Perceptual Evaluation of Audio-Visual Synchrony Grounded in Viewers' Opinion Scores

Lucas Goncalves*[1], Prashant Mathur[2], Chandrashekhar Lavania[2], Metehan Cekic[2], Marcello Federico[2], and Kyu J. Han[2]

[1] The University of Texas at Dallas
[2] Amazon
pramathu@amazon.com

Abstract. Recent advancements in audio-visual generative modeling have been propelled by progress in deep learning and the availability of data-rich benchmarks. However, the growth is not attributed solely to models and benchmarks. Universally accepted evaluation metrics also play an important role in advancing the field. While there are many metrics available to evaluate audio and visual content separately, there is a lack of metrics that offer a quantitative and interpretable measure of audio-visual synchronization for videos 'in the wild'. To address this gap, we first created a large scale human annotated dataset (100+ hrs) representing nine types of synchronization errors in audio-visual content and how human perceive them. We then developed a PEAVS (Perceptual Evaluation of Audio-Visual Synchrony) score, a novel automatic metric with a 5-point scale that evaluates the quality of audio-visual synchronization. We validate PEAVS using a newly generated dataset, achieving a Pearson correlation of 0.79 at the set level and 0.54 at the clip level when compared to human labels. In our experiments, we observe a relative gain 50% over a natural extension of Fréchet based metrics for Audio-Visual synchrony, confirming PEAVS' efficacy in objectively modeling subjective perceptions of audio-visual synchronization for videos 'in the wild'.

1 Introduction

Audio-visual (AV) generative modeling has made rapid progress in recent years due to advancements in deep learning models [38] and the availability of large-scale datasets like VGGSound [9] and AudioSet [15]. However, evaluation of these models remains an open challenge. Existing automatic metrics often focus on specific aspects, such as image quality [19, 37] or audio fidelity [10, 23, 32], lacking a satisfactory measure for assessing audio-visual synchronization of 'in the wild' videos. In fact, lack of well-established metrics for AV synchrony issues can already be observed in literature. For instance, recent works in the realm of audio-visual generation, like MM-Diffusion [33], had to solely rely on single-modality metrics to evaluate their outputs. The work on Diff-Foley [26] had

* Work done during an internship at Amazon.

to train their own audio-visual synchrony classifier. More advanced evaluation metrics are needed to provide a holistic assessment of audio-visual coherence. The development of such metrics is essential to accurately judge model performance, identify failure cases, and drive further progress in this evolving field.

Recent studies such as AV Synchrony Transformer (AVST) [8], SparseSync [21] and Diff-Foley [26] have defined the synchronization problem between audio-visual modalities primarily as an issue of temporal offset determination. However, this perspective addresses only one aspect of synchronization challenges. In our work, we extend the scope to a broader range of synchronization issues, encompassing audio/visual speed variations, intermittent muting, fragment shuffling, AV flickering, and temporal shifts. We also delve into developing a perceptual, automatic metric that aligns with human judgment. To achieve this, we initiated a large scale annotation study, gathering human assessments on audio-visual synchrony across over 100 hours of both original and distorted content, featuring the aforementioned synchronization issues.

In this endeavor, our objective is to create a metric that accomplishes the following: (a) attains a high correlation with human assessments of synchronization issues, (b) offers interpretable scoring for end users, and (c) operates in a reference-free manner, meaning it does not rely on ground truth for quality prediction. The key innovation in our approach lies in the development of an interpretable scale based on detailed perceptual guidelines provided to human raters. By optimizing for concordance with human-annotated scores during training, our metric produces scores directly aligned with predefined levels of audio-visual (a)synchrony. This emphasis on interpretability and alignment with human perception addresses the need for a more insightful evaluation of AV synchronization.
Our main contributions are as follows:

- We present a new Audio-Visual Synchrony human perception (AVS) benchmark data set with over 120K annotations and over 100+ hrs of content. Each video in this benchmark is annotated by three different annotators and we see an agreement of 0.71 Krippendorff's alpha [25].
- We propose a new audio-visual synchrony evaluation metric - PEAVS - that is reference-free and has interpretable scoring on a scale of 1 to 5.
- PEAVS shows high correlation of 0.79 with human judgements on the benchmark data, significantly outperforming a Fréchet distance based AV synchrony metric.

2 Related Works

Unimodal Metrics: Several factors have fueled progress in single modality generation, including advancement in deep learning model architectures and training/optimization methodology, increased computational capabilities, a rich availability of multimodal data, and the development of consistent metrics. Established metrics like Inception Score (ISc) [34], Fréchet Inception Distance (FID) [20], Fréchet Audio Distance (FAD) [23], Fréchet Video Distance (FVD) [37],

inter alia, are frequently employed to gauge the quality of the content generated by unimodal generative models, providing a standard for comparison. However, these metrics are reliant only on one modality - either audio [10,23,27,31,32,35, 39], video [5,19,24,37,42–44], or just applied to images [20,34]. In the evaluation of generated multimodal content, ensuring the quality of individual modalities is only half the battle. It is crucial to evaluate the synchronization between audio and visual elements, especially in the context of audio-visual generation.

Multimodal Metrics: For AV synchrony evaluation, it is important to determine the misalignment between audio and visual inputs. In the past, researchers have explored handling synchrony by statistical approaches and specifically designed features to pinpoint where the sound comes from, like locating the source of a voice in a scene [17]. However, with the rise of deep learning recent studies are harnessing distinct connections between audio and visuals, like those in human speech [11], audio-visual correspondence to check for matching content [3], or musical performances [4]. Some recent research studies, such as those by Chen *et al.* [8] and Iashin *et al.* [21], propose methods to determine audio-visual synchronization in a natural diverse video environment. However, their focus is solely on tackling the synchronization problem arising from delays in audio or video streams, specifically addressing differences in timing (offsets). In these existing approaches, the model is trained with the objective of either 1) detecting synchronization errors as a binary classification or 2) predicting the precise offset values through regression. In contrast, our aim is to have the model replicate the way humans perceive synchronization issues across a range of distortion types (see more in Sec 3.1).

Metrics for "in the wild" videos: The assessment of multimodal content has been extensively investigated within the computer vision community however most of the work has been tailored to lip-synchrony for instance, SyncNet [11] quantifies lip-sync errors by comparing mouth movements to the corresponding speech; AlignNet [40] focuses on aligning video content with ground truth audio, and predicts frame-wise offsets. In this work, we focus on evaluation of AV synchrony for "in the wild" videos which apart from Chen *et al.* [8], Iashin *et al.* [21] has been explored in Diff-Foley [26] where the authors train their own alignment classifier to detect synchrony. However, none of these works directly target modeling of human perception.

