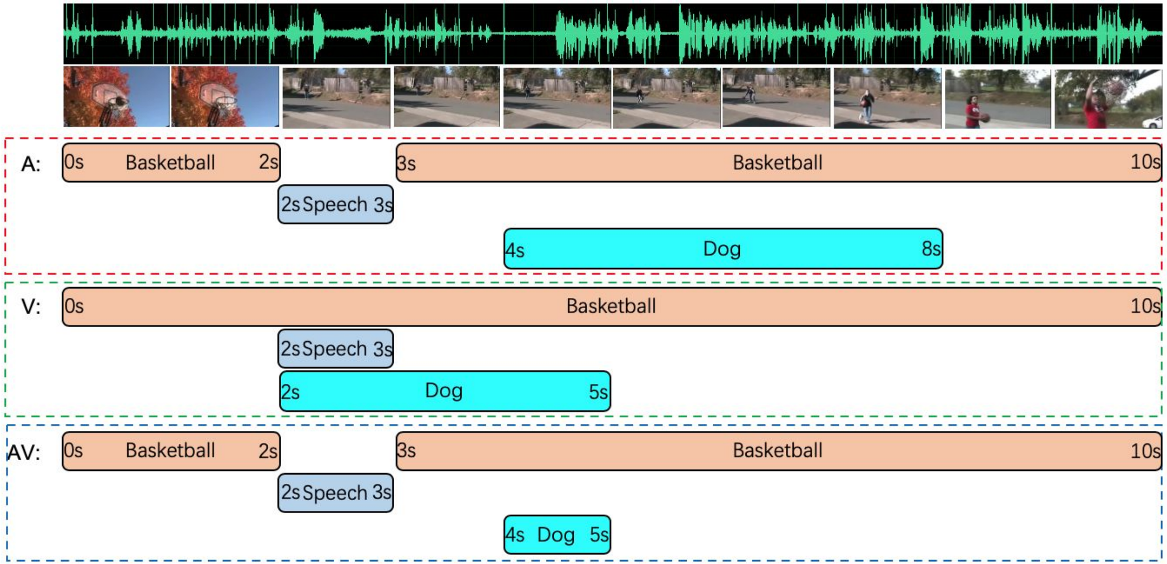
Exploring Heterogeneous Clues for Weakly-Supervised Audio-Visual Video Parsing

By :Yu Wu, Yi Yang

Glorious Purpose-
Prathamesh Sachin Pilkhane
P Balasubramanian
Vaibhav Raj

AVVP - Audio-Visual Video Parsing Task

- Different modalities in a video give us different information
- The task demands labelling of both streams for activities with time boundaries



Audio stream

Visual stream

Audio-Visual stream

Problem Statement

Input

- T seconds audio-visual sequence $S = \{V_t, A_t\}_{t=1}^{T}$ where V_t, A_t are the Visual and Audio data .

Task

- Predict target labels for each segment and over each modality. i.e. estimate $\mathbf{y}_t = (y_t^a, y_t^v, y_t^{av})$

Training data

- Provided with weak labels of events in video without actual details of timings and/or presence of info in modalities

Challenge!!

Method in AVVP:

01 Data Preprocess

- Compute $\mathbf{f}_t^a, \mathbf{f}_t^v$ using pre-trained models.

02 Feature Aggregation

- Compute $\hat{\mathbf{f}_t^a}, \hat{\mathbf{f}_t^v}$ using attention for self & cross modality
- Use of aggregated features proves better results

03 Prediction

- Predict data based on the features using a classifier using Sigmoid function

Attention Function :

Attention function over matrices **q** (of dimension **d**), **K** and **V** is given by :

$$att(q, K, V) = sigmoid(\frac{\mathbf{qK^T}}{\sqrt{d}})\mathbf{V}$$

And the corresponding equations for the **features aggregates** are given by :

