

Text-to-feature diffusion for audio-visual few-shot learning

Otniel-Bogdan Mercea¹, Thomas Hummel¹, A. Sophia Koepke¹, and
Zeynep Akata^{1,2}

¹ University of Tübingen

² MPI for Intelligent Systems

{otniel-bogdan.mercea, thomas.hummel,
a-sophia.koepke,zeynep.akata}@uni-tuebingen.de

Abstract. Training deep learning models for video classification from audio-visual data commonly requires vast amounts of labelled training data collected via a costly process. A challenging and underexplored, yet much cheaper, setup is few-shot learning from video data. In particular, the inherently multi-modal nature of video data with sound and visual information has not been leveraged extensively for the few-shot video classification task. Therefore, we introduce a unified audio-visual few-shot video classification benchmark on three datasets, i.e. the VGGSound-FSL, UCF-FSL, ActivityNet-FSL datasets, where we adapt and compare ten methods. In addition, we propose AV-DIFF, a text-to-feature diffusion framework, which first fuses the temporal and audio-visual features via cross-modal attention and then generates multi-modal features for the novel classes. We show that AV-DIFF obtains state-of-the-art performance on our proposed benchmark for audio-visual (generalised) few-shot learning. Our benchmark paves the way for effective audio-visual classification when only limited labelled data is available. Code and data are available at <https://github.com/ExplainableML/AVDIFF-GFSL>.

Keywords: audio-visual learning, few-shot learning.

1 Introduction

The use of audio-visual data can yield impressive results for video classification [56, 62, 85]. The complementary knowledge contained in the two modalities results in a richer learning signal than using unimodal data. However, video classification frameworks commonly rely on significant amounts of costly training data and computational resources. To mitigate the need for large amounts of labelled data, we consider the few-shot learning (FSL) setting where a model is tasked to recognise new classes with only a few labelled examples. Moreover, the need for vast computational resources can be alleviated by operating on the feature level, using features extracted from pre-trained visual and sound classification networks.

In this work, we tackle the task of few-shot action recognition in videos from audio and visual data which is an understudied problem in computer vision. In

the few-shot setting, a model has to learn a transferable audio-visual representation which can be adapted to new classes with few annotated data samples. In particular, we focus on the more practical generalised FSL (GFSL) setting, where the aim is to recognise samples from both the base classes, i.e. classes with many training samples, and from novel classes which contain only few examples. Additional modalities, such as text and audio, are especially useful for learning transferable and robust representations from few samples.

To the best of our knowledge, the FSL setting with audio-visual data has only been considered for speech recognition [88], and for learning an acoustic model of 3D scenes [50]. Moreover, existing video FSL benchmarks are not suitable for the audio-visual setting. In particular, the SomethingV2 and HMDB51 benchmarks proposed in [15] and [87] do not contain audio and about 50% of the classes in the UCF101 benchmark from [83] have no sound either. The Kinetics split in [90] suffers from an overlap with the classes used to pre-train the feature extractors [83], and [56, 85] show that the audio modality in Kinetics is less class-relevant than the visual modality. Existing audio-visual zero-shot learning benchmarks [51, 52] cannot directly be used for few-shot learning due to their distinct training and testing protocols. Moreover, the baselines in both settings differ significantly as state-of-the-art few-shot learning methods usually necessitate knowledge of novel classes through classification objectives and generative models, a condition that is not possible in zero-shot learning. Thus, we introduce a new benchmark for generalised audio-visual FSL for video classification that is comprised of three audio-visual datasets and ten methods carefully adapted to this challenging, yet practical task.

To tackle our new benchmark, we propose AV-DIFF which uses a novel hybrid cross-modal attention for fusing audio-visual information. Different to various attention fusion techniques in the audio-visual domain [51, 52, 56] which use a single attention type or different transformers for each modality, our model makes use of a novel combination of within-modality and cross-modal attention in a multi-modal transformer. This allows the effective fusion of information from

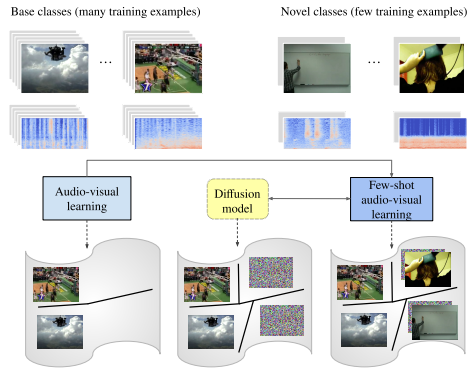


Fig. 1. AV-DIFF learns to fuse the audio-visual inputs into multi-modal representations in the audio-visual learning stage (left). In the few-shot learning stage (right), the multi-modal representations from the previous stage are used to concurrently train (double arrow line) a text-conditioned diffusion model on all the classes (middle) and a classifier. The classifier is trained on real features from base classes and real and synthetic features from novel classes.

both modalities and across the temporal dimension of the inputs. Furthermore, we introduce a novel text-conditioned diffusion model for generating audio-visual features to augment the few samples in the novel classes. In the image and video domain, generative adversarial networks (GANs) have been used to generate uni-modal features for data augmentation in the FSL setting [32, 46, 58, 83, 84]. However, we are not aware of prior works that have used diffusion models for multi-modal (audio-visual) feature generation in FSL. Both, cross-modal fusion and text-to-feature diffusion contribute to significant boosts in performance on our proposed benchmark.

To summarise, our contributions are: 1) We introduce the audio-visual generalised few-shot learning task for video classification and a benchmark on three audio-visual datasets. We additionally adapt and compare ten methods for this task. 2) We propose a hybrid attention mechanism to fuse multi-modal information and a diffusion model for multi-modal feature generation to augment the training dataset with additional novel-class samples. 3) We obtain state-of-the-art performance across all three datasets, outperforming the adapted multi-modal zero-shot learning and video FSL models.

2 Related work

We discuss prior works in learning from audio-visual data, FSL, and feature generation in low-shot learning.

Audio-visual learning. Multi-modal inputs, such as audio and visual data, provide significantly more information than unimodal data, resulting in improved overall performance for video classification and acoustic scene classification [7, 10, 45, 60–62]. Approaches, such as [21, 25], use class-label supervision between modalities without requiring temporal alignment between the input modalities. Besides audio and video classification, other domains also benefit from multi-modal data, such as lip reading [4, 5], audio synthesis based on visual information [27, 30, 43, 44, 57, 72, 89], and localisation and separation of sounds in videos [3, 6, 8, 18, 28, 59, 75]. Recently, transformer models have gained popularity in audio-visual learning, e.g. for classification [14], event localization [48], dense video captioning [36], and text-based video retrieval [26, 80]. As shown in these works, transformers can effectively process multi-modal input. Thus, our proposed framework fuses audio-visual information using a transformer-based mechanism.

FSL has been explored in the image domain [20, 23, 32, 47, 49, 64, 65, 68, 70, 73, 79, 81, 82, 86] and in the video domain [11, 15, 41, 83, 90]. The popular meta-learning paradigm in FSL [11, 15, 47, 49, 65, 73, 79, 81, 86, 90] has been criticised by recent works [20, 39, 81, 83]. In the video domain, commonly a query and support set is used and each query sample is compared to all the support samples [11, 15, 63, 90]. The number of comparisons grows exponentially with the number of ways and shots. These methods become prohibitively expensive for GFSL, where models are evaluated on both the base and the novel classes. Hence, we focus on the non-meta-learning approach in this work. Some non-meta-learning approaches have

addressed the more challenging and practical GFSL setting for videos [46, 83] using unimodal visual data. In contrast, we propose to use multi-modal data in our novel (G)FSL benchmark for audio-visual video classification which provides the possibility to test a model in both scenarios (FSL and GFSL).

