Slide 1:
Good morning everybody , our team Glorious Purpose will be presenting the research, Exploring Heterogeneous Clues for Weakly-Supervised Audio-Visual Video Parsing or rather, Clued AVVP. Our team includes me, Prathamesh, Vaibhav , and Balasubramanian.

Slide 2: Before we begin on Clued AVVP, we must get a clear idea on what Clued AVVP came from , basically the AVVP task and the corresponding research paper. So what is Audio Visual Video Parsing? It basically means gathering and identifying data on not just visual information but using audio data as well. For example, while we may miss out on whether a basketball player missed a shot or not visually,but the ambient cheering or booing confirms us about the result. The task of AVVP basically deals with labeling such events in the video. Another reason why AVVP is important is when we have a single mode of data for some duration of time, say the video got blurred / absence of sounds, we can still predict and label the events in the video sequence. Given this background, lets state the problem now more formally,

Slide 3-
In the task, we have been given the video sequence S of Visual data V_t , and audio data A_t for time segment t. We want to label each of these individual segments of 1sec duration. Specifically, we label each segment over each modality of events. Now at a given instance of time we can ofcourse have different events happening simultaneously and hence this problem belongs to multi-class labelling.
Now the existing ideas before AVVP paper assumed a strong correlation that Audio and Visual data at a given segment belongs to same event. AVVP instead assumes the contrary.

Slide 4-
So getting onto the implementation of AVVP . The first step in our method is the preprocessing of our data. We first preprocess our audio and visual data separately using pre existing models , for example Resnet-152 and VGG. We will use the outputs of these networks as our inputs. We assume that the dimension of each feature is of the same dimension (we can always pad one with 0 to match the dimensions for both, but we try to avoid padding (wastage of space)). Now , to further enrich the features to be used, what we do is we find how much the feature is relevant on a global scale. For global scale, we mean in both the modalities of data. To do so, we use the attention function developed in AVVP paper. We will look up into what attention function is for this case. Assuming we have such a feature which captures information over a global scale, we now predict over evaluations using a classifier network as a Sigmoid instead of a softmax , emphasising that our task is a multi-class label task. This outlines our original procedure as introduced in AVVP.
Slide 5-
Coming to the Attention Function, it basically represents whether or not a query is present in the entire scale, and also whether how relevant it is in the segment. The function is defined as such very simplistically ,sigmoid (.....) . Which basically translates to saying, given a vector q of the form 1,1,1,-1,1-,1… we take dot product of it with every vector and take the dot product value as weightage over some set of values. To have a normalising effect of this weights , we use the sigmoid activation. Thus, the new feature aggregates basically will emphasise on the feature more if it is also present on a global scale in the self modality as well in the crossed modality score.
Slide 6-

So what was the issue then? Wasnt AVVP doing fine by itself? Well yes and no. It did perform moderately good, but major problem was that it struggled when data had noise. Basically , when data of one modality was sufficient to convey some information but the other modality became a noise to this. To deal with this is what was the task of Clued AVVP. Handing over to Vaibhav

Slide 7

One solution to deal with this problem is the usage of heterogeneous clues - those which exist only in 1 modality. As you might have seen in the image on the previous slide, basketball commentary was never present in the visual stream, could only be heard. The general approach counted it towards both modalities. All we need is a method to adapt it so that only the correct modality counts it.

For this treatment, we would require refinement of labels and this is a major contribution coming from this paper. If some media stream has no information related to the event, it won't predict that event independently. However, our model treats only audio-visual combinations. So we could add in an unrelated audio and we would still be getting poor probability for that event. We call this swap **Channel Exchanging**.

Slide 8

Whatever we have discussed until now, we will make it more concrete here. Suppose the video S_i has visual and audio streams V_i and A_i. So now we choose 2 unrelated videos. Unrelated as in they don't have any labels in common overall. After swapping their audio-visual streams, we use our good-enough baseline to predict events. The first equation tells us the prediction based on the video when we swapped out our audio. You see two probabilities there - those are predictions based on visual and audio features respectively. When only audio had the clues for the events, we would expect these two probabilities in the first equation to be low. The same applies to the second equation in the case when video had clues. We keep this cutoff as 0.5, if below it, we scrap off the low-probability labellings. The new labels thus obtained are hereby called MODALITY AWARE LABELS.