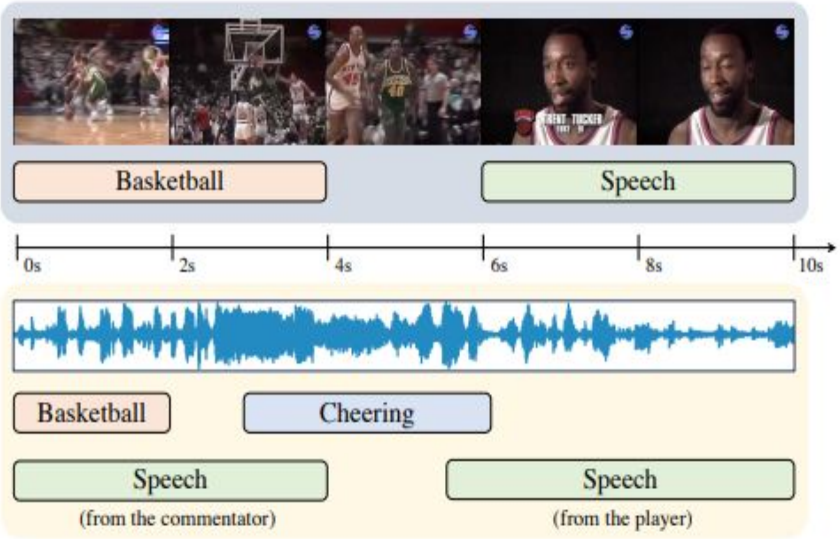
$$\hat{\mathbf{f}_t^a} = \mathbf{f}_t^a + att(f_t^a, F_a, F_a) + att(f_t^a, F_v, F_v)$$
$$\hat{\mathbf{f}_t^v} = \mathbf{f}_t^v + att(f_t^v, F_a, F_a) + att(f_t^v, F_v, F_v)$$

The feature aggregates tend to cover the **presence of similar features** over a global scale as well as across the modalities.

What was wrong then?

The **basic AVVP** model indeed works moderately good , but faces difficulty in the presence of **noise** in the modality data.

Even though the inputs were treated individually, if an **event is almost uncertain** in one data mode, then its computed probability gets **significantly** reduced.



Proposed Solution :

- **Explore heterogeneous (present in only 1 modality) clues in the video.**
  - Example : Commentary occurs in audio alone, while a boxing match has information predominantly in visual form

Main Idea:

- Train and classify events, give more importance if one modality conveys the information evidently.
- Enforce this using the method of **Channel Exchanging**

Exchanging Channels

1. Given two audio-visual sequences $S^i = (V^i, A^i)$. $S^j = (V^j, A^j)$ with different event label tags. Construct $\hat{S}^i_j = (V^i, A^j)$, $\hat{S}^j_i = (V^j, A^i)$, with **exchanged Audio-Visual** data.
2. Thus, event labels for S^i can only be derived from the visual data for $\hat{S}^i_j$ and only from audio data in $\hat{S}^j_i$
3. Evaluate the event labels based on the base model over each of the mixed construction.

$$p^v_{\hat{a}}, p^a_{\hat{a}} = \phi(V^i, A^j)/E_c$$
$$p^v_{\hat{v}}, p^a_{\hat{v}} = \phi(V^j, A^i)/E_c$$

4. Find the predictions for the events - In case less than 0.5 , the corresponding modality is weaker & does not refer to the event assigned.
5. The event tags would now by relabelled , called as **Modality-aware labels.**