Feature generation. Due to the progress of generative models, such as GANs [2, 29, 31, 37, 55] and diffusion models [12, 24, 67], different works have tried to adapt these systems to generate features as a data augmentation mechanism. GANs have been used in zero-shot learning (ZSL) and FSL [46, 58, 83, 84] to increase the number and diversity of samples, especially for unseen or novel classes. Diffusion models have also been applied to image generation in the feature space, such as [67, 77], but not in the ZSL or FSL setting. It is known that GANs are hard to optimize [69] while diffusion models appear to be more stable, leading to better results [22]. Therefore, our proposed framework uses a text-conditioned diffusion model to generate features for the novel classes in the FSL setting.

3 Audio-visual (G)FSL benchmark

We describe the audio-visual (G)FSL setting, present our proposed benchmark that we construct from audio-visual datasets, and explain the methods that we used to establish baselines for this task.

3.1 Audio-visual (G)FSL setting

We address the tasks of (G)FSL using audio-visual inputs. The aim of FSL is to recognise samples from classes that contain very few training samples, so-called *novel classes*. In addition, the goal of GFSL is to recognise both *base classes*, which contain a significant amount of samples, and novel classes.

Given an audio-visual dataset \mathcal{V} with M samples and C classes, containing base and novel classes, we have $\mathcal{V} = \{\mathcal{X}_{\mathbf{a}[i]}, \mathcal{X}_{\mathbf{v}[i]}, y_{[i]}\}_{i=1}^M$, where $\mathcal{X}_{\mathbf{a}[i]}$ represents the audio input, $\mathcal{X}_{\mathbf{v}[i]}$ the video input and $y_{[i]} \in \mathbb{R}^C$ the ground-truth class label. Both the audio and the video inputs contain temporal information. Two frozen, pre-trained networks are used to extract features from the inputs, VGGish [34] for the audio features $a_{[i]} = \{a_1, \dots, a_t, \dots, a_{F_a}\}_i$ and C3D [76] for video features $v_{[i]} = \{v_1, \dots, v_t, \dots, v_{F_v}\}_i$. We use these specific feature extractors to ensure that there is no leakage to the novel classes from classes seen when training the feature extractors (Sports1M [40] for the visual and YouTube-8M [1] for the audio modality), similar to [52]. A potential leakage is harmful as it would artificially increase the performance and will not reflect the true performance.

All models are evaluated in the FSL and GFSL settings for k samples in the novel classes (called shots), with $k \in \{1, 5, 10, 20\}$. During inference, in the FSL setting, the class search space is composed only of the novel class labels and the samples belonging to these classes. In the GFSL setting, the search space contains both the novel and base class labels and their corresponding samples.

Meta-learning approaches commonly use the notion of episodes, where each episode only uses P novel classes randomly sampled from the total number of

Table 1. Statistics for our VGGSound-FSL (1), UCF-FSL (2), and ActivityNet-FSL (3) benchmark datasets, showing the number of classes and videos in our proposed splits in the 5-shot setting. $\mathcal{V}_{B_1} \cup \mathcal{V}_{N_1}$ are used for training, Val_B and Val_N for validation in the first training stage. $\mathcal{V}_{B_2} \cup \mathcal{V}_{N_2}$ serves as the training set in the second stage, and evaluation is done on $Test_B$ and $Test_N$.

	# classes				# videos stage 1				# videos stage 2			
	all	\mathcal{V}_{B_1}	\mathcal{V}_{N_1}	\mathcal{V}_{N_2}	\mathcal{V}_{B_1}	\mathcal{V}_{N_1}	Val_B	Val_N	\mathcal{V}_{B_2}	\mathcal{V}_{N_2}	$Test_B$	$Test_N$
(1)	271	138	69	64	70351	345	7817	2757	81270	320	9032	2880
(2)	48	30	12	6	3174	60	353	1407	4994	30	555	815
(3)	198	99	51	48	9204	255	1023	4052	14534	240	1615	3812

novel classes in a dataset, usually $P \in \{1, 5\}$ (coined P -way). However, similar to [83], we suggest using higher values for P (e.g. all the classes in the dataset), so that the evaluation is closer to the real-world setting, as argued in [32, 83]. In our proposed FSL setting, P corresponds to the total number of novel classes $P = N$, while for GFSL $P = C$. Our evaluation protocol is in line with [32].

3.2 Dataset splits and training protocol

We provide training and evaluation protocols for audio-visual (G)FSL along with splits for UCF-FSL, ActivityNet-FSL and VGGSound-FSL. These are based on the UCF-101 [71], ActivityNet [33] and VGGSound [19] datasets.

Our proposed training and evaluation protocol is similar to [32, 51, 52]. The training protocol is composed of two stages, indicated by subscripts $1, 2$. In the first stage, a model is trained on the training set $Train_1 = \mathcal{V}_{B_1} \cup \mathcal{V}_{N_1}$ where \mathcal{V}_{B_1} consists of dataset samples from base classes, and \mathcal{V}_{N_1} contains k samples for each of the classes N_1 . The trained model is then evaluated on $Val = Val_B \cup Val_N$, where Val is the validation dataset which contains the same classes as $Train_1$. In the first stage, the hyperparameters for the network are determined, such as the number of training epochs and the learning rate scheduler parameters.

In the second stage, the model is retrained on the training set $Train_2$, using the hyperparameters determined in the first stage. Here, $Train_2 = \mathcal{V}_{B_2} \cup \mathcal{V}_{N_2}$ with $\mathcal{V}_{B_2} = Train_1 \cup Val$, and \mathcal{V}_{N_2} contains k samples for the novel classes in the $Test$ set. The final model is evaluated on $Test = Test_B \cup Test_N$ with $Train_2 \cap Test = \emptyset$. With a small number of shots, e.g. $k = 1$, models risk a bias towards the novel samples in $Train_2$. To obtain robust evaluation results, the second stage is repeated three times with k randomly selected, but fixed samples from \mathcal{V}_{N_2} . We provide dataset statistics in Table 1.

3.3 Benchmark comparisons

To establish benchmark performances for the audio-visual GFSL task, we adapt ten recent state-of-the-art methods for video FSL from visual information only, from audio-visual representation learning, and from audio-visual ZSL.

We provide results with several few-shot video recognition frameworks that are adapted to the multimodal audio-visual setting.

ProtoGan [46] uses GANs conditioned on the visual prototypes of classes that are obtained by averaging the features of all videos in that class. We adapt it to audio-visual inputs by concatenating the visual and audio features before passing them into the model.

SLDG [13] is a multi-modal video FSL that uses video frames and optical flow as input. It weighs the frame features according to normal distributions. We replace the optical flow in [13] with audio features.

TSL [83] is the current state-of-the-art video FSL which uses a GAN to generate synthetic samples for novel classes. It does not fully use temporal information, as the final score is the average of scores obtained on multiple short segments. We adapt it to the multi-modal setting by concatenating input features from the audio and visual modalities.

Moreover, we have adapted audio-visual representation learning methods to the few-shot task as can be seen below.

Perceiver [38], **Hierarchical Perceiver (HiP)** [16], and **Attention Fusion** [25] are versatile video classification methods and we provide comparisons with them. We use the implementations of the adapted Perceiver and Attention Fusion frameworks provided by [51] and we implement HiP in a similar way.

MBT [56] learns audio-visual representations for video recognition. It uses a transformer for each modality and these transformers can only exchange information using bottleneck attention.

Zorro [66], in contrast to MBT, uses two transformers that do not have access to the bottleneck attention. We adapt it by using a classifier on top of the averaged bottleneck attention tokens.

Finally, we have adapted the state-of-the-art methods in the audio-visual zero-shot learning domain, as shown below.

AVCA [52] is an audio-visual ZSL method which uses temporally averaged features for the audio and visual modalities. We adapt it by using a classifier on the video output, which is the strongest of the two outputs in [52].

