods on the LLP dataset, reporting AVVP metrics.

scaling factor for $\hat{w}_t$. We use cosine similarity between the visual features of the current segment t, and the previous segment $t-1$ as our scaling factor. The estimated ratio $\hat{w}_t$ is computed as follows:

$$\hat{w}_t = M^{-1} P_{av}^t \cdot \frac{x_t \cdot x_{t-1}}{\|x_t\| \|x_{t-1}\|} \quad , \tag{7}$$

where M is the estimated soft confusion matrix, P_{av}^t is the score vector for segment t and x_t is denoted as the corresponding visual feature vector produced from the foundation model. The estimated ratio $\hat{w}_t$, computed independently for each test video at inference time, quantifies the change in the score for each category c between consecutive segments $t-1$ and t. Specifically, the ratio captures how the score vector P_{av}^t for the current segment t has shifted relative to the previous segments. The cosine similarity serves as a scaling factor, providing a measure of alignment between the feature vectors of adjacent segments. This ensures that if two segments are highly similar in content, any score changes are emphasized less, while more significant score changes in dissimilar segments are weighted more heavily. Thus, the ratio $\hat{w}_t$ not only reflects changes in category scores but also incorporates segment continuity, offering a refined measure of temporal dynamics in the event patterns across segments.

Following this, we update the category thresholds $\boldsymbol{\tau}_t$ for each segment t. This update employs an exponential decay function to gradually reduce the thresholds, ensuring that the adjustments become smaller over time and preventing excessive reductions as the process progresses. Here, $\boldsymbol{\tau}_1$ represents the initial threshold vector, while $\mathbf{z}$ is a vector tracking the cumulative count of occurrences for each category $c \in \hat{\mathcal{C}}_{av}$ up to time $t-1$. The exponential decay, $e^{-\lambda \mathbf{z}}$, is applied element-wise, meaning that each component z_i in $\mathbf{z}$ undergoes a separate transformation $e^{-\lambda z_i}$, allowing each category's threshold adjustment to be scaled individually according to its occurrence frequency. Here, λ is the decay constant in the exponential decay function, controlling the rate at which threshold adjustments decrease.

$$\boldsymbol{\tau}_t = \boldsymbol{\tau}_{t-1} - \boldsymbol{\tau}_1 e^{-\lambda \mathbf{z}} \cdot \hat{w}_t \tag{8}$$

$$\boldsymbol{\tau}_t, \mathbf{z}, w_t \in \mathbb{R}^{1 \times |\hat{\mathcal{C}}_{av}|} \quad ,$$

After obtaining all $\mathcal{Y}_1, \ldots, \mathcal{Y}_T$, we define a candidate S_k, which is the predicted audio-visual localization for event category c, where k represents the index of the multiple predicted candidates. S_k is constructed to capture the maximal-length interval during which the score $P_{av}^t[c]$ remains above $\boldsymbol{\tau}_t[c]$ to each segment t, meaning that $y_c^t = 1$ holds true continuously throughout the entire candidate duration, such as:

$$S_k = (i_c^n, j_c^n, c) \quad \text{such that} \quad \exists n \leq T,\ \exists c \in \hat{\mathcal{C}}_{av} \quad ,$$

Where:

$$i_c^n = \min\{t \mid y_c^t = 1,\ t > j_c^{n-1}\}$$

$$j_c^n = \max\{t \mid y_c^t = 1,\ t \geq i_c^n\}$$

$$\text{such that} \quad y_c^t = 1, \quad \forall t \in [i_c^n, j_c^n] \quad .$$

Here, i_c^n and j_c^n denote the start and end indices of the candidate S_k, respectively, and c represents the event category associated with the candidate S_i and n is the number of candidates for the same category c.

To further refine the predicted candidates, each candidate S_k is validated based on its associated audio A' and video frames V', where:

$$A' = \{a_i, \ldots, a_j\}, \quad V' = \{v_i, \ldots, v_j\} \quad .$$

Here, i and j represent the start and end timestamps of the predicted candidate. The candidate scores are computed as:

$$P_a' = \sigma(FM(\hat{\mathcal{C}}_{av}, A')), \quad P_v' = \sigma(FM(\hat{\mathcal{C}}_{av}, V')) \quad , \tag{9}$$

where P_a' and P_v' denote the audio and visual scores for all categories $\hat{\mathcal{C}}_{av}$ between timestamps i, j. We compute the audio-visual score P_{av}', which is computed as described in Equation 3 with P_a' and P_v'. We apply a confidence threshold, τ_r, to the audio-visual score $P_{av}'[c]$ for the predicted category c associated with a candidate. If $P_{av}'[c]$ surpasses τ_r, the candidate is deemed valid; otherwise, it is discarded. Our framework employs a total of $T \cdot |\hat{C}_{av}| + 2$ thresholds, specifically $\{\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_T\}, \tau_f$, and τ_r. The output of our framework consists of these valid candidates.