Perceptual Metrics: Previous studies on audio metrics [28,29] demonstrated that a learned metric trained on data collected via crowd-sourced human annotations, particularly on just noticeably different samples, correlates well with mean opinion scores. For video modality, Wang et. al. [41] designed a perceptual metric — learnt over a large scale human annotated data — designed to distinguish user generated video content in different dimensions such as semantic content, technical quality, and compression level. Our work in contrast centers on the synchronization between audio and video than solely on video or audio quality.

ID	Distortion	Parameter	Levels
1	Temporal Shift Audio	shift length (sec)	-1, -.5, -.125, .045, .1, .125, .25, .5, 1, 2
2	Audio Speed Change up	speed factor %	.025, .05, .10, .15, .20, .25, .30, .4, .5, .75
3	Video Speed Change up	speed factor %	.025, .05, .10, .15, .20, .25, .30, .4, .5, .75
4	Audio Speed Change down	speed factor %	.025, .05, .10, .15, .20, .25, .30, .4, .5, .75
5	Video Speed Change down	speed factor %	.025, .05, .10, .15, .20, .25, .30, .4, .5, .75
6	Intermittent Muting	duration (sec) of mute for every 1 sec	.01, .025, .05, .1, .2, .3, .5, 1, 2.5, 4
7	Randomly Sized Gaps Video	gap duration (sec) and prob. 40%	.01, .025, .05, .1, .2, .3, .5, 1, 2.5, 4
8	Fragment Shuffling	duration (sec) of segments to shuffle	.3, .4, .5, 1, 1.5, 2, 2.5, 3, 3.5, 4
9	AV Flickering	gap duration (sec) and prob. 40%	.01, .025, .05, .1, .2, .3, .5, 1, 2.5, 4

Table 1: Summary of Distortions and their Parameters for Audio-Visual Content.

3 AVS Benchmark Dataset

To develop an effective metric that aligns closely with human perception, it is important to have a dataset representative of human judgments. With this in mind, our study was designed to create a benchmark for our metric training and to assess its correlation with human perception, especially in the context of AV synchronization.

3.1 Data Preparation and Distortions

Audio-visual synchrony, the alignment of audio and visual elements, is important for an immersive multimedia experience. Even minor disruptions can notably degrade the user's experience, as witnessed in scenarios like watching a video where a mere mismatch in drums playing (visuals) and beats (audio) can become distracting. Various real-world challenges, from unreliable network connection to encoding mishaps, can lead to such anomalies. To produce realistic data representing such scenarios, we first collected 200 videos from the AudioSet corpus' *evaluation* split [15] ensuring that there is an aspect of synchrony in each of the samples. The samples selected are examples of "in the wild" videos, and does not include any samples with faces. We only include videos from actions such as: car driving by, dogs barking, instruments being played, manual labor being

done, etc.[3] Each video then underwent nine synchrony related distortions at ten varying levels as shown in Table 1, totaling in 18,200 distorted videos.
Following [37], we apply static and temporal noises in the AV content. **Temporal noise** is where the distortions in either or both modalities cause de-synchronization in the AV content:

- **Temporal Misalignment:** The audio track was offset, either forwards or backwards relative to the video [13], creating de-synchronizations spanning from -1 to 2 seconds.
- **Audio Speed Change:** Similar to FAD [23], the audio playback speed was manipulated, without corresponding changes in the video, resulting in gradual de-synchronization in the AV content. Depending on the speed change (up/down), we either clip the audio or visual frames to keep the same duration as the original video. In this setting, we ensure the sound characteristics is preserved by keeping essential features such as pitch and timbre while altering speed to minimize changes on sound texture.
- **Video Speed Change:** Similar to FVD [37], the speed of the video is adjusted while leaving the audio track untouched. This would cause the visuals to slowly move out of sync with the audio. As in the audio speed change, we clip the content to keep the same duration as the original one.
- **Fragment Shuffling:** Similar to the global swap distortion in FVD [37], both audio and video tracks were segmented and rearranged, preserving segment-level synchrony but disrupting overall synchronization of the video.

Static noise is where the distortions do not necessarily change the synchrony in the AV content but it can be perceived as synchronization errors by humans:

- **Intermittent Muting:** Intermittent periods of silences with varying duration is introduced in the audio track. This distortion disrupts the continuity, which can pose a challenge for some synchronization metrics. This distortion is similar to Pops as described in Kilgour et. al. [23].
- **Randomly Sized Gaps:** We randomly pick a frame with a certain probability and add black out visuals with varying number of frames in the video track. This distortion disrupts the continuity of the video. This distortion is similar to black rectangle noise as introduced in FVD.
- **AV Flickering:** Periods of silence and blackout were introduced in the audio-visual track. Although flickering by itself is not a synchronization issue, it can however be perceived as synchronization errors by humans and poses a challenging scenario for metrics [22].

3.2 Data Annotation

Our aim in this task is to **NOT** evaluate the standalone quality of either audio or video. Instead, our focus is on synchrony issues and how humans perceive

[3] Appendix B.4 contains a brief overview of top-30 label frequencies contained within the videos subset selected for our dataset.

them. Since pairwise stimulus experiments have shown to be faster than single-stimuli [12] and have been used in dubbing evaluation to foster relative measures in the rating [2], we also present our participants with two videos side-by-side. They are asked to play both videos, compare them, and rate each one on a scale from 1-5 based on the provided guidelines.[4]

From 18.2K videos, we randomly sampled (with replacement) 20K video pairs for comparison while taking into account the different levels of distortion (c.f. Table 1). Participants conducted pairwise comparisons on these videos and each pair was rated at least three times by different annotators. This resulted in a total of 60,000 pairwise ratings or 120,000 ratings for individual videos. Each video duration being at least 10 seconds, this amounts to 100+ hours of annotated videos with each video annotated at least 3 times. When looking at the inter-annotator agreement, we divided our data based on the severity of disagreement between the annotators. If all annotators annotated different scores those samples were put in "disagreement" group which constitute nearly 10% of the total annotated data. Within the disagreement group, we saw about 16% of cases where disagreements were equal or higher than 3, e.g., one video was rated as 1/4/5 by three annotators. These high-disagreement samples were then sent back to the annotators for QA, after which we received their revised annotations. We are unable to directly calculate inter-annotator agreement (like Cohen's or Fleiss Kappa), as we had a large pool of annotators (80+) and not all annotators annotated all samples. However, we do observe a Krippendorff's alpha value of 0.71 for the ratings assigned to all the samples, which corresponds to a moderate agreement [25].

After removing outliers and de-duplication we are left with 15.2K videos out of 40K.[5] We divide this 15.2K into 10.6K for training, 2.3K for development and 2.3K for evaluation. While creating these splits, we ensured that no original files and their derived (i.e., distorted) files are present in more than one set.