TCAF [51] is the state-of-the-art audio-visual ZSL method. It utilizes a transformer architecture with only cross-modal attention, leveraging temporal information in both modalities. As it does not use a classifier, TCAF outputs embeddings, and we determine the class by computing the distance to the semantic descriptors and selecting the closest one.

4 AV-DIFF framework

In this section, we provide details for our proposed cross-modal AV-DIFF framework which employs cross-modal fusion (Section 4.1) and a diffusion model to generate audio-visual features (Section 4.2). Then, we describe the training curriculum in Section 4.3. Figure 2 illustrates AV-DIFF’s full architecture.

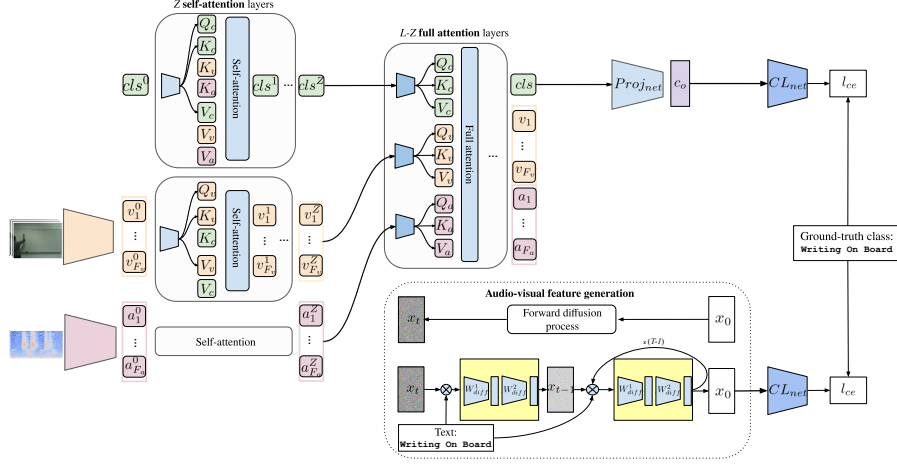


Fig. 2. Our AV-DIFF model for audio-visual (G)FSL takes audio and visual features extracted from pre-trained audio and video classification models as inputs. During training, the features from both modalities are fused into a classification token, denoted by cls . At the same time, our diffusion model (bottom) generates additional synthetic features for the novel classes (denoted by x_0). Finally, we train our classifier CL_{net} (right) on fused real features c_o of both novel and base classes and synthetic features of novel classes. \otimes is the concatenation operator.

4.1 Audio-visual fusion with cross-modal attention

Audio-visual fusion. We project the audio $a_{[i]}$ and visual features $v_{[i]}$ to a shared embedding space. Then we use Fourier features [74] as temporal positional embeddings and modality embeddings respectively and obtain positional aware video v_t^E and audio a_t^E tokens for timestep t . We prepend a classification token $cls^0 \in \mathbb{R}^{d_{dim}}$ to the audio and visual tokens. The output token cls corresponding to cls^0 is the final fused audio-visual representation which is input to $Proj_{net}$. Our audio-visual fusion mechanism contains L layers, which are based on multi-head attention [78] Att^l , followed by a feed forward function $FF^l : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{dim}}$. The input to the first layer is $x_{in}^1 = [cls^0, a_1^E, \dots, a_{T_a}^E, v_1^E, \dots, v_{T_v}^E]$. The output of a layer is:

$$x_{out}^l = FF^l(Att^l(x_{in}^l) + x_{in}^l) + Att^l(x_{in}^l) + x_{in}^l. \quad (1)$$

In the following, we describe the first layer of the audio-visual fusion. The other layers work similarly. Our input x_{in}^1 is projected to queries, keys and values with linear maps $s : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{dim}}$ for $s \in \{q, k, v\}$. The outputs of the projection are written as zero-padded query, key and value features. For the keys we get:

$$\mathbf{K}_c = [k(cls^0), 0, \dots, 0], \quad (2)$$

$$\mathbf{K}_a = [0, \dots, 0, k(a_1^E), \dots, k(a_{F_a}^E), 0, \dots, 0], \quad (3)$$

$$\mathbf{K}_v = [0, \dots, 0, k(v_1^E), \dots, k(v_{F_v}^E)]. \quad (4)$$

The final keys are obtained as $\mathbf{K} = \mathbf{K}_c + \mathbf{K}_a + \mathbf{K}_v$. The queries and values are obtained in a similar way. We define full attention as $\mathbf{A} = \mathbf{A}_c + \mathbf{A}_{cross} + \mathbf{A}_{self}$:

$$\begin{aligned} \mathbf{A}_c &= \mathbf{Q}_c \mathbf{K}^T + \mathbf{K} \mathbf{Q}_c^T, & \mathbf{A}_{cross} &= \mathbf{Q}_a \mathbf{K}_v^T + \mathbf{Q}_v \mathbf{K}_a^T, \\ \mathbf{A}_{self} &= \mathbf{Q}_a \mathbf{K}_a^T + \mathbf{Q}_v \mathbf{K}_v^T. \end{aligned} \quad (5)$$

The novelty in the attention mechanism in AV-DIFF is that it exploits a hybrid attention mechanism composed of two types of attention: within-modality self-attention and full-attention. The first Z layers use self-attention $\mathbf{A}_{self} + \mathbf{A}_c$, the subsequent $L - Z$ layers leverage full attention \mathbf{A} .

Audio-visual classification. We project cls to $\mathbb{R}^{d_{out}}$ by using a projection network, $c_o = Proj_{net}(cls)$. Then, we apply a classification layer to c_o , $logits = CL_{net}(c_o)$. Given the ground-truth labels gt , we use a cross-entropy loss, $L_{ce} = CE(logits, gt)$ to train the full architecture.

4.2 Text-conditioned feature generation

AV-DIFF uses a diffusion process to generate audio-visual features which is based on the Denoising Diffusion Probabilistic Models (DDPM) [35]. In particular, we condition the generation of features for novel classes on a conditioning signal, such as the word embedding (e.g. word2vec [53]) of a class name. The diffusion framework consists of a forward process and a reverse process.

The forward process adds noise to the data sample x_0 for T timesteps:

$$q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1}) = \prod_{t=1}^T \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}), \quad (6)$$

where β_1, \dots, β_T is the variance schedule.

As the **reverse process** $q(x_{t-1}|x_t)$ is intractable, we approximate it with a parameterised model p_θ :

$$p_\theta(x_{0:T}) = p_\theta(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t) = p_\theta(x_T) \prod_{t=1}^T \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \quad (7)$$

We condition the model on the timestep t and the class label embedding w ,

$$L_{diff, w} = E_{x_0, t, w, \epsilon} [||\epsilon - \epsilon_\theta(\sqrt{a_t}x_0 + \sqrt{1 - a_t}\epsilon, w, t)||^2], \quad (8)$$

where ϵ is the noise added at each timestep and ϵ_θ is a model that predicts this noise. The sample at timestep $t - 1$ is obtained from timestep t as:

$$p_\theta(x_{t-1}|x_t, w) = \mathcal{N}(x_{t-1}; \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}}\epsilon_\theta(x_t, w, t)), \sigma_t^2 \mathbf{I}). \quad (9)$$

The input to ϵ_θ at timestep t is obtained by concatenating x_t, w , and t . We optimize $L_{\text{diff},w}$ to learn p_θ .

4.3 Training curriculum and evaluation

Each training stage (explained in Section 3.2) is split into two substages. In the first substage, we train the full architecture (the fusion mechanism, the diffusion model, $Proj_{net}$ and the classifier CL_{net}) on base classes \mathcal{V}_{B_1} (or \mathcal{V}_{B_2} in the second stage) by minimizing $L_{ce} + L_{\text{diff},w}$. The classifier CL_{net} is trained only on real features for the base classes in \mathcal{V}_{B_1} (or \mathcal{V}_{B_2} for the second stage) in the first substage.