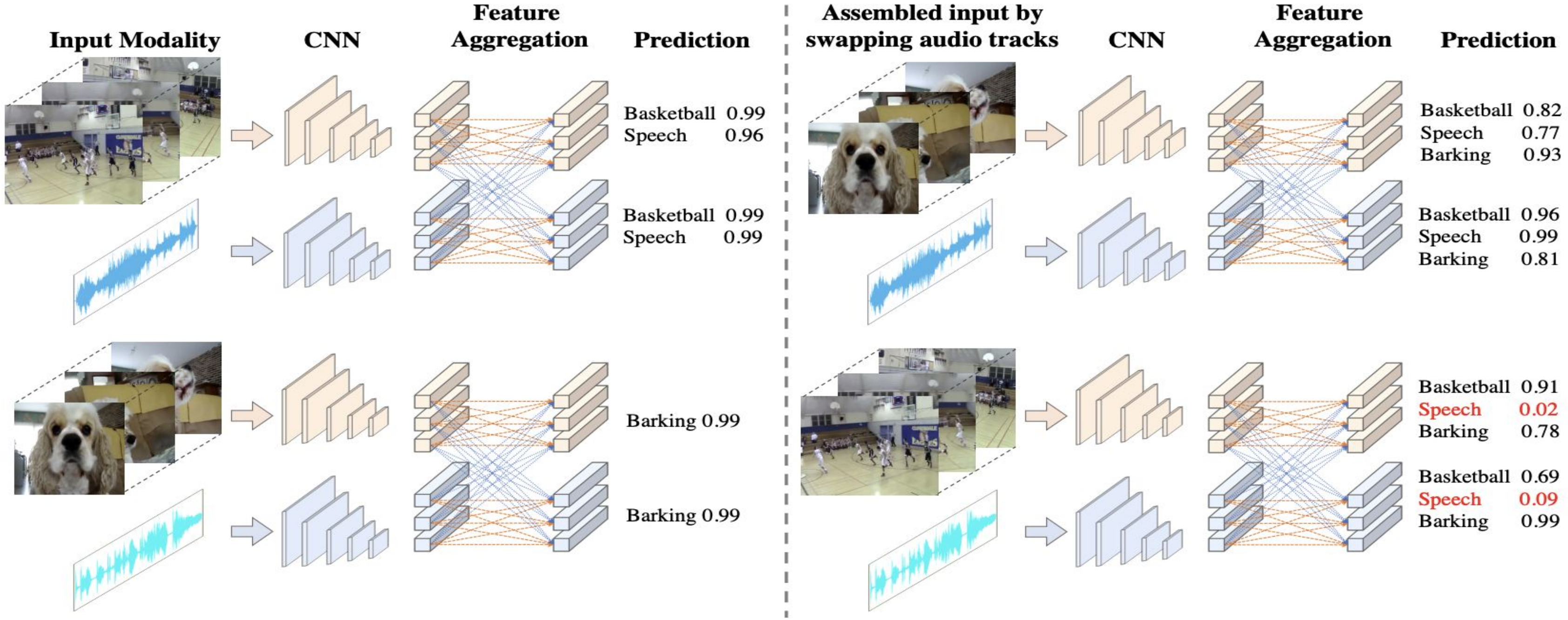


Figure 2. Our modality-aware label refining (MA) pipeline. Our model aggregates feature by self- and cross-modality attention, and then predicts the event labels for each modality. The figure's left shows the prediction on normal training videos, which would have relatively high confidence in their event predictions. We then exchange audio and visual tracks of these two unrelated videos (whose labels are not *disjoint*). The newly assembled videos are further input to the model for checking prediction confidences (right figure). We believe the confidences should still be high if the remaining visual/audio track does contain the target event. Otherwise, the event is not visible/audible in this modality. In this way, we could obtain modality-aware event labels and protect models from being misled by the ambiguous overall labels. In the case shown in the figure, we filter out the "Speech" event that is not visible in the original basketball video.