To gain insights into which distortion types posed the most perceptual challenges for humans in terms of synchronization, we examined videos that underwent direct comparisons and analyzed the variations in the annotated scores.[6] Discrepancies in these scores (absolute values) offered insights into how differently these distortion types were perceived. Figure 1 highlights that intermittent muting (type 6) is consistently the most noticeable distortion amongst all types (all high peaks). This observation aligns with the score distribution we observed in AVS benchmark (see Appendix Section B.5) and our manual analysis where audio muting was the most disruptive distortion and humans could easily differentiate audio muting from other distortions.

[4] Annotation guidelines are shown in Appendix B.2. Upon publication, we will release the data set and code for creating distorted samples.

[5] Since we randomly sampled with replacement, one video was annotated approx. 2.4 times by annotators, which led to duplication.

[6] We leave out audio-shift distortion in this analysis as the levels are not linearly increasing or decreasing.

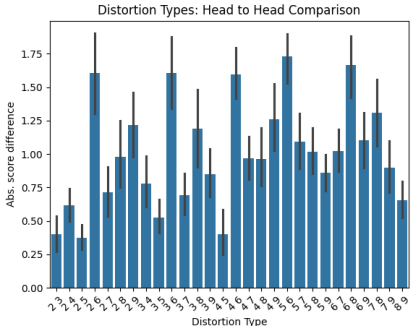


Fig. 1: We compare absolute differences in annotation score across the distortion types. In this plot, x-axis shows distortion types that were compared, i.e., '2 4' represents distortion type 2 v/s 4 and y-axis represents the difference in scores across annotation tasks. For ID to distortion type mapping see Table 1.

4 Preliminary Experiments

Studies by Unterthiner *et al.* [37] (Fréchet Video Distance) and Kilgour *et al.* [23] (Fréchet Audio Distance) have demonstrated the efficacy of embeddings from I3D [7] and VGGish [18] models in capturing qualitative and temporal characteristics from audio and video inputs, respectively. In the absence of dedicated metrics for assessing AV synchrony of 'in the wild' content and the lack of baseline models, we formulated a new Fréchet audio-visual distance (FAVD) metric.[7] Additionally, we want to assess if the combination of embeddings from I3D and VGGish models can be useful for detecting synchrony issues.

Fréchet Distance metrics—FID, FAD, FVD, and now FAVD—compare the statistics of generated samples to those of the ground truth or a large collection of samples. In our preliminary experiment, we aimed to explore how FAD, FVD, and FAVD metrics respond to temporal discrepancies in the input data. We conducted our study on 18.2K videos, generating two sets of audio-visual embeddings: one from disrupted content and another representing aligned content (ground truth). For each type and level of distortion, we created a distinct "evaluation set" to analyze the impact of each specific distortion type and level on these metrics. In our experiments, we deliberately chose not to use FID (Fréchet Inception Distance) due to its limitation in evaluating temporal aspects. FID is designed for static images and cannot adequately assess content with time-dependent elements, such as videos, where scene continuity and flow are essential for quality evaluation. By focusing on metrics capable of capturing these temporal dynamics, we aim to provide a more accurate and relevant evaluation of the audio-visual content's quality.

[7] SparseSync [21] and AVST [8] are only able to detect audio shift but our benchmark contain many more synchrony issues which render these metrics unusable.

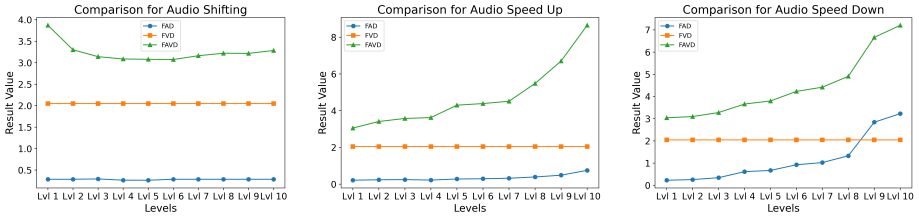
We use I3D and VGGish to extract features from audio-visual content within the same time span (of 0.96 seconds - a default time span from VGGish).

These representations are concatenated time-wise and we calculate multivariate Gaussians for both evaluation set embeddings, $N_e(\mu_e, \Sigma_e)$, and the ground truth AV collection embeddings, $N_{av}(\mu_{av}, \Sigma_{av})$.

We then compute the Fréchet distance between these two Gaussians, as described by Dowson *et al.* [14]:

$$F(N_{av}, N_e) = |\mu_{av} - \mu_e|^2 + tr(\Sigma_{av} + \Sigma_e - 2(\Sigma_{av}\Sigma_e)^{\frac{1}{2}}) \quad (1)$$

where tr is the trace of the matrix. For FAD and FVD, we compute the embeddings from respective modalities.



(a) Audio Shift **(b)** Audio Speed Up **(c)** Audio Speed Down

Fig. 2: In these plots, we show the effect of distortions with varying levels on **FAD**, **FVD**, and **FAVD** for three distortion types. Flat trend line implies that a metric is not able to capture the distortion type. Increasing or decreasing trends show that a metric is susceptible to varying levels in a distortion. Distortion levels are taken from Table 1. Detailed plots for all distortion types are available in Appendix C.

Examining the plots in Figure 2, it becomes evident that only FAVD (depicted in green) is responsive to various temporal distortions associated with AV synchrony across different levels. This is mainly due to use of covariance matrices (Σ_{av}, Σ_e) from audio-visual features (shown in Eq 1) which helps FAVD track the interaction of AV modality. Additionally, in cases like audio shift (a) where neither FAD nor FVD can detect synchronization issues (indicated by flat trend lines), and audio speed-up where FAD shows minimal movement, FAVD formulation exhibits a more pronounced reaction, highlighting that the fusion of features from I3D and VGGish models effectively captures temporality in an audio-visual context. Based on these observations, we chose to use I3D and VGGish features as inputs for our metric training, with FAVD serving as a baseline metric for the remainder of our work.

5 PEAVS Score Formulation

This section describes the architecture we have used to train our 'PEAVS' score metric and its training details.

5.1 Architecture

The input to our proposed model are the video and audio streams, as seen in Figure 3. For feature extraction, we leverage I3D and VGGish. In addition to this, our design draws inspiration from the audio-visual framework presented in [16]. Central to our architecture is the utilization of the cross-modal transformer formulation [36], which excels in capturing temporal information and relationships between different modalities. To ensure temporal consistency across modalities, both audio and visual input features are extracted over matched time spans and are then relayed through cross-modal transformer layers.

There are several key components in the our architecture. First, the audio and visual inputs are processed through our frozen backbone feature extractors. Both features are extracted over 0.96 second windows. The I3D extracted inputs x_v have dimensionality $x_v \in \mathbb{R}^{N_{av} \times 1024}$, where N_{av} is the dimension of the audiovisual feature sequence, and 1,024 is the feature vector dimension. The acoustic feature vector x_a extracted from the VGGish model is $x_a \in \mathbb{R}^{N_{av} \times 128}$, where N_{av} is the dimension of the acoustic feature sequence, and 128 is the feature vector dimension. We ensure that all inputs to the cross-modal layers have same dimensionality, by projecting the visual features from 1,024 to 128 through the use of a 1D convolutional layer to produce $\bar{x}_v \in \mathbb{R}^{N_v \times 128}$.