Special Note - Considering the first case, when we use the cutoff we ensure that both the audio and visual representations give low probability. One might think that the audio was already from an unrelated video and need not be checked. However, there might be some trickling of features from the visual stream to this representation, thanks to feature aggregation step.

Slide 9

We now give an example here - As you can see, we have represented the feature aggregation step with those cross-modal lines. To the left, we have our usual weak labels. However, once we swap the streams, we now have very low probability of speech when the barking sound

replaces this audio. This means that the images, which anyway showed just the basketball players in action, didn't have any audio related to speaking.

Slide 10

This is not just the only novel idea this paper comes up with. Remember from the Problem Statement that we had to specify time boundaries. However, when we used attention, the contrast between audio and visual frames at different times would decrease and time boundaries predicted would be blurry and less accurate. We need to increase this contrast.

To implement this, as is done in a lot of self-supervised models, we use Contrastive Learning. Specifically, we use Noise Contrastive Estimation - we try to differentiate our prime candidate compared to other distractors. Our aim is to make sure the audio data at this timepoint is close specifically to the visual representation here and not to those at other timepoints. Notice that these distractors can be very similar to the ground truth frame.

The loss that we use here is specified herewith, we try increasing the cosine similarity for the prime candidate compared to all frames.

Slide 11

We had a parameter TAU for the temperature there. It controls the shape of the distribution. When it is high, the distribution is steep. Up ahead, we also discuss ablation studies for it.

Finally, we add the contribution of this loss to the binary cross-entropy loss for event labels. It won't so happen that some temporal segment which was favored highly in the attention model goes on to dominate the results.

Bala will now be explaining specific details about the model including dataset and ablation studies.

Slide 12

Implementation Details… The LLP(Look,listen,parse) dataset is used for this AVVP model. It is the most used dataset or baseline for AVVP tasks. It has great range of daily life scenes from animal videos to cars and human, with weak labels for the same. A pre-trained Resnet-152 model trained on the Imagenet dataset and a 18-layer deep R(two plus one)D model pretrained on the Kinetics-400 Dataset is used to extract the 2D and 3D visual features. A VGGlike model pretrained on the AudioSet Dataset is used to get a 128 dimension audio feature vector. Concatenating these features, we get the input feature for our model. F-Socres metrics used for evaluation and model trained with the Adam optimizer. Training happens in 3 stages:

- Train a basic **AVVP** model including Contrastive Learning
- Freeze the model and evaluate based on **Exchanging Channels**
- **Re-train** the model from scratch using modality aware labels

Slide 13

We name the final model as “MA” (Modality Aware).We compare our model MA with weakly-supervised sound detection method TALNet, temporal action localization methods STPN, and state- of-the-art audio-visual event parsing methods are shown in the table. All the models, including ours, are trained using the LLP training dataset only.

The table shows the performances of our method MA and state-of-the-art methods on the LLP test set. It can be seen from the table that our method outperforms the SOTA methods by a large margin on all AVVP subtasks for both the segment-level and event-level metrics.

The most significant improvement is found for visual event parsing, which validates our motivation that previous methods are suffered from the ambiguous overall labels of invisible events.

Comparison with the SOTA methods demonstrates that our model is able to predict significantly better event categories with accurate temporal locations.

Slide 14

Table shows 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. By leveraging clues between the audio and visual tracks and assigning different labels for the two modalities, we find the model performance gets significantly improved. Specifically, for the visual event parsing, the model with the modality-aware refinement significantly improves the performance by 4.6 points at the segment-level prediction. It validates that ambiguous video-level labels harm model training since some events only appear in one modality.

Slide 15

We further uncover the effect of modality-aware refinement by looking into modalities. We conduct experiments including 1) only refining audio labels, 2) only refining visual labels, and 3) refining both modalities labels. The results are reported in the shown Table. We can find the most significant improvement is brought by refining event labels for visual parsing prediction. By refining visual parsing labels, we significantly improve the performance on segment-level visual parsing evaluation.

The reason is that the visual content could only be captured for specific camera views, whether the object of interest might usually be outside of the field-of-view. In contrast, the audio signals are collected by microphones, which are able to perceive all the event information of the scenes. Therefore, unmatched event labels are more common for visual modalities. We achieve further performance improvement by refining event labels for both modalities.

Slide 16

Analysis on different τ values used in contrastive learning. Smaller τ leads to sharper probability distribution and There is a very slight increase in performance when tau decreases. Segment-level audio-visual video parsing results are reported.