Defining temporal difference and Losses

- The use of attention may harm the model , specifically this would introduce temporal difference during the weakly-supervised training.
- The absence of temporal annotation can be met up by using **Contrastive learning**.
- A **proxy task** that urges model to pick the correct temporal segment from all distractor segments.
- Our target now becomes to align $\hat{\mathbf{f}}_t^a$ and $\hat{\mathbf{f}}_t^v$ to be much similar at time step t and be comparatively far off in segments with temporal difference.
- **Loss description :** This model uses the following loss :

$$\mathcal{L}_c = -\log \frac{\exp(\mathbf{f}_t^{v\top}\hat{\mathbf{f}}_t^a/\tau)}{\sum_j \exp(\mathbf{f}_j^{v\top}\hat{\mathbf{f}}_t^a/\tau)}$$

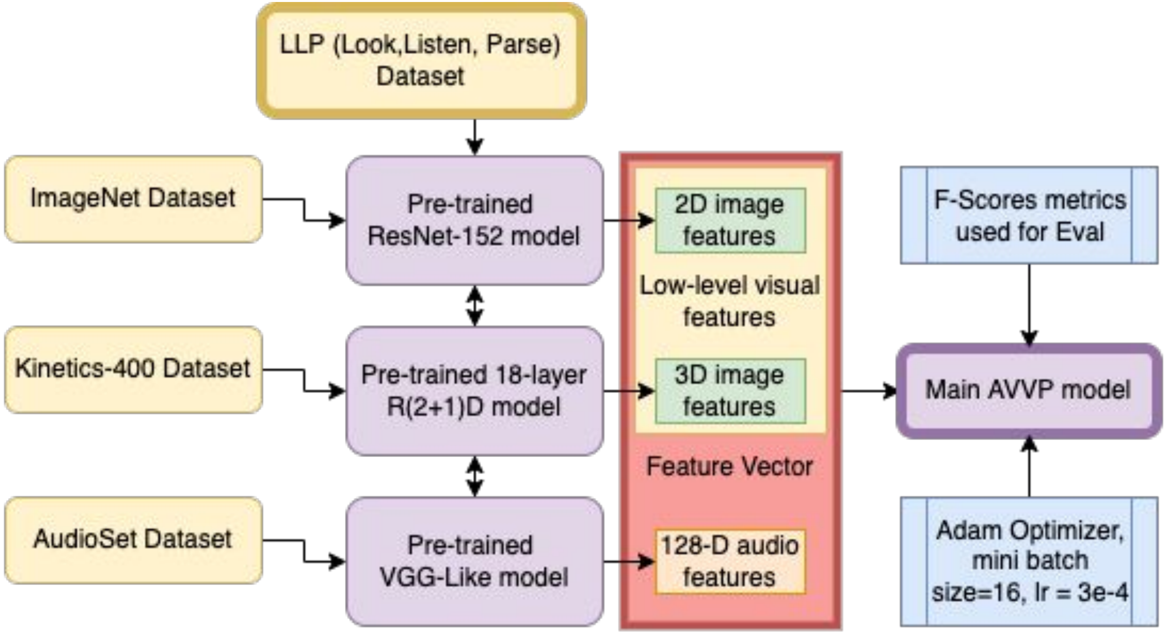
Defining temporal difference and Losses contd…

- In using the previous mentioned formula, we try to implement the fact that the loss function is much higher for values with higher temporal difference.

- τ is the **temperature parameter** and basically determines the concentration of the pdf distribution. Lower value -> Softer distribution.
- We add a component of this loss with the model cross entropy loss to get the total loss.

Implementation Details:

Datasets and Feature Extractors



LLP (Look,Listen, Parse) Dataset

ImageNet Dataset

Pre-trained ResNet-152 model

Kinetics-400 Dataset

Pre-trained 18-layer R(2+1)D model

AudioSet Dataset

Pre-trained VGG-Like model

2D image features

Low-level visual features

3D image features

Feature Vector

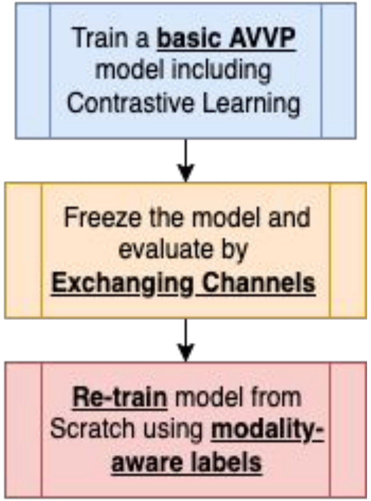
128-D audio features

F-Scores metrics used for Eval

Main AVVP model

Adam Optimizer, mini batch size=16, lr = 3e-4

Training Process

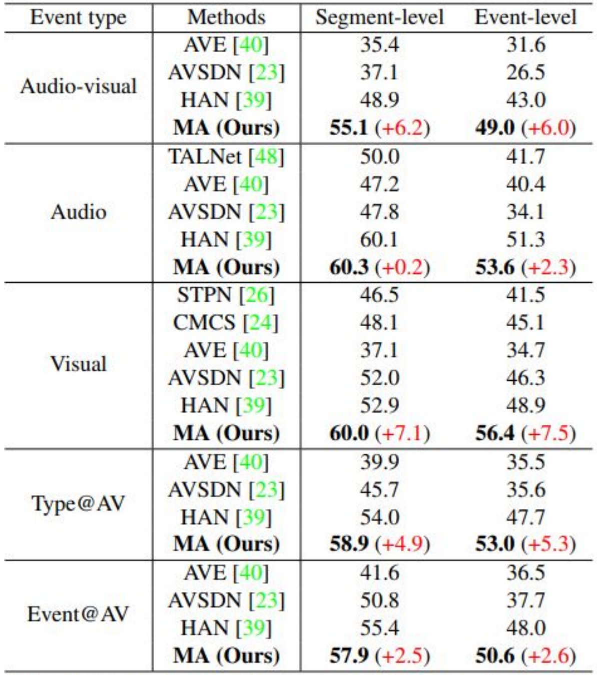


Train a **basic AVVP** model including Contrastive Learning

Freeze the model and evaluate by **Exchanging Channels**

Re-train model from Scratch using **modality-aware labels**

Results and Ablation Study (1)



Event type	Methods	Segment-level	Event-level
Audio-visual	AVE [40]	35.4	31.6
	AVSDN [23]	37.1	26.5
	HAN [39]	48.9	43.0
	MA (Ours)	**55.1** (+6.2)	**49.0** (+6.0)
Audio	TALNet [48]	50.0	41.7
	AVE [40]	47.2	40.4
	AVSDN [23]	47.8	34.1
	HAN [39]	60.1	51.3
	MA (Ours)	**60.3** (+0.2)	**53.6** (+2.3)
Visual	STPN [26]	46.5	41.5
	CMCS [24]	48.1	45.1
	AVE [40]	37.1	34.7
	AVSDN [23]	52.0	46.3
	HAN [39]	52.9	48.9
	MA (Ours)	**60.0** (+7.1)	**56.4** (+7.5)
Type@AV	AVE [40]	39.9	35.5
	AVSDN [23]	45.7	35.6
	HAN [39]	54.0	47.7
	MA (Ours)	**58.9** (+4.9)	**53.0** (+5.3)
Event@AV	AVE [40]	41.6	36.5
	AVSDN [23]	50.8	37.7
	HAN [39]	55.4	48.0
	MA (Ours)	**57.9** (+2.5)	**50.6** (+2.6)

Table 1. Comparisons with the state-of-the-art methods of the audio-visual video parsing task on the LLP test dataset. Note that we use the same input features as the compared methods.