Next, we introduce positional embeddings to the feature vectors, following the step adopted in the original transformer paper [38]. The resultant features are used by the model to compute Query, Key, and Value vectors for each modality, as illustrated in Figure 3. In our model, the Query vector from one modality is paired in the transformer layer with the Key and Value vectors of the other, ensuring a cross-modal interaction. This design choice allows the cross-modal transformer encoder layers for each modality branch to produce representations for a given modality that are influenced by the other modality. Following this, the output of the cross-modal layers are refined through the self-attention layers to enhance the cross-modal representations. The resultant features are then processed by averaging the outputs from the self-attention layers followed by concatenation of these averaged representations from both modality branches. Finally, a *multilayer perceptron* (MLP) with three fully-connected layers is integrated to produce a prediction score based on these concatenated features.

5.2 Training

As seen in Figure 3, our training method consists of a two-stage strategy. In stage 1, we pre-train the model using a contrastive learning task, distinguishing between aligned and non-aligned pairs. Afterwards, in Stage Two, we fine-tune the model on our primary task, which involves predicting scores based on human annotators scores.

Stage 1 In stage 1, we pre-train our model using a contrastive learning objective. In this phase, we used AudioSet's *Balanced train* split [15], which after downloading all videos possible and removing corrupted ones resulted in 17K

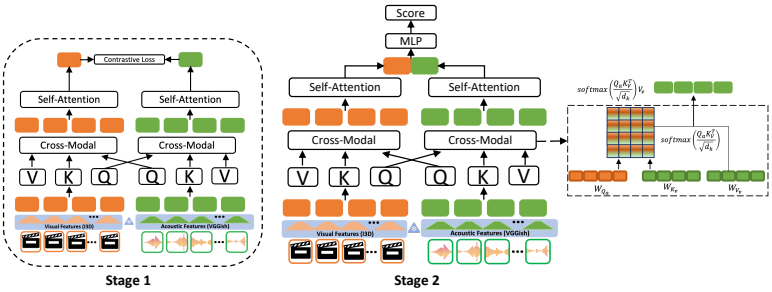


Fig. 3: Framework Overview

videos for our contrastive training stage.[8] We then subjected each of these files to ten levels of temporal misalignment (audio shift), as specified in section 3.1. This pre-processing resulted in a total of ∼190K files for the contrastive learning phase. Similar to SyncNet [11], the training objective ensures that the outputs of both the audio and video branches are similar for aligned pairs and dissimilar for misaligned pairs. Specifically, we either minimize or maximize the Euclidean distance between the network outputs for these cases, respectively. Our objective consists of a contrastive loss (Equation 2). The primary intention behind this pre-training is to embed regularization into the audio and visual features extracted from our frozen backbone networks. This ensures that the model prioritizes temporal attributes of the features over their qualitative characteristics.

$$\mathcal{L}_c = \frac{1}{2N}\sum_{n=1}^{N}(y_n)D_n^2 + (1-y_n)(\max(0, m - D_n))^2 \tag{2}$$

where $y \in \{0, 1\}$ is the label (1 for similar pairs and 0 for dissimilar pairs), D_w is the Euclidean distance between the two data points in the pair after passing them through the network, and m is the margin.

Stage 2 In stage 2 the output modality heads of the pre-trained model are concatenated together to produce audio-visual representations. Three fully-connected layers are added on top of these concatenated heads. We then fine tune the entire model for the downstream task of predicting human-alignment scores. Since the aggregated human annotation scores are non-negative real numbers, we approach model training as a regression task and utilize *concordance correlation coefficient* (CCC) to measure the agreement between the true and predicted scores. The CCC measurement is illustrated in Equation 3.

$$\mathcal{L}_{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{3}$$

where μ_x and μ_y denote the means of the true and predicted scores of a batch, respectively; σ_x and σ_y represent the standard deviations of the true and predicted

[8] Note that this is different from AVS benchmark.

scores of the batch, respectively; and ρ is their Pearson's correlation coefficient. Our training objective is to maximize CCC, ensuring that the predicted scores closely correlate with the true scores.

Training Details In stage 1, we divided the video files obtained from AudioSet's *Balanced train* split [15] into an 80% training set and a 20% validation set. This 80/20 split is carried out on the 17K original videos and extrapolated to distorted samples. We pre-trained the model in stage 1 until the validation loss plateaued. In stage 2, we use the train/validation/testing sets provided in the AVS benchmark corpus. We fine-tuned the model on training data for 20 epochs, saving the best-performing model based on the validation set performance.[9]

6 Experiments

In this section, we evaluate the performance of our trained model using human-evaluation scores. We compare the model's scores with those provided by human evaluators on the **test split** of AVS benchmark. To further assess the effectiveness of our cross-modal transformer architecture for metric training, we run two ablation experiments. First, to assess the impact of pre-training in stage 1, we directly train our model from scratch as in stage 2. This variant is termed "Cross-Modal Base" as it incorporates the same audiovisual cross-modal framework but excludes the contrastive learning pre-training phase. Secondly, to assess the importance of cross-modal framework, we test a version without cross-modal layers called "Basic Transformer". This version employs only the self-attention branches, differently from the cross-modal approach of sharing query vector from one modality to another.

6.1 Human Scores Correlation

For the above models, we compute the correlation of their outputs with the human judgements using pearson correlation coefficient (PCC). Since Fréchet based metrics require a set of videos to calculate distance between multivariate Gaussians we compute set level correlations. As PEAVS and its variants produce a score for each video separately, we average the scores across the test set to compute set level scores.[10] We also provide correlations at a much fine-grained clip level. Examining the results presented in Table 2, FAVD shows a relative improvement of 11% (from 0.475 to 0.527) and 64% (from 0.321 to 0.527) over FAD and FVD baselines setting a strong baseline for future work. This highlights FAVD's effectiveness in capturing AV synchrony issues. PEAVS model achieves a Pearson correlation of 0.794 at the Set Level, marking a substantial 51% improvement over the robust FAVD baseline. Notably, when we analyze the

[9] For further details on the pre-training and fine-tuning settings, refer to appendix A.
[10] Each set is defined as a collection of videos showing same distortion type and level. In total we had 90 sets, i.e. 9 distortions x 10 levels.