During the second substage, we freeze the fusion mechanism and continue to train the diffusion model, $Proj_{net}$ and CL_{net} with the same training objective $L_{ce} + L_{\text{diff},w}$. Here we consider both base and novel classes \mathcal{V}_{B_1} and \mathcal{V}_{N_1} classes (or \mathcal{V}_{B_2} and \mathcal{V}_{N_2} in the second stage), unlike in the first substage where we only used base classes. For each batch composed of real samples from novel classes, we generate a corresponding batch of the same size with synthetic samples using our diffusion model. CL_{net} is then trained on real features from \mathcal{V}_{B_1} (or \mathcal{V}_{B_2} in the second stage) and on real and synthetic features for the classes in \mathcal{V}_{N_1} (or \mathcal{V}_{N_2} in the second stage). Freezing the audio-visual transformer ensures that its fusion mechanism does not overfit to the few samples from the novel classes.

The diffusion model is not used for inference, and the output of the classifier CL_{net} for c_0 provides the predicted score for each class (including the novel classes). The class with the highest score is selected as the predicted class.

5 Experiments

In this section, we first provide the implementation details for obtaining the presented results (Section 5.1). We then report results for our proposed AV-DIFF in our benchmark study (Section 5.2). Finally, we analyse the impact of different components of AV-DIFF (Section 5.3).

5.1 Implementation details

AV-DIFF uses features extracted from pre-trained audio and visual classification networks as inputs (details provided in the suppl. material). AV-DIFF is trained using $d_{dim} = 300$ and $d_{out} = 64$. Our fusion network has $L = 5, 4, 8$ transformer layers, the layer after which the attention changes is set to $Z = 3, 2, 5$ on ActivityNet-FSL, UCF-FSL and VGGSound-FSL respectively. We train all models on a single NVIDIA RTX 2080-Ti GPU. The first substage uses 30 epochs while the second one uses 20 epochs. We use the Adam optimizer [42], and $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of $1e^{-5}$. We use a learning rate of $7e^{-5}$ for UCF-FSL and ActivityNet-FSL, and $6e^{-5}$ for VGGSound-FSL. For ActivityNet-FSL and UCF-FSL, we use a scheduler that reduces the learning rate by a factor of 0.1 when the performance has not improved for 3 epochs. We use a batch size of

Table 2. Our benchmark study for audio-visual (G)FSL: 1,5,10-shot performance of our AV-DIFF and compared methods on (G)FSL. The harmonic mean (HM) of the mean class accuracies for base and novel classes are reported for GFSL. For the FSL performance, only the test subset of the novel classes is considered. Base, novel, and 20-shot performances are included in the suppl. material.

Model ↓	VGGSound-FSL						UCF-FSL						ActivityNet-FSL					
	1-shot		5-shot		10-shot		1-shot		5-shot		10-shot		1-shot		5-shot		10-shot	
	HM	FSL	HM	FSL	HM	FSL	HM	FSL	HM	FSL	HM	FSL	HM	FSL	HM	FSL	HM	FSL
Att. Fusion [25]	15.46	16.37	28.22	31.57	30.73	39.02	37.39	36.88	51.68	47.18	57.91	52.19	4.35	5.82	6.17	8.13	10.67	10.78
Perceiver [38]	17.97	18.51	29.92	33.58	33.65	40.73	44.12	33.73	48.60	40.47	55.33	47.86	17.34	12.53	25.75	21.50	29.88	26.46
MBT [56]	14.70	21.96	27.26	34.95	30.12	38.93	39.65	27.99	46.55	34.53	50.04	39.73	14.26	12.63	23.26	22.38	26.86	26.03
TCAF [51]	19.54	20.01	26.09	32.22	28.95	36.43	44.61	35.90	46.29	37.39	54.19	47.61	16.50	13.01	22.79	21.81	24.78	23.33
ProtoGan [46]	10.74	14.08	25.17	28.87	29.85	34.80	37.95	28.08	42.42	33.63	51.01	40.68	2.77	4.40	2.67	7.81	4.05	8.81
SLDG [13]	16.83	17.57	20.79	25.17	24.11	29.48	39.92	28.91	36.47	28.56	34.31	26.96	13.57	10.30	22.29	19.16	27.81	25.35
TSL [83]	18.73	22.44	19.49	29.50	21.93	31.29	44.51	35.17	51.08	42.42	60.93	55.63	9.53	10.77	10.97	12.77	10.39	12.18
HiP [16]	19.27	18.64	26.82	30.67	29.25	35.13	21.79	34.88	36.44	42.23	50.69	43.29	13.80	10.31	18.10	16.25	19.37	17.06
Zorro [66]	18.88	21.79	29.56	35.17	32.06	40.66	44.35	34.52	51.86	42.59	58.89	49.06	14.56	11.94	23.14	21.94	27.35	26.33
AVCA [52]	6.29	10.29	15.98	20.50	18.08	28.27	43.61	31.24	49.19	36.70	50.53	39.17	12.83	12.22	20.09	21.65	26.02	26.76
AV-DIFF	20.31	22.95	31.19	36.56	33.99	41.39	51.50	39.89	59.96	51.45	64.18	57.39	18.47	13.80	26.96	23.00	30.86	27.81

32 for ActivityNet-FSL, and 64 for UCF-FSL and VGGSound-FSL. Each epoch consists of 300 batches. As ActivityNet-FSL has very long videos, we randomly trim the number of features during training to 60. During evaluation, we also trim the videos to a maximum length of 300 features, and the trimmed features are centred in the middle of the video. To reduce the bias towards base classes, we use calibrated stacking [17] on the search space composed of the interval [0,1] with a step size of 0.1. This value is obtained on the validation dataset.

5.2 Audio-visual GFSL performance

For each of the models featured in our benchmark, we report results for three different numbers of shots, i.e. 1-shot, 5-shot, 10-shot on all three datasets in Table 2. AV-DIFF outperforms all the methods across all shots and datasets for few-shot learning (FSL) and generalised few-shot learning (HM).

For 1-shot, AV-DIFF achieves a HM/FSL of 20.31%/22.95% vs. HM of 19.54% for TCAF and FSL score of 22.44% for TSL on VGGSound-FSL. On 5-shot, our model obtains a HM/FSL of 31.19%/36.56% vs. 29.92% for the Perceiver and FSL of 35.17% for Zorro. Furthermore, AV-DIFF yields slightly better results than the Perceiver in both HM and FSL for 10 shots, with HM/FSL of 33.99%/41.39% vs. 33.65%/40.73% for the Perceiver. Thus, combining our hybrid attention and the diffusion model is superior to systems that rely solely on powerful attention mechanisms without incorporating generative modelling (Perceiver, TCAF) and systems that incorporate generative modelling, but that do not employ powerful attention mechanisms (TSL, ProtoGan).

Similar trends are observed on UCF-FSL, while on ActivityNet-FSL, the ranking of methods changes dramatically. Methods that perform well on UCF-FSL and VGGSound-FSL, but which do not fully use the temporal information (e.g. Attention Fusion, ProtoGan and TSL) perform weakly on ActivityNet-FSL which contains videos with varying lengths, including some very long videos,

Table 3. Impact of different audio-visual fusion mechanisms in the 5-shot setting.

Model ↓	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
A	28.56	31.52	29.98	36.55	78.95	42.07	54.90	43.75	23.10	22.06	22.57	22.53
$\mathbf{A}_{cross} + \mathbf{A}_c$	28.44	32.48	30.33	36.85	82.89	44.33	57.77	47.02	27.02	21.25	23.79	21.98
$\mathbf{A}_{self} + \mathbf{A}_c$	26.68	33.23	29.60	37.06	50.10	44.58	47.18	45.03	31.61	21.48	25.58	22.65
Alternate AV-DIFF	27.40	32.60	29.78	36.82	80.25	43.01	56.00	45.81	31.15	21.57	25.49	22.59
AV-DIFF	30.88	31.50	31.19	36.56	74.11	50.35	59.96	51.45	35.84	21.61	26.96	23.00

making the setting more challenging. Our AV-DIFF can process temporal information effectively, resulting in robust state-of-the-art results on ActivityNet-FSL.