	Audio		Visual		Audio-Visual		Type@AV		Event@AV	
	segment	event	segment	event	segment	event	segment	event	segment	event
MGN-MA [18]	60.7	51.0	55.5	52.4	50.6	44.4	55.6	49.3	57.2	49.2
with AV^{2}A's dynamic thresholds	60.8	51.0	55.6	52.7	51.9	45.5	56.1	49.7	57.3	49.3
JoMoLD [2]	61.3	53.9	63.8	59.9	57.2	49.6	60.8	54.5	59.9	52.5
with AV^{2}A's dynamic thresholds	61.4	53.9	64.1	61.0	57.7	50.9	61.1	55.2	60.0	52.7
CMPAE [5]	64.4	57.5	66.1	63.3	59.5	52.4	63.3	57.7	62.9	56.3
with AV^{2}A's dynamic thresholds	**64.7**	**58.0**	**66.3**	**63.4**	**59.9**	**52.7**	**63.6**	**58.0**	**63.0**	**56.7**

Table 3. Performance of state-of-the-art weakly supervised methods and their improvements with AV^{2}A's dynamic thresholds mechanism on the LLP dataset.

Method	Accuracy(%)
DG-SCT [3]	64.7
CLIP [22] + CLAP [4]	38.2 $\pm$ 2.87
with AV^{2}A	<u>68.5</u> $\pm$ 0.16
LanguageBind [37]	32.9 $\pm$ 1.27
with AV^{2}A	**72.8** $\pm$ 0.14

Table 4. Comparison with zero-shot methods on AVE datasets, reporting AVEL metrics.

4. Experiments

We assess the performance of AV^{2}A on two widely adopted datasets, AVE [27] and LLP [28], through comprehensive experiments demonstrating its improvements in both training-free and zero-shot settings. Additionally, we apply our novel dynamic thresholds technique to state-of-the-art weakly supervised methods, achieving significant performance enhancements without additional training. An extensive ablation study further highlights the contribution of each component in our approach.

4.1. Experimental setup

Datasets. We evaluate our approach on two datasets, AVE [27] and LLP [28].

The Audio-Visual Event (AVE) dataset [27] consists of 4,143 YouTube videos spanning 28 categories derived from AudioSet [7], with each video clip lasting 10 seconds. The AVE [27] dataset focuses on recognizing audio-visual events that are both visible and audible across multiple time segments within a video, that is to say, it provides synchronized audio and visual signals, along with precise frame-level annotations that indicate the temporal locations of these events.

The Look, Listen, and Parse (LLP) dataset [28] is a comprehensive audio-visual benchmark designed for studying event perception in unconstrained, real-world videos. It comprises 11,849 YouTube video clips, each lasting 10 seconds and covering 25 diverse event categories that involve various audio-visual interactions. The dataset provides video-level event labels, including both single and multiple overlapping events. The LLP dataset suits weakly-supervised learning, providing video-level annotations without precise event boundaries.

Evaluation Metrics. Consistent with previous work, we evaluate using F1-scores for audio, visual, and audio-visual events at both segment and event levels. Additionally, we report aggregate metrics, Type@AV and Event@AV, computed at segment and event levels as well. For a comprehensive explanation of these metrics, refer to [28]. For the AVE dataset, in line with [27], we use overall accuracy as the evaluation metric.

For hyperparameter optimization, we performed a search over 200 samples from the LLP validation set for the training-free methods, while using the full validation set of LLP for the weakly-supervised methods. Since only AV^{2}A's dynamic threshold mechanism is applied, optimizing the τ_r and τ_f values is unnecessary for the MGN-MA [18], JoMoLD [2], and CMPAE [5] methods. All hyperparameters used for each baseline are listed in Table 1.

4.2. Comparison with State-of-the-art Methods

Baselines. In our approach to training-free compression, we employ three foundational models to create two training-free baseline methods, upon which we implement AV^{2}A. The first baseline employs the LanguageBind model [37], which provides both text-visual and text-audio similarity measurements. The second baseline combines the CLIP [22] and CLAP [4] models, where CLIP captures text-visual similarities, and CLAP supplies text-audio similarities; we refer to this baseline as CLIP-CLAP. For the evaluation of the CLIP-CLAP and LanguageBind baselines, a single static threshold is applied to predict categories across segments, creating two naive baselines. For weakly-supervised compression, we adopt three methodologies: JoMoLD [2],

Figure 3. Performance analysis of AV^{2}A, based on the LanguageBind [37] model, showcasing predictions on audio-visual events, specifically those occurring simultaneously in both audio and video. Comparisons highlight (a) improvements and (b) failure-cases relative to state-of-the-art weakly supervised baselines.

which performs joint modality learning for improved cross-modal representations; MGN-MA [18], which leverages multi-granularity networks to enhance modality alignment; and the state-of-the-art method CMPAE [5], that integrates adaptive encoding to capture complex, high-quality cross-modal associations.