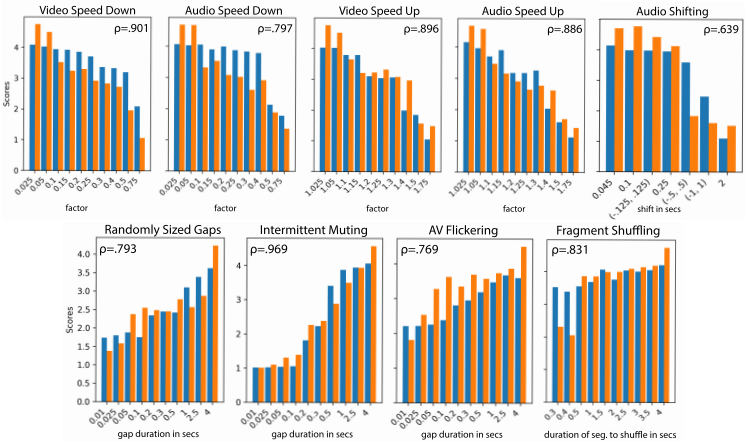


Fig. 4: PEAVS results with Human Evaluation Scores compared side-by-side for each distortion type and at different levels. PEAVS scores are represented by **blue** bars and Human Evaluation scores are represented by **orange** bars. The x-axis are marked by the parameter and its levels as shown in Table 1. ρ is the correlation of PEAVS with human judgements.

impact of pre-training on PEAVS (c.f. Cross-Modal Base), performance drops by 7% & 11% at set and clip level, respectively, underscoring the significance of pre-training in PEAVS. Furthermore, removing the cross-modal layers (c.f. Base Transformer) results in a performance decrease of 9.4% & 19% at set and clip level, respectively, emphasizing the critical role of cross-modality layers.

Model Type	Set Level	Clip Level
FVD [37]	0.321	-
FAD [23]	0.475	-
FAVD (Ours)	0.527	-
Basic Transformer	0.720	0.434
Cross-Modal Base	0.736	0.477
PEAVS (Ours)	**0.794**	**0.536**

Table 2: Correlation scores of different model types at Set and Clip levels.

6.2 Per Distortion Analysis

In this section, we analyze the response of PEAVS to varying distortions at different levels. Figure 4 factorizes the results in Table 2, by each distortion type and level. PEAVS scores are shown side-by-side with the scores from human annotations on AVS benchmark test set. Generally, PEAVS metric aligns well with scores from human judgements across the board except in a few cases.

In top-right corner of Figure 4, we present the effects of audio shift. For ease of analysis, we display these results in terms of absolute shift values. We start with a shift of 0.045 based on a report from ITU which indicated that the threshold for audio-visual shift detection by human ranges from -125ms to +45ms. [13]. At this level, both the metric's output and human scores peak (>4), and they subsequently decrease as the shift increases.

In contrast to audio-shift, PEAVS shows the highest correlation for intermittent muting distortion which implies it is easier for the metric to capture this distortion. This confirms the notion that we observed in Section 3.2 where intermittent muting stood out as the most noticeable distortion and it was easier for humans to detect it as well.

In case of the fragment shuffling (bottom-right corner), initial levels of perturbation — shuffled fragments of 0.3-0.4 seconds — is significantly more disruptive for humans than the metric, i.e. the metric fails to capture this insight.

		Predicted	
		Pos.	Neg.
Actual	Pos	39	227
	Neg	4	130

		Predicted	
		Pos.	Neg.
Actual	Pos	92	174
	Neg	24	110

Table 3: Confusion matrix for SparseSync (left) and PEAVS (right). Accuracy of SparseSync = 42.3% v/s PEAVS = 50.5%

6.3 PEAVS vs. SparseSync

In this section, we report an experiment comparing PEAVS with SparseSync, which was introduced by Iashin *et al.* [21] for detecting audio-visual shifts inside 'in the wild' videos. While SparseSync's goal is to objectively quantify specific synchronization issues, by estimating an audio-visual offset shift value, PEAVS evaluates synchronization under multiple facets and according to a perceptual scale. Due to their diverging objectives, a direct comparison is challenging. Moreover, SparseSync is trained on 21 classes representing distinct audio-shift levels from -2.0 sec (left) to 2.0 sec (right) with increments of 0.2, with 0.0 representing no audio shift; while PEAVS is trained to predict float values between 1 and 5.

In this experiment, we assess both model's performance in accurately classifying 400 videos, including 200 ground truth videos and 200 randomly audio-shifted versions of them. [11] We consider distortions of 0.045, 0.1 and ±0.125 as positive cases (i.e. ground truth) as ITU recommend an acceptability threshold for AV shift up to 185 ms [6]. This also aligns well with SparseSync buckets where first distortion starts at ±0.2 seconds (i.e. SparseSync's step-size). To ensure a fair assessment, we divided the PEAVS scale into 21 bins, matching the output structure of SparseSync.[12] For ground truth videos, we anticipate that

[11] More details about this evaluation set are in Appendix Section D.

[12] We use increments of 0.238 ($=\frac{5}{21}$).

SparseSync will classify them as having zero shifts (0.0 class), while PEAVS is expected to assign them to the highest scoring bin (i.e. $(4.76, 5]$).

In Table 3, we show the confusion matrices for SparseSync (left) and PEAVS (right). In terms of accuracy, PEAVS outperforms SparseSync by 19%, i.e. 50.5% vs. 42.3%. In particular, SparseSync shows an 18% edge in detecting distortions while PEAVS shows a 136% higher rate in detecting clean content.

7 Limitations

Like any research, our work is not without its limitations, and we have identified a couple of them here:
Dataset: Due to budget constraints (30K USD), we could only source 200 videos and conduct 60K pair-wise annotations. In selecting these 200 videos, we ensured diversity. Our dataset comprises 215 labels out of the 512 present in AudioSet, representing only 40% of labels. Expanding this dataset for broader video coverage is left for future work.
Metric: The limited size of the AVS benchmark highlights the necessity of pre-training in model-based metrics (in our case, PEAVS). This is further evidenced by the results in Table 2, where pre-training improves correlation from 0.736 (Cross-Modal Base) to 0.794 (PEAVS). In this study, we only pre-trained PEAVS on audio-shift noise, leaving pre-training on other noise types for future work.
"in the wild" videos: Due to proprietary issues, we could not experiment with datasets featuring talking faces like VoxCeleb [30] or Lip Reading Sentences [1]. Our study is confined to addressing synchrony issues for "in the wild" videos.

8 Conclusions

We introduced PEAVS, a novel metric designed for measuring synchrony in audio-visual content, an important advancement in the field of evaluation of audio-visual content. While the domain of audio-visual generative modeling is ever-evolving, assessing synchronization remains crucial for enhancing user experience. Although numerous metrics cater to either audio or visual content separately, a noticeable gap exists in evaluating their synchronization with a focus on viewers opinions. Our study addresses this gap with PEAVS, a metric based on a vast human-annotated dataset. The strong correlation between PEAVS scores and human evaluations confirms its efficacy in capturing perceptions surrounding audio-visual synchronization. As the scope of audio-visual content generation expands, tools like PEAVS are essential in maintaining a balanced interplay between sight and sound, guiding both creators and researchers.