Interestingly, VGGSound-FSL contains the most classes among the datasets considered, resulting in a significantly lower N (suppl. material, Tab. 1) than FSL. This also lowers the HM (computed from B, N). On VGGSound-FSL, methods tend to be biased towards novel classes ($N \geq B$) due to calibration [17]. In this case, $HM \leq N \leq FSL$. Moreover, some baselines that were also used in audio-visual zero-shot learning [51, 52] (e.g. TCAF) exhibit significant increases in performance even in the 1-shot setting. This is expected as for 1-shot learning, one training example is used from each novel class. This reduces the bias towards base classes, leading to more balanced B and N scores, and thereby better HM and FSL results. Base, novel, and 20-shot performances are included in the suppl. material.

5.3 AV-DIFF model ablations

Here, we analyse the benefits of the main components of AV-DIFF, i.e. our proposed audio-visual fusion mechanism, and the diffusion model for feature generation. Furthermore, we analyse the importance of using multiple modalities, and the effect of different semantic representations.

Audio-visual fusion mechanism. Table 3 ablates our cross-modal fusion mechanism for generating rich audio-visual representations. As shown in Section 4.1, AV-DIFF uses two types of attention: $\mathbf{A}_{self} + \mathbf{A}_c$ for the first few layers and **A** for the later layers. For *Alternate AV-DIFF*, we alternate the two types of attention used in AV-DIFF in subsequent layers. We also show our model with $\mathbf{A}_{cross} + \mathbf{A}_c$ which is the same attention used by the SOTA audio-visual GZSL framework [51]. On ActivityNet-FSL, AV-DIFF obtains a HM/FSL of 26.96%/23.00% vs. 25.58%/22.65% for $\mathbf{A}_{self} + \mathbf{A}_c$. The same trend is seen on UCF-FSL. On VGGSound-FSL, we outperform *Alternate AV-DIFF* on HM but are slightly weaker than $\mathbf{A}_{self} + \mathbf{A}_c$ in FSL. Overall, our fusion mechanism is the best across both metrics and datasets.

Feature generation model. In Table 4, we investigate the impact of different generative models to produce audio-visual features for the novel classes. We compare the diffusion model in AV-DIFF to a GAN similar to the one used by TSL [83], which optimizes a Wasserstein GAN loss [9]. On ActivityNet-FSL, we observe that AV-DIFF outperforms the GAN variant, with a HM/FSL of

Table 4. Influence of using different feature generators in the 5-shot setting.

Model ↓	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
AV-GAN	27.80	31.75	29.64	36.53	83.79	36.20	50.56	37.33	35.12	19.53	25.10	21.35
AV-DIFF	30.88	31.50	31.19	36.56	74.11	50.35	59.96	51.45	35.84	21.61	26.96	23.00

Table 5. Influence of using multi-modal input in the 5-shot setting.

Model ↓	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
Audio	28.30	30.56	29.39	36.64	55.31	39.18	45.87	44.44	13.74	15.23	14.45	17.58
Visual	7.83	8.92	8.35	9.51	67.13	30.70	42.14	30.98	20.80	17.49	19.01	17.84
AV-DIFF	30.88	31.50	31.19	36.56	74.11	50.35	59.96	51.45	35.84	21.61	26.96	23.00

26.96%/23.00% vs. 25.10%/21.35% for the GAN. The same can be seen on UCF-FSL and VGGSound-FSL. This shows that our generative diffusion model is better suited for audio-visual GFSL than a GAN.

Multi-modal input. We explore the impact of using multi-modal inputs for AV-DIFF in Table 5. For unimodal inputs, we adapt AV-DIFF to only employ full attention which is identical to self-attention in this case. On ActivityNet-FSL, using multi-modal inputs provides a significant boost in performance compared to unimodal inputs, with a HM/FSL of 26.96%/23.00% vs. 19.01%/17.84% when using only visual information. The same trend can be observed on UCF-FSL. In contrast, on VGGSound-FSL, using multi-modal inputs gives stronger GFSL but slightly weaker results in FSL than using the audio modality. This might be due to the focus on the audio modality in the data curation process for VGGSound. As a result, significant portions of the visual information can be unrelated to the labelled class. Overall, the use of multi-modal inputs from the audio and visual modalities significantly boosts the (G)FSL performance for AV-DIFF.

However, one interesting aspect is that using both modalities leads to better B and N performances across all three datasets. For example, on ActivityNet-FSL, AV-DIFF obtains a B score of 35.84% and an N score of 21.61% compared to 20.80% and 17.49% when using only the visual modality. On UCF-FSL, AV-DIFF achieves a score of 74.11% for B and 50.35% for N compared to 67.13% and 39.18% for the visual and audio modalities respectively. Finally, on VGGSound-FSL, AV-DIFF achieves a B score of 30.88% and an N score of 31.50% compared to 28.30% and 30.56% for unimodal audio inputs. This shows that using multi-modal inputs decreases the bias towards either of the metrics, leading to a more robust and balanced system.

Semantic class representations. We consider using different semantic class representations in Table 6. In FSL, the most common semantic descriptor is word2vec [53] which is used to condition the audio-visual feature generation in AV-DIFF. However, related works (e.g. ProtoGan [46]), use prototypes which average the visual features of all the training videos in a class to obtain the semantic representation of that class. In the multi-modal setting, we can concatenate the

Table 6. Influence of different semantic class representations in the 5-shot setting.

Model ↓	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
AV-DIFF av_{prot}	25.74	33.00	28.92	35.76	83.38	42.46	56.26	44.78	32.22	21.50	25.79	22.73
AV-DIFF	30.88	31.50	31.19	36.56	74.11	50.35	59.96	51.45	35.84	21.61	26.96	23.00

audio and visual prototypes to obtain multi-modal prototypes av_{prot} which is used as a conditioning signal for our diffusion model. On ActivityNet-FSL, using word2vec embeddings leads to better results than using the audio-visual prototypes av_{prot} , with a HM/FSL of 26.96%/23.00% vs. 25.79%/22.73% for av_{prot} . The same can be seen on UCF-FSL and VGGSound-FSL, demonstrating that the word2vec embeddings provide a more effective conditioning signal.

6 Conclusion

In this work, we propose an audio-visual (generalised) few-shot learning benchmark for video classification. Our benchmark includes training and evaluation protocols on three datasets, namely VGGSound-FSL, UCF-FSL and ActivityNet-FSL, and baseline performances for ten state-of-the-art methods adapted from different fields. Moreover, we propose AV-DIFF which fuses multi-modal information with a hybrid attention mechanism and uses a text-conditioned diffusion model to generate features for novel classes. AV-DIFF outperforms all related methods on the new benchmark. Finally, we provided extensive model ablations to show the benefits of our model’s components. We hope that our benchmark will enable significant progress for audio-visual generalised few-shot learning.