AVVP Evaluation. In this experiment, we employ two baselines, CLIP-CLAP and LanguageBind, which are constructed as described in Section 4.2. As shown in Table 2, applying AV^{2}A to the baselines CLIP-CLAP and LanguageBind results in metric improvements across the board. Specifically, when AV^{2}A is integrated with the naive CLIP-CLAP baseline, it achieves a 21.7% improvement in segment-level Type@AV, which measures average segment performance across audio, visual, and audio-visual settings. Additionally, AV^{2}A enhances performance by 19.9% in event-level Type@AV, which evaluates the average event performance across these setups. Furthermore, AV^{2}A improves the naive LanguageBind baseline by 30.4% and 28% in segment-level and event-level Type@AV, respectively.

We further validate AV^{2}A's dynamic thresholds mechanism on the AVVP benchmark by applying it on top of current state-of-the-art weakly supervised methods, specifically, CMPAE [5], JoMoLD [2], and MGN-MA [18]. In this experiment, we apply the dynamic thresholds mechanism (Eq. 8) to these methods, as they already employ a relevant class selection stage that closely aligns with our approach; however, these methods used static thresholds to predict events over time. The results, presented in Table 3, indicate that our dynamic thresholds outperform the static thresholds technique without additional training, establishing a new state-of-the-art in video parsing. To imple-

Method	Audio		Visual		Audio Visual		Type@AV		Event@AV	
	segment	event	segment	event	segment	event	segment	event	segment	event
AV^{2}A(LanguageBind [37])	**40.9**	**35.9**	**57.4**	**54.4**	**59.1**	**52.3**	**52.4**	**47.5**	**43.4**	**38.9**
w/o cosine similarity	40.8	**35.9**	57.3	54.3	58.8	51.8	52.3	47.3	43.3	38.8
w/o dynamic thresholds	40.7	35.1	56.9	55.2	55.1	47.2	52.2	47.2	43.0	37.8
w/o refine segments	38.9	34.1	53.6	50.9	57.9	51.1	50.2	45.4	41.4	37.0
w/o relevant class selection	33.4	29.3	40.2	38.4	39.7	34.3	37.8	34.0	34.8	31.4

Table 5. Ablation study results on the LLP dataset.

ment our dynamic threshold mechanism in the audio-visual regime, it is essential to apply our score-level fusion stage (Eq.4), as the audio-visual scores required for the dynamic thresholds mechanism are unavailable in the methods standard form due to their late-fusion strategy.

AVE Evaluation. As shown in Table 4, we compare AV^{2}A with three zero-shot baselines: the CLIP-CLAP baseline [4, 22], the LanguageBind baseline [37], and the state-of-the-art DG-SCT method [3]. Notably, AV^{2}A outperforms DG-SCT [3], achieving improvements of 3.8% and 8.1% when built upon the CLIP-CLAP and LanguageBind baselines, respectively. To ensure a fair zero-shot comparison, we used the same hyperparameters identified during tuning on the LLP validation set, as detailed in Section 4.2. Please note that we did not apply AV^{2}A on the DG-SCT baseline, as the trained checkpoints and features were not made available.

Method Efficiency. In our experiments, inference times across all weakly-supervised methods were of the same order of magnitude, with our method achieving $\sim$ 2.49 seconds per video. However, our primary efficiency advantage lies in the fact that AV^{2}A requires no additional training, unlike the weakly-supervised baselines. For comparison, the weakly-supervised methods trainable parameters are as follows: MGN-MA [18] has $\sim$ $4.3M$, JoMoLD [2] has $\sim$ $4.5M$, and CMPAE [5] has $\sim$ $5.6M$. In addition to the trainable parameters, the above methods employ three pretrained, fixed models to generate input features: ResNet152 [8], VGGish [9], and R(2+1)D [29]. This procedure adds $\sim$ 2.68 seconds to the inference time per video.