Furthermore, we introduced the Audio-Visual Synchrony human perception benchmark, which provides a new dataset detailing human-perceptual scores on various audio-visual synchrony challenges for real-world videos. By isolating specific challenges in a controlled audio-visual environment, we aim to simplify the task for researchers to propose, assess, and analyze potential solutions while maintaining agreement with human perception.

References

1. Afouras, T., Chung, J.S., Zisserman, A.: Lrs3-ted: a large-scale dataset for visual speech recognition (2018)
2. Agarwal, M., Agrawal, S., Anastasopoulos, A., Bentivogli, L., Bojar, O., Borg, C., Carpuat, M., Cattoni, R., Cettolo, M., Chen, M., Chen, W., Choukri, K., Chronopoulou, A., Currey, A., Declerck, T., Dong, Q., Duh, K., Estève, Y., Federico, M., Gahbiche, S., Haddow, B., Hsu, B., Mon Htut, P., Inaguma, H., Javorský, D., Judge, J., Kano, Y., Ko, T., Kumar, R., Li, P., Ma, X., Mathur, P., Matusov, E., McNamee, P., P. McCrae, J., Murray, K., Nadejde, M., Nakamura, S., Negri, M., Nguyen, H., Niehues, J., Niu, X., Kr. Ojha, A., E. Ortega, J., Pal, P., Pino, J., van der Plas, L., Polák, P., Rippeth, E., Salesky, E., Shi, J., Sperber, M., Stüker, S., Sudoh, K., Tang, Y., Thompson, B., Tran, K., Turchi, M., Waibel, A., Wang, M., Watanabe, S., Zevallos, R.: FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In: Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023). pp. 1–61. Association for Computational Linguistics, Toronto, Canada (in-person and online) (Jul 2023). https://doi.org/10.18653/v1/2023.iwslt-1.1, https://aclanthology.org/2023.iwslt-1.1
3. Arandjelović, R., Zisserman, A.: Look, listen and learn. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 609–617 (2017), https://api.semanticscholar.org/CorpusID:10769575
4. Arandjelović, R., Zisserman, A.: Objects that sound. In: Computer Vision – ECCV 2018. pp. 451–466. Springer International Publishing, Cham (2018)
5. Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD gans. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=r1lUOzWCW
6. BT.1359, R.I.R.: BT.1359 - Relative timing of sound and vision for broadcasting (1998)
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733. IEEE Computer Society, Los Alamitos, CA, USA (jul 2017). https://doi.org/10.1109/CVPR.2017.502, https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.502
8. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Audio-visual synchronisation in the wild (2021)
9. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset (2020)
10. Chinen, M., Lim, F.S.C., Skoglund, J., Gureev, N., O'Gorman, F., Hines, A.: Visqol v3: An open source production ready objective speech and audio metric. In: 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–6 (2020). https://doi.org/10.1109/QoMEX48832.2020.9123150
11. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Workshop on Multi-view Lip-reading, ACCV (2016)
12. Clark, A.P., Howard, K.L., Woods, A.T., Penton-Voak, I.S., Neumann, C.: Why rate when you could compare? Using the "EloChoice" package to assess pairwise comparisons of perceived physical strength. PLOS ONE **13**(1), e0190393 (Jan 2018). https://doi.org/10.1371/journal.pone.0190393, https://dx.plos.org/10.1371/journal.pone.0190393

13. Committee, A.T.S.: Atsc implementation subcommittee finding: Relative timing of sound and vision for broadcast operations. Tech. rep., Advanced Television Systems Committee, 1750 K Street, N.W., Suite 1200, Washington, D.C. 20006 (6 2003), doc. IS-191
14. Dowson, D., Landau, B.: The fréchet distance between multivariate normal distributions. Journal of Multivariate Analysis **12**(3), 450–455 (1982). https://doi.org/https://doi.org/10.1016/0047-259X(82)90077-X, https://www.sciencedirect.com/science/article/pii/0047259X8290077X
15. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780 (2017). https://doi.org/10.1109/ICASSP.2017.7952261
16. Goncalves, L., Busso, C.: Auxformer: Robust approach to audiovisual emotion recognition. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7357–7361 (2022). https://doi.org/10.1109/ICASSP43922.2022.9747157
17. Hershey, J., Movellan, J.: Audio vision: Using audio-visual synchrony to locate sounds. In: Solla, S., Leen, T., Müller, K. (eds.) Advances in Neural Information Processing Systems. vol. 12. MIT Press (1999), https://proceedings.neurips.cc/paper_files/paper/1999/file/b618c3210e934362ac261db280128c22-Paper.pdf
18. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K.: Cnn architectures for large-scale audio classification (2017)
19. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: a reference-free evaluation metric for image captioning. In: EMNLP (2021)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Neural Information Processing Systems (2017)
21. Iashin, V., Xie, W., Rahtu, E., Zisserman, A.: Sparse in space and time: Audio-visual synchronisation with trainable selectors. In: British Machine Vision Conference (BMVC) (2022)
22. ITU-T RECOMMENDATION, P.: Subjective video quality assessment methods for multimedia applications (1999)
23. Kilgour, K., Zuluaga, M., Roblek, D., Sharifi, M.: Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In: Interspeech (2019), https://api.semanticscholar.org/CorpusID:202725406
24. Korhonen, J., You, J.: Korhonen. In: 2012 Fourth International Workshop on Quality of Multimedia Experience. pp. 37–38 (2012). https://doi.org/10.1109/QoMEX.2012.6263880
25. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology (second edition). Sage Publications (2004)
26. Luo, S., Yan, C., Hu, C., Zhao, H.: Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models (2023)
27. Luo, Y., Mesgarani, N.: Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 696–700 (2018). https://doi.org/10.1109/ICASSP.2018.8462116

28. Manocha, P., Finkelstein, A., Zhang, R., Bryan, N.J., Mysore, G.J., Jin, Z.: A Differentiable Perceptual Audio Metric Learned from Just Noticeable Differences. In: Interspeech 2020. pp. 2852–2856. ISCA (Oct 2020). https://doi.org/10.21437/Interspeech.2020-1191, https://www.isca-speech.org/archive/interspeech_2020/manocha20_interspeech.html
29. Manocha, P., Kumar, A., Xu, B., Menon, A., Gebru, I.D., Ithapu, V.K., Calamia, P.: Dplm: A deep perceptual spatial-audio localization metric. In: 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). pp. 6–10 (2021). https://doi.org/10.1109/WASPAA52581.2021.9632781
30. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. In: INTERSPEECH (2017)
31. Rix, A., Beerends, J., Hollier, M., Hekstra, A.: Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221). vol. 2, pp. 749–752 vol.2 (2001). https://doi.org/10.1109/ICASSP.2001.941023
32. Roux, J.L., Wisdom, S., Erdogan, H., Hershey, J.R.: Sdr – half-baked or well done? In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 626–630 (2019). https://doi.org/10.1109/ICASSP.2019.8683855
33. Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N., Jin, Q., Guo, B.: Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10219–10228. IEEE Computer Society, Los Alamitos, CA, USA (jun 2023). https://doi.org/10.1109/CVPR52729.2023.00985, https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00985
34. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. p. 2234–2242. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (2016)
35. Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing **19**(7), 2125–2136 (2011). https://doi.org/10.1109/TASL.2011.2114881
36. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Florence, Italy (7 2019)
37. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges (2019)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
39. Vincent, E., Gribonval, R., Fevotte, C.: Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing **14**(4), 1462–1469 (2006). https://doi.org/10.1109/TSA.2005.858005
40. Wang, J., Fang, Z., Zhao, H.: Alignnet: A unifying approach to audio-visual alignment. In: WACV (2020)