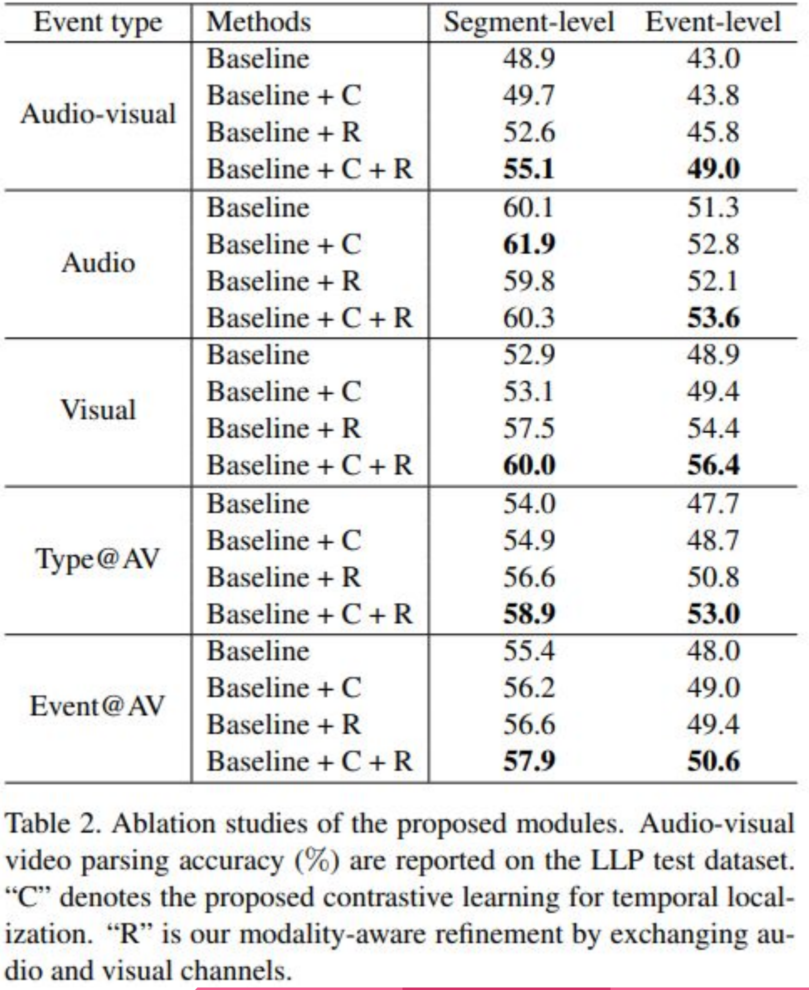
Comparison with other models:

This model outperforms the state-of-the-art audio-visual parsing methods by significant margins, showing the effectiveness of **Exchanging Channels** & **Contrastive Learning**.

Result and Ablation Study (2)

Based on the **Baseline+R** score of our model, we conclude that our separation of audio and visual labelling indeed helps the model perform better.

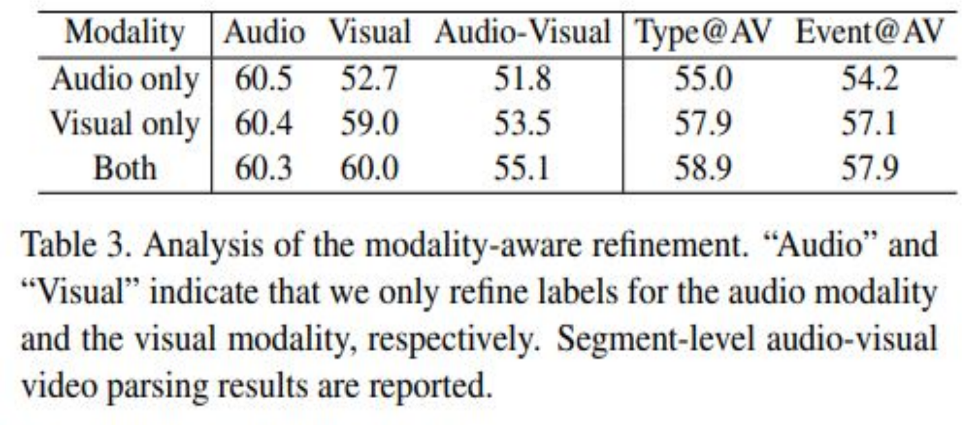
Notice that **C** also contributes to the performance of the model by a decent margin.