4.3. Ablation study

We conducted an ablation study using the LanguageBind baseline [37] to investigate the contributions of each component of our method. The results, detailed in Table 5, illustrate the impact of these components on performance. The first row presents our complete method, with the LanguageBind model [37]. The second row of the table represents the model configuration without cosine similarity in the label shift method, effectively demonstrating the method's performance when excluding the cosine similarity between adjacent frames. The third row highlights the effect of omitting the dynamic threshold mechanism, employing fixed thresholds instead. Additionally, the fourth row reports the outcome when the refinement of segments is removed, resulting in candidate event segments being used as final predictions. Lastly, we examined the model without the relevant class selection process, using the entire class set from the outset. Note that in each row, only the specified stage is pruned, while the remainder of the pipeline remains unchanged. The findings indicate that each component of our method contributes positively to the overall performance, reinforcing the importance of their inclusion.

5. Limitations

Our reliance on foundation models for extracting audio-visual similarities may introduce biases and limit the accuracy of short-duration or rapid events (Figure 3). These models, trained on longer, distinct events, struggle with subtle changes. Since AV^{2}A requires no training or fine-tuning, improvements in foundation models should enhance audio-visual perception.

6. Conclusion

In this paper, we introduce the first training-free, open-vocabulary approach for audio-visual perception via AV^{2}A, a model-agnostic method for robust event perception. AV^{2}A improves weakly-supervised baselines by replacing static thresholds with a within-video label shift mechanism, enhancing adaptability to temporal variations in event distribution. Applied to zero-shot and weakly-supervised settings, AV^{2}A achieves new state-of-the-art results without additional training.

7. Acknowledgments

This work was supported by the Tel Aviv University Center for AI and Data Science (TAD), and by a Vatat data-science grant to GC and the Bar-Ilan University data-science institute.

References

[1] Yee Seng Chan and Hwee Tou Ng. Word sense disambiguation with distribution estimation. In *IJCAI*, pages 1010–5, 2005. 3

[2] Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision*, pages 431–448. Springer, 2022. 1, 2, 4, 6, 7, 8

[3] Haoyi Duan, Yan Xia, Zhou Mingze, Li Tang, Jieming Zhu, and Zhou Zhao. Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 6, 8

[4] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2, 4, 5, 6, 8

[5] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18827–18836, 2023. 1, 2, 4, 6, 7, 8

[6] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300, 2020. 3

[7] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 6

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8

[9] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 8

[10] Xun Jiang, Xing Xu, Zhiguo Chen, Jingran Zhang, Jingkuan Song, Fumin Shen, Huimin Lu, and Heng Tao Shen. Dhhn: Dual hierarchical hybrid network for weakly-supervised audio-visual video parsing. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 719–727, 2022. 1, 2

[11] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[12] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. *Advances in Neural Information Processing Systems*, 34:11449–11461, 2021. 1, 2

[13] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018. 3

[14] Tanvir Mahmud and Diana Marculescu. Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5158–5167, 2023. 2

[15] Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P Namboodiri. Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3090–3099, 2021. 2

[16] Otniel-Bogdan Mercea, Thomas Hummel, A Sophia Koepke, and Zeynep Akata. Temporal and cross-modal attention for audio-visual zero-shot learning. In *European Conference on Computer Vision*, pages 488–505. Springer, 2022. 2

[17] Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10553–10563, 2022. 2

[18] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. *Advances in Neural Information Processing Systems*, 35: 34722–34733, 2022. 1, 2, 4, 6, 7, 8

[19] Shentong Mo and Yapeng Tian. Semantic-aware multi-modal grouping for weakly-supervised audio-visual video parsing. In *ECCV Workshop*, 2022. 1, 2

[20] Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3251–3260, 2020. 2

[21] Kranthi Kumar Rachavarapu, Kalyan Ramakrishnan, et al. Weakly-supervised audio-visual video parsing with prototype-based pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18952–18962, 2024. 1, 2

[22] 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. 2, 4, 5, 6, 8

[23] Janani Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4372–4376. IEEE, 2020. 2

[24] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *Proceedings of the IEEE/CVF Winter Confer-*

ence on Applications of Computer Vision, pages 2970–2979, 2020.

[25] Varshanth Rao, Md Ibrahim Khalil, Haoda Li, Peng Dai, and Juwei Lu. Dual perspective network for audio-visual event localization. In *European Conference on Computer Vision*, pages 689–704. Springer, 2022. 2

[26] Amos Storkey. When training and test sets are different: characterizing learning transfer. 2008. 3

[27] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. 1, 2, 6

[28] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 436–454. Springer, 2020. 2, 6

[29] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 8

[30] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1326–1335, 2021. 1, 2

[31] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6292–6300, 2019. 2

[32] Yiling Wu, Xinfeng Zhang, Yaowei Wang, and Qingming Huang. Span-based audio-visual localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1252–1260, 2022. 2

[33] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19989–19998, 2022. 2

[34] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 279–286, 2020. 2

[35] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. Pmlr, 2013. 3

[36] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021. 2

[37] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 2, 4, 5, 6, 7, 8