41. Wang, Y., Ke, J., Talebi, H., Yim, J.G., Birkbeck, N., Adsumilli, B., Milanfar, P., Yang, F.: Rich features for perceptual quality assessment of ugc videos. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13430–13439 (2021). https://doi.org/10.1109/CVPR46437.2021.01323
42. Wang, Z., Simoncelli, E., Bovik, A.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402 Vol.2 (2003). https://doi.org/10.1109/ACSSC.2003.1292216
43. Wang, Z., Bovik, A.: A universal image quality index. IEEE Signal Processing Letters **9**(3), 81–84 (2002). https://doi.org/10.1109/97.995823
44. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068

A Model Training Settings

Table 4: Training Settings

Parameter	**Stage 1**	**Stage 2**
LR	0.001	0.0001
Optimizer	Adam	Adam
Loss	Contrastive	CCC
Batch Size	128	64
Epochs	60	20
LR Scheduler	.1 (Patience 10)	.1 (Patience 3)
Margin	1.0	-

Table 5: Transformer Layers Settings

Parameter	**Value**
Heads	8
Layers	3
Embed Dim	128
Attention Dropout	0.1
Relu Dropout	0.1
Embed. Dropout	0.25
Residual Block Dropout	0.1

B Annotation Process

This section includes complementary information from the main paper, detailing human annotation guidelines and an analysis of the AVS benchmark. The analysis specifically addresses: 1) the extent of the benchmark's generality, and 2) the distribution of annotated scores within the benchmark.

B.1 Human Annotation Guidelines - Background

In this annotation task, we want to evaluate the synchronization quality between audio and visual modalities in the video, and get an understanding of how the two videos presented to you compare to each other in terms of quality of synchronization. Please note that:

– We are NOT looking at the perceived quality of audio or video independently, but only how well they are aligned.

- We added certain types of distortions in these videos that disrupt the AV synchrony like speeding up the audio or speeding up the video, intermittently muting the audio, blacking out parts of video, flickering, etc.
- You might find it difficult to decide if a document should receive a 3 or a 4. Always focus on making sure that the relative scores of both videos reflect your relative judgement of their audio/video synchronization quality. For example, if you scored two videos 2 or 3 points apart, it should indicate that you have a clear preference for the higher scored video. Conversely, if the difference in their scores is 1 point, it should indicate that you have only a slight preference for a audio/video quality.
- Note that the scores might be a tie, and that is acceptable.

B.2 Annotation Criteria

In each comparison view, you will see two videos side by side. Play both videos and rate them on a scale of 1-5 in terms of disruption caused by these distortions based on the following likert scale:

1. Score 1 when there is complete misalignment between audio and video OR the video/audio is totally incomprehensible due to disruptions.
2. Score 2 when only a few parts of audio/video are in alignment OR large portion of video/audio is incomprehensible due to disruptions.
3. Score 3 when there is moderate mis-alignment between audio/video OR some portion of video/audio is comprehensible but there are visible disruptions.
4. Score 4 when there is almost perfect alignment with minor mis-alignments in some parts of video OR most of the video and audio is comprehensible with minor disruptions.
5. Score 5 when there is perfect alignment and audio/video are flawlessly in sync, AND/OR have no disruptions at all.

B.3 Annotation Reliability

To ensure the reliability of our annotation process, we implemented a three-stage approach. The first stage was a warm-up phase designed to train the annotators, familiarize them with the guidelines and provide examples (including corner cases) that might appear in the full annotation task. After receiving the annotations for the warm-up set and refining the guidelines, we sent them two additional batches of annotation tasks in succession.

In each batch of annotation tasks, we included original videos to evaluate the annotators' performance. Specifically, we added original videos to approximately 150 pairs for comparison. The annotators were not informed that they were being evaluated or that these specific pairings were included. This was done to identify and filter out annotations from inexperienced annotators. Ultimately, we decided not to remove any annotators' work.

After each batch of annotations were completed, there was a QA step where the vendor reviewed the annotations at their end. Once the annotations were sent

to us, we reviewed a sample of annotators' work. If their annotations did not align with the expected outcomes, we returned these samples for re-annotation.

Additionally, as described in section 3.2, we conducted an overall analysis of annotator agreement to identify further samples needing re-annotation. This systematic approach ensured that our annotations were reliable and consistent.

B.4 Dataset Labels Information

To generate our dataset, we selected 200 videos from AudioSet [15] ensuring that each sample contained well-aligned audio-visual content. The chosen samples are representative of videos "in the wild"; we excluded any samples that featured talking faces. Instead, our selection primarily contains videos depicting actions such as: cars driving by, dogs barking, instruments being played, and manual labor, among others. Based on the labels provided in AudioSet, our dataset comprises 215 labels. The number of labels exceeds the number of files because some files may be associated with multiple classes, resulting in co-occurrence of labels in many files within AudioSet. To provide an overview of the labels in our selected dataset, Figure 5 presents a histogram of the top 30 label frequencies from our AudioSet subset.

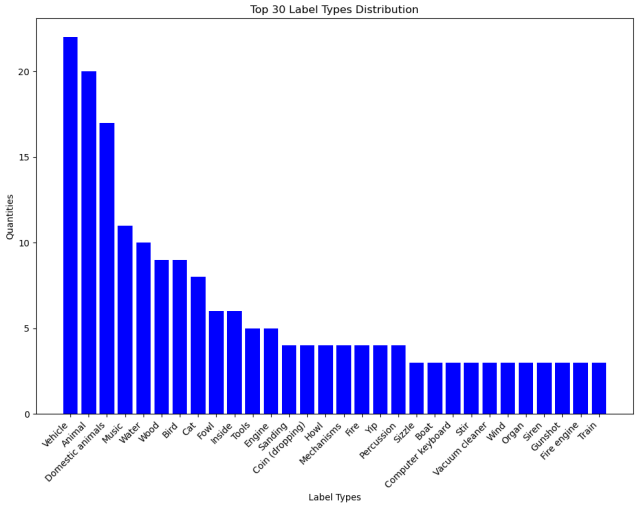


Fig. 5: Histogram of the top 30 label frequencies from the AVS benchmark sub-sampled from the AudioSet corpus.