Acknowledgements: This work was supported by BMBF FKZ: 01IS18039A, DFG: SFB 1233 TP 17 - project number 276693517, by the ERC (853489 - DEXIM), and by EXC number 2064/1 – project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting O.-B. Mercea and T. Hummel.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv:1609.08675 (2016) 4, 18
2. Adler, J., Lunz, S.: Banach wasserstein gan. NeurIPS (2018) 4
3. Afouras, T., Asano, Y.M., Fagan, F., Vedaldi, A., Metze, F.: Self-supervised object detection from audio-visual correspondence. In: CVPR (2022) 3
4. Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. IEEE TPAMI (2018) 3
5. Afouras, T., Chung, J.S., Zisserman, A.: Asr is all you need: Cross-modal distillation for lip reading. In: ICASSP (2020) 3

6. Afouras, T., Owens, A., Chung, J.S., Zisserman, A.: Self-supervised learning of audio-visual objects from video. In: ECCV (2020) [3](#)
7. Alwassel, H., Mahajan, D., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. In: NeurIPS (2020) [3](#)
8. Arandjelovic, R., Zisserman, A.: Objects that sound. In: ECCV (2018) [3](#)
9. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. ICML (2017) [11](#)
10. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: NeurIPS (2016) [3](#)
11. Bishay, M., Zoumpourlis, G., Patras, I.: Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. In: BMVC (2019) [3](#)
12. Blattmann, A., Rombach, R., Oktay, K., Müller, J., Ommer, B.: Semi-parametric neural image synthesis. In: NeurIPS (2022) [4](#)
13. Bo, Y., Lu, Y., He, W.: Few-shot learning of video action recognition only based on video contents. In: WACV (2020) [6](#), [10](#), [21](#)
14. Boes, W., Van hamme, H.: Audiovisual transformer architectures for large-scale classification and synchronization of weakly labeled audio events. In: ACM MM (2019) [3](#)
15. Cao, K., Ji, J., Cao, Z., Chang, C.Y., Niebles, J.C.: Few-shot video classification via temporal alignment. In: CVPR (2020) [2](#), [3](#)
16. Carreira, J., Koppula, S., Zoran, D., Recasens, A., Ionescu, C., Henaff, O., Sheldhamer, E., Arandjelovic, R., Botvinick, M., Vinyals, O., et al.: Hierarchical perceiver. arXiv preprint arXiv:2202.10890 (2022) [6](#), [10](#), [21](#)
17. Chao, W.L., Changpinyo, S., Gong, B., Sha, F.: An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: ECCV (2016) [10](#), [11](#), [20](#)
18. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: CVPR (2021) [3](#)
19. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP (2020) [5](#)
20. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. arXiv:1904.04232 (2019) [3](#)
21. Chen, Y., Xian, Y., Koepke, A.S., Shan, Y., Akata, Z.: Distilling audio-visual knowledge by compositional contrastive learning. In: CVPR (2021) [3](#)
22. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NeurIPS (2021) [4](#)
23. Douze, M., Szlam, A., Hariharan, B., Jégou, H.: Low-shot learning with large-scale diffusion. In: CVPR (2018) [3](#)
24. Esser, P., Rombach, R., Blattmann, A., Ommer, B.: Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. NeurIPS (2021) [4](#)
25. Fayek, H.M., Kumar, A.: Large scale audiovisual learning of sounds with weakly labeled data. In: IJCAI (2020) [3](#), [6](#), [10](#), [21](#)
26. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: ECCV (2020) [3](#)
27. Gan, C., Huang, D., Chen, P., Tenenbaum, J.B., Torralba, A.: Foley music: Learning to generate music from videos. In: ECCV (2020) [3](#)
28. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: ICCV (2019) [3](#)
29. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv:1508.06576 (2015) [4](#)

30. Goldstein, S., Moses, Y.: Guitar music transcription from silent video. In: BMVC (2018) [3](#)
31. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *Communications of the ACM* (2020) [4](#)
32. Hariharan, B., Girshick, R.: Low-shot visual recognition by shrinking and hallucinating features. In: ICCV (2017) [3](#), [5](#)
33. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR (2015) [5](#)
34. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: ICASSP (2017) [4](#), [18](#)
35. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *NeurIPS* (2020) [8](#)
36. Iashin, V., Rahtu, E.: A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In: BMVC (2020) [3](#)
37. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017) [4](#)
38. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: ICML (2021) [6](#), [10](#), [21](#)
39. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: ICLR (2020) [3](#)
40. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014) [4](#), [18](#)
41. Kim, S., Choi, D.W.: Better generalized few-shot learning even without base data. *arXiv:2211.16095* (2022) [3](#)
42. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014) [9](#)
43. Koepke, A.S., Wiles, O., Moses, Y., Zisserman, A.: Sight to sound: An end-to-end approach for visual piano transcription. In: ICASSP (2020) [3](#)
44. Koepke, A.S., Wiles, O., Zisserman, A.: Visual pitch estimation. In: SMC (2019) [3](#)
45. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: *NeurIPS* (2018) [3](#)
46. Kumar Dwivedi, S., Gupta, V., Mitra, R., Ahmed, S., Jain, A.: Protogan: Towards few shot learning for action recognition. In: ICCVW (2019) [3](#), [4](#), [6](#), [10](#), [12](#), [21](#)
47. Li, X., Sun, Q., Liu, Y., Zhou, Q., Zheng, S., Chua, T.S., Schiele, B.: Learning to self-train for semi-supervised few-shot classification. *NeurIPS* (2019) [3](#)
48. Lin, Y.B., Wang, Y.C.F.: Audiovisual transformer with instance attention for audio-visual event localization. In: ACCV (2020) [3](#)
49. Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S.J., Yang, Y.: Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv:1805.10002* (2018) [3](#)
50. Majumder, S., Chen, C., Al-Halah, Z., Grauman, K.: Few-shot audio-visual learning of environment acoustics. In: *NeurIPS* (2022) [2](#)
51. Mercea, O.B., Hummel, T., Koepke, A.S., Akata, Z.: Temporal and cross-modal attention for audio-visual zero-shot learning. In: ECCV (2022) [2](#), [5](#), [6](#), [10](#), [11](#), [20](#), [21](#)

52. Mercea, O.B., Riesch, L., Koepke, A.S., Akata, Z.: Audio-visual generalised zero-shot learning with cross-modal attention and language. In: CVPR (2022) [2](#), [4](#), [5](#), [6](#), [10](#), [11](#), [20](#), [21](#)
53. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (2013) [8](#), [12](#)
54. Min, S., Yao, H., Xie, H., Wang, C., Zha, Z.J., Zhang, Y.: Domain-aware visual bias eliminating for generalized zero-shot learning. In: CVPR (2020) [20](#)
55. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv:1411.1784 (2014) [4](#)
56. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. NeurIPS (2021) [1](#), [2](#), [6](#), [10](#), [21](#)
57. Narasimhan, M., Ginosar, S., Owens, A., Efros, A.A., Darrell, T.: Strumming to the beat: Audio-conditioned contrastive video textures. arXiv:2104.02687 (2021) [3](#)
58. Narayan, S., Gupta, A., Khan, F.S., Snoek, C.G., Shao, L.: Latent embedding feedback and discriminative features for zero-shot classification. In: ECCV (2020) [3](#), [4](#)
59. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV (2018) [3](#)
60. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: ECCV (2016) [3](#)
61. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Learning sight from sound: Ambient sound provides supervision for visual learning. IJCV (2018) [3](#)
62. Patrick, M., Asano, Y.M., Fong, R., Henriques, J.F., Zweig, G., Vedaldi, A.: Multi-modal self-supervision from generalized data transformations. In: NeurIPS (2020) [1](#), [3](#)
63. Perrett, T., Masullo, A., Burghardt, T., Mirmehdi, M., Damen, D.: Temporal-relational crosstransformers for few-shot action recognition. In: CVPR (2021) [3](#)
64. Qi, H., Brown, M., Lowe, D.G.: Low-shot learning with imprinted weights. In: CVPR (2018) [3](#)
65. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR (2017) [3](#)
66. Recasens, A., Lin, J., Carreira, J., Jaegle, D., Wang, L., Alayrac, J.b., Luc, P., Miech, A., Smaira, L., Hemsley, R., et al.: Zorro: the masked multimodal transformer. arXiv preprint arXiv:2301.09595 (2023) [6](#), [10](#), [21](#)
67. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) [4](#)
68. Roy, A., Shah, A., Shah, K., Roy, A., Chellappa, R.: Diffalign: Few-shot learning using diffusion based synthesis and alignment. arXiv preprint arXiv:2212.05404 (2022) [3](#)
69. Saxena, D., Cao, J.: Generative adversarial networks (gans) challenges, solutions, and future directions. ACM Computing Surveys (CSUR) (2021) [4](#)
70. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. NeurIPS (2017) [3](#)
71. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402 (2012) [5](#)
72. Su, K., Liu, X., Shlizerman, E.: Multi-instrumentalist net: Unsupervised generation of music from body movements. arXiv:2012.03478 (2020) [3](#)
73. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: CVPR (2018) [3](#)

74. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS* (2020) 7
75. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: *ECCV* (2018) 3
76. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: *ICCV* (2015) 4, 18
77. Vahdat, A., Kreis, K., Kautz, J.: Score-based generative modeling in latent space. *NeurIPS* (2021) 4
78. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *NeurIPS* (2017) 7
79. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. *NeurIPS* (2016) 3
80. Wang, X., Zhu, L., Yang, Y.: T2vlad: global-local sequence alignment for text-video retrieval. In: *CVPR* (2021) 3
81. Wang, Y., Chao, W.L., Weinberger, K.Q., van der Maaten, L.: Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv:1911.04623* (2019) 3
82. Wang, Y.X., Girshick, R., Hebert, M., Hariharan, B.: Low-shot learning from imaginary data. In: *CVPR* (2018) 3
83. Xian, Y., Korbar, B., Douze, M., Torresani, L., Schiele, B., Akata, Z.: Generalized few-shot video classification with video retrieval and feature generation. *IEEE TPAMI* (2021) 2, 3, 4, 5, 6, 10, 11, 21
84. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: f-vaegan-d2: A feature generating framework for any-shot learning. In: *CVPR* (2019) 3, 4
85. Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C.: Audiovisual slowfast networks for video recognition. *arXiv:2001.08740* (2020) 1, 2
86. Ye, H.J., Hu, H., Zhan, D.C., Sha, F.: Few-shot learning via embedding adaptation with set-to-set functions. In: *CVPR* (2020) 3
87. Zhang, H., Zhang, L., Qi, X., Li, H., Torr, P.H., Koniusz, P.: Few-shot action recognition with permutation-invariant attention. In: *ECCV* (2020) 2
88. Zhang, Y.K., Zhou, D.W., Ye, H.J., Zhan, D.C.: Audio-visual generalized few-shot learning with prototype-based co-adaptation. *Proc. Interspeech 2022* (2022) 2
89. Zhou, H., Liu, Z., Xu, X., Luo, P., Wang, X.: Vision-infused deep audio inpainting. In: *ICCV* (2019) 3
90. Zhu, L., Yang, Y.: Compound memory networks for few-shot video classification. In: *ECCV* (2018) 2, 3

Supplementary: Text-to-feature diffusion for audio-visual few-shot learning

Otniel-Bogdan Mercea¹, Thomas Hummel¹, A. Sophia Koepke¹, and Zeynep Akata^{1,2}

¹ University of Tübingen

² MPI for Intelligent Systems

{otniel-bogdan.mercea, thomas.hummel,
a-sophia.koepke,zeynep.akata}@uni-tuebingen.de

In Section 1, we describe the procedure used to extract the audio and visual features that are used as inputs to our AV-DIFF framework. In Section 2, we provide additional experimental results for (G)FSL with 20 shots, along with reporting the GFSL performance on base and novel classes across all shots and datasets. Finally, we provide additional ablations on the hybrid attention and diffusion model.

1 Feature extraction

We train AV-DIFF on already pre-extracted temporal features for the audio and visual modalities. We used C3D [76] which was pre-trained on Sports1M [40] and VGGish [34] pre-trained on YouTube-8M [1] to extract audio and visual features respectively. Each audio feature is represented by a 128-dimensional vector corresponding to one second of audio data. To extract the visual features, we first resampled the videos to 25fps and then extracted a 4096-dimensional vector for 16 consecutive video frames.

2 Additional experimental results

We present (G)FSL results for 20 shots on the UCF-FSL, VGGSound-FSL and ActivityNet-FSL datasets in Section 2.1. In Section 2.2, we discuss the 1-,5-,10- and 20-shot (G)FSL performance on base and novel classes across all three datasets (which complements Section 5.2 of the main paper). Finally, Section 2.3 shows additional ablations on the hybrid attention and diffusion model.

2.1 (G)FSL in the 20-shot setting

In Table 1 (bottom), we provide additional (G)FSL results for the 20-shot setting with AV-DIFF and related methods. Similar to our observations in the main paper with 1, 5, and 10 shots, AV-DIFF achieves state-of-the-art performance for 20 shots, outperforming all related methods in the FSL and GFSL (HM) settings.

Similar to the conclusions for ActivityNet-FSL in the main paper, it can be observed that the ranking of baselines changes dramatically on ActivityNet-FSL, while AV-DIFF still remains the best, showing that our model is also more robust on 20 shots.

The HM and FSL performances on 20 shots for AV-DIFF and for the related methods are higher compared to the lower shots. The increase in performance for AV-DIFF from 10 to 20 shots is similar to the one from 5 to 10 shots. However, the most significant boost in performance happens between the 1-shot and 5-shot settings, showing that the gain in performance decreases as more training samples for novel classes are added. Similar trends can also be observed for the related methods.

2.2 Performance on base and novel classes

In the main paper, we only presented the GFSL results in terms of the harmonic mean of the performance on the B (base) and N (novel) classes (Table 2 in the main paper). The harmonic mean is crucial as it evaluates how robust a system is, and it also provides higher scores to systems which are very balanced and which are less biased towards either B or N . In this section, we are going to analyse the performance of the components that are used to calculate the HM, namely the B and N performance, to have a better idea of the model’s strengths and weaknesses. It can be seen in Table 1 that in the majority of cases, AV-DIFF obtains state-of-the-art performance on B and N , but there are still some exceptions, as presented below.

In the 1-shot setting, it can be observed that MBT outperforms AV-DIFF on N in VGGSound-FSL and B in UCF-FSL, with scores of 21.34% and 79.89% compared to 21.25% and 77.94% for AV-DIFF. However, MBT is very biased towards one of the metrics. On VGGSound-FSL, the bias is towards N , and MBT obtains a very low score on B , only 11.21%, compared to 19.44% for AV-DIFF. The same applies to UCF-FSL, where MBT is very biased towards B . For B on VGGSound-FSL, AV-DIFF obtains a performance of 19.44% compared to 28.55% SLDG. While AV-DIFF scores similarly on both metrics in VGGSound-FSL, SLDG obtains a B score which is more than twice that of N , showing how unbalanced and biased SLDG is. An interesting observation that can be made in the 1-shot setting is that on VGGSound-FSL, AV-DIFF is not able to attain state-of-the-art performance in B or N , but it still performs overall much better than the systems that outperform AV-DIFF in these two metrics.

In the 5-shot setting, AV-DIFF is outperformed on B in both VGGSound-FSL and UCF-FSL by the Perceiver, with scores of 31.46% and 83.56% compared to 30.88% and 74.11% for AV-DIFF. Moreover, on VGGSound-FSL, AV-DIFF is also outperformed on N by MBT with scores of 31.79% vs 31.50% for AV-DIFF. However, both MBT and Perceiver have a bigger bias towards one of the metrics, leading to a lower HM on VGGSound-FSL. On UCF-FSL, it can be clearly observed that Perceiver is biased towards B , obtaining a score which is more than twice that of N . For AV-DIFF this is not the case, as scores for both B and N are much more balanced.

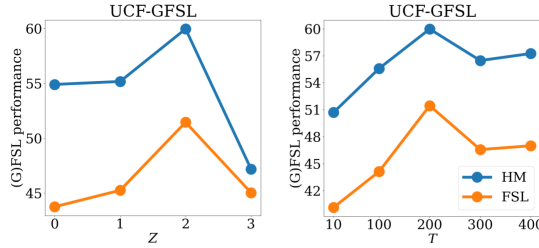


Fig. 1. (G)FSL performance (5-shot) for different numbers of self- (Z) and full attention layers (*left*), and different amounts of noise addition time steps T on UCF-FSL (*right*).