Event type	Methods	Segment-level	Event-level
Audio-visual	Baseline	48.9	43.0
	Baseline + C	49.7	43.8
	Baseline + R	52.6	45.8
	Baseline + C + R	**55.1**	**49.0**
Audio	Baseline	60.1	51.3
	Baseline + C	**61.9**	52.8
	Baseline + R	59.8	52.1
	Baseline + C + R	60.3	**53.6**
Visual	Baseline	52.9	48.9
	Baseline + C	53.1	49.4
	Baseline + R	57.5	54.4
	Baseline + C + R	**60.0**	**56.4**
Type@AV	Baseline	54.0	47.7
	Baseline + C	54.9	48.7
	Baseline + R	56.6	50.8
	Baseline + C + R	**58.9**	**53.0**
Event@AV	Baseline	55.4	48.0
	Baseline + C	56.2	49.0
	Baseline + R	56.6	49.4
	Baseline + C + R	**57.9**	**50.6**

Table 2. Ablation studies of the proposed modules. Audio-visual video parsing accuracy (%) are reported on the LLP test dataset. "C" denotes the proposed contrastive learning for temporal localization. "R" is our modality-aware refinement by exchanging audio and visual channels.

Result and Ablation Study (3)



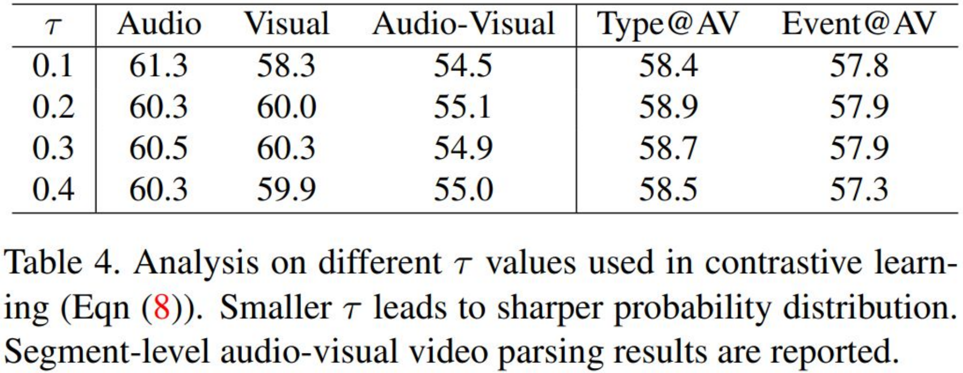
Modality	Audio	Visual	Audio-Visual	Type@AV	Event@AV
Audio only	60.5	52.7	51.8	55.0	54.2
Visual only	60.4	59.0	53.5	57.9	57.1
Both	60.3	60.0	55.1	58.9	57.9

Table 3. Analysis of the modality-aware refinement. “Audio” and “Visual” indicate that we only refine labels for the audio modality and the visual modality, respectively. Segment-level audio-visual video parsing results are reported.

Modality-aware refinement results:

We observed the improvement in performance when measuring segment level visual parsing evaluations. The use of audio information clearly shows how videos can be understood further.

Result and Ablation Study (4)



τ	Audio	Visual	Audio-Visual	Type@AV	Event@AV
0.1	61.3	58.3	54.5	58.4	57.8
0.2	60.3	60.0	55.1	58.9	57.9
0.3	60.5	60.3	54.9	58.7	57.9
0.4	60.3	59.9	55.0	58.5	57.3

Table 4. Analysis on different τ values used in contrastive learning (Eqn (8)). Smaller τ leads to sharper probability distribution. Segment-level audio-visual video parsing results are reported.

Analysis of Contrastive Parameter: There is a very slight increase in performance when τ decreases.

Overall, the effect is not so significant and the value 0.2 is used for all other testing.

Thank You!