B.5 Score Distribution, Filtering and Aggregation

In this section, we will analyse the score distribution on the AVS benchmark data and the effect of filtering. We wanted to do a followup analysis from the

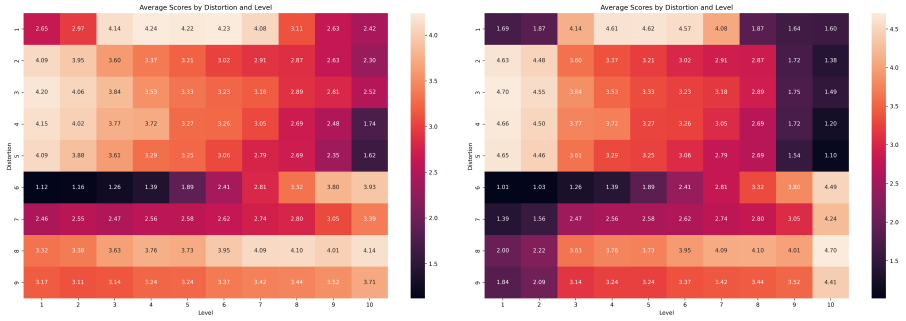


Fig. 6: Heatmap of score distribution on AVS benchmark, across distortion types and levels before (left) and after (right) filtering of data. The lighter the color the higher the scores and vice-versa.

main paper, especially by looking at the distribution of human annotation scores in the AVS benchmark data. Figure 6 provides more insights in our benchmark. Here, we observe a similar trend of intermittent muting (distortion type 6) being highly disruptive perceptually (colored with dark cells in the figure) compared to other distortion types especially in the lower levels of distortion - where the distorted videos are more jittery.

To remove the outliers in the benchmark, we filtered the dataset on the following criteria:

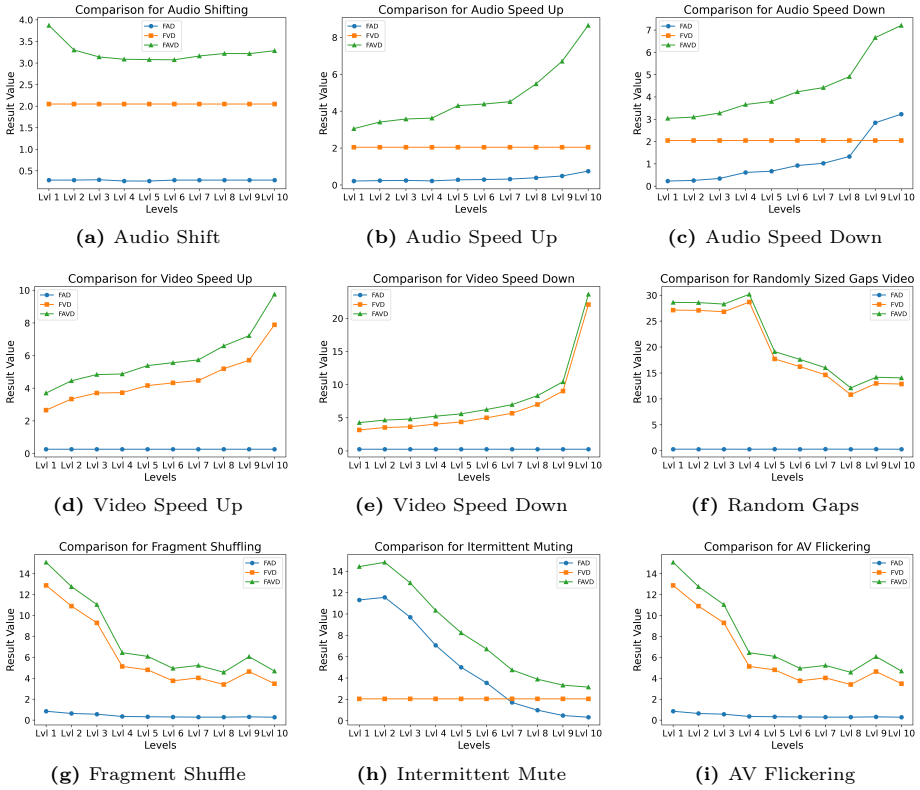
1. We aggregate the score of three annotators by averaging them and filter out all samples where ground truth videos are scored 3.5 or lower.
2. We filter out all samples where extreme distortion level samples are scored as 5.

Filtering the data increases the variance across different levels within each distortion type and a clear trend emerges. After filtering, less distorted videos have higher scores as compared to those without filtering. See distortion type IDs 7,8,9 (in Figure 6) where values in initial level shows an increasing trend in the right plot (filtered) as opposed to the left one (unfiltered).

Note that in Figure 6, distortion type IDs are the same as in Table 1. Levels in the audio-shift distortion (top row) is sorted in increasing order (from left shift of -1 to right shift of +2) this is why we see lower scores at the end of spectrums.

C FAD v/s FVD v/s FAVD

Plots depicted in Figure 7 provide more detailed comparisons of the effect of all distortion types with varying levels on **FAD**, **FVD**, and **FAVD**.



(a) Audio Shift **(b)** Audio Speed Up **(c)** Audio Speed Down

(d) Video Speed Up **(e)** Video Speed Down **(f)** Random Gaps

(g) Fragment Shuffle **(h)** Intermittent Mute **(i)** AV Flickering

Fig. 7: In these plots, we show the effect of all distortion types with varying levels on **FAD**, **FVD**, and **FAVD**. Flat trend line implies that a metric is not able to capture the distortion type. Increasing or decreasing trends show that a metric is susceptible to varying levels in a distortion. Distortion levels are taken from Table 1. These results are extension of Figure 2.

D PEAVS v/s SparseSync Benchmark

In our comparison for PEAVS v/s SparseSync, we created a held out evaluation set specifically for the head to head comparison. In addition to the 37 videos (annotated from AVS benchmark), we selected 163 more diverse set of videos from AudioSet [15], ensuring that they do not contain talking faces but showcase various real-world scenarios, such as a car driving by, an instrument being played, a dog barking, and many more. We manually looked into each video to ensure if they are well aligned. We then changed the offset of audio stream of these 200 videos by randomly sampling from the different levels of audio-shift distortion (as described in Table 1). This resulted in a total of 400 videos with an equal mix of ground truth and distorted videos. This portion of evaluation set will also be released as a part of AVS benchmark.

Levels of Distortion ITU recommends an acceptability thresholds for AV shift up to 185 ms [6], this is why we consider distortions of 0.045, 0.1 and ± 0.125 as positive cases (i.e. ground truth). This also aligns well with SparseSync buckets where the first distortion starts at ± 0.2 seconds (i.e. SparseSync's step-size).