The same observations can be made in the 10- and 20-shot settings where sometimes AV-DIFF is outperformed in one of the B or N , but still achieves a higher HM overall. While most of the baselines that outperform AV-DIFF in one of the metrics are usually very biased towards that metric, this is not always the case. For example, in the 20-shot setting on UCF-FSL, Att. Fusion slightly outperforms AV-DIFF on N with a score of 61.02% compared to 59.94% for AV-DIFF. However, on B , AV-DIFF significantly outperforms Att. Fusion with a score of 86.51% compared to 79.39% for Att. Fusion. While in this case Att. Fusion is very well balanced, it is still worse overall than AV-DIFF, as it only slightly outperforms AV-DIFF in N but it is significantly outperformed in B .

Interestingly, for different methods, the N score is sometimes higher than B . This is likely due to the use of calibrated stacking [17]. Similar behaviour has been observed by several other works, such as [51, 52, 54]

Overall, AV-DIFF is not necessarily the best in both B and N every single time. However, across all shots and datasets, AV-DIFF achieves state-of-the-art GFSL performance in terms of the HM. This shows that AV-DIFF is the most balanced and robust among all the methods, as it can consistently score very high on both B and N .

2.3 Ablation on hybrid attention and diffusion.

In Fig 1 (left), we analyse the impact of the number of self-attention layers Z and full-attention layers used. For values of $Z < 2$ the performance increases consistently and reaches a peak performance at $Z = 2$ for both metrics on UCF-FSL. It appears that changing the attention in late layers of the network is beneficial. Finally, we ablate over the timesteps T for adding noise to the original feature in the diffusion model in Fig 1 (right). The (G)FSL performance maximizes for $T = 200$ on UCF-FSL which corresponds to the number of timesteps used in AV-DIFF.

Table 1. Novel (N) and base (B) performance for audio-visual (G)FSL: 1-shot, 5-shot, 10-shot, and 20-shot performance of AV-DIFF and compared methods on the VGGSound-FSL, UCF-FSL and ActivityNet-FSL datasets. The harmonic mean (HM) of the mean class accuracies for base and novel classes are reported for GFSL. The FSL performance considers only the test subset of novel classes.

1-shot	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
Att. F. [25]	15.16	15.77	15.46	16.37	38.91	35.98	37.39	36.88	3.48	5.78	4.35	5.82
Perc. [38]	18.46	17.51	17.97	18.51	74.57	31.33	44.12	33.73	30.32	12.14	17.34	12.53
MBT [56]	11.21	21.34	14.70	21.96	79.89	26.37	39.65	27.99	17.07	12.24	14.26	12.63
TCAF [51]	20.93	18.34	19.54	20.01	66.18	33.64	44.61	35.90	23.85	12.62	16.50	13.01
Proto [46]	8.85	13.65	10.74	14.08	60.12	27.72	37.95	28.08	2.02	4.40	2.77	4.40
SLDG [13]	28.55	11.94	16.83	17.57	73.15	27.45	39.92	28.91	23.22	9.58	13.57	10.30
TSL [83]	17.09	20.72	18.73	22.44	68.18	33.04	44.51	35.17	8.96	10.18	9.53	10.77
HiP [16]	23.39	16.39	19.27	18.64	16.20	33.26	21.79	34.88	25.02	9.53	13.80	10.31
Zorro [66]	17.49	20.51	18.88	21.79	67.85	32.94	44.35	34.52	19.67	11.55	14.56	11.94
AVCA [52]	4.53	10.28	6.29	10.29	82.86	29.59	43.61	31.24	14.15	11.73	12.83	12.22
AV-DIFF	19.44	21.26	20.31	22.95	77.94	38.46	51.50	39.89	32.77	12.86	18.47	13.80

5-shot	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
Att. F. [25]	28.64	27.82	28.22	31.57	63.27	43.69	51.68	47.18	5.00	8.05	6.17	8.13
Perc. [38]	31.46	28.52	29.92	33.58	83.56	34.27	48.60	40.47	35.66	20.15	25.75	21.50
MBT [56]	23.86	31.79	27.26	34.95	80.61	32.72	46.55	34.53	25.36	21.48	23.26	22.38
TCAF [51]	24.34	28.11	26.09	32.22	73.76	33.73	46.29	37.39	24.45	21.35	22.79	21.81
Proto [46]	25.27	25.08	25.17	28.87	63.69	31.79	42.42	33.63	1.61	7.81	2.67	7.81
SLDG [13]	29.74	15.98	20.79	25.17	65.44	25.28	36.47	28.56	29.40	17.95	22.29	19.16
TSL [83]	15.02	27.75	19.49	29.50	68.80	40.62	51.08	42.42	9.93	12.27	10.97	12.77
HiP [16]	30.01	24.18	26.82	30.67	33.65	39.74	36.44	42.23	21.98	15.39	18.10	16.25
Zorro [66]	29.06	30.07	29.56	35.17	69.13	41.49	51.86	42.59	25.72	21.03	23.14	21.94
AVCA [52]	13.24	20.15	15.98	20.50	84.80	34.64	49.19	36.70	19.18	21.09	20.09	21.65
AV-DIFF	30.88	31.50	31.19	36.56	74.11	50.35	59.96	51.45	35.84	21.61	26.96	23.00

10-shot	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
Att. F. [25]	26.87	35.89	30.73	39.02	73.53	47.77	57.91	52.19	12.58	9.27	10.67	10.78
Perc. [38]	32.64	34.73	33.65	40.73	71.88	44.97	55.33	47.86	37.06	25.03	29.88	26.46
MBT [56]	26.76	34.43	30.12	38.93	84.07	35.62	50.04	39.73	29.06	24.98	26.86	26.03
TCAF [51]	26.62	31.73	28.95	36.43	84.28	39.93	54.19	47.61	27.86	22.32	24.78	23.33
Proto [46]	30.48	29.26	29.85	34.80	70.28	40.03	51.01	40.68	2.63	8.81	4.05	8.81
SLDG [13]	28.32	20.99	24.11	29.48	49.35	26.29	34.31	26.96	34.69	23.20	27.81	25.35
TSL [83]	17.96	28.15	21.93	31.29	74.31	51.63	60.93	55.63	9.31	11.76	10.39	12.18
HiP [16]	28.43	30.12	29.25	35.13	75.54	38.14	50.69	43.29	24.32	16.10	19.37	17.06
Zorro [66]	28.48	36.68	32.06	40.66	82.88	45.67	58.89	49.06	30.11	25.05	27.35	26.33
AVCA [52]	13.39	27.83	18.08	28.27	71.96	38.93	50.53	39.17	26.36	25.68	26.02	26.76
AV-DIFF	32.15	36.05	33.99	41.39	84.62	51.69	64.18	57.39	37.91	26.02	30.86	27.81

20-shot	VGGSound-FSL				UCF-FSL				ActivityNet-FSL			
	B	N	HM	FSL	B	N	HM	FSL	B	N	HM	FSL
Att. F. [25]	31.43	37.88	34.35	44.08	79.39	61.02	69.00	63.20	15.51	11.41	13.15	13.22
Perc. [38]	33.11	37.66	35.24	43.77	77.81	48.29	59.59	52.66	32.30	31.06	31.67	32.21
MBT [56]	28.41	37.95	32.49	43.19	81.73	42.35	55.80	44.58	36.21	28.60	31.96	30.76
TCAF [51]	32.48	29.41	30.87	38.89	75.71	47.38	58.29	51.99	35.87	27.61	31.20	29.88
Proto [46]	31.44	32.66	32.04	38.42	61.07	49.32	54.57	50.48	25.05	8.17	12.32	14.65
SLDG [13]	33.20	19.53	24.59	33.30	81.08	39.52	53.14	43.95	32.60	30.80	31.68	32.44
TSL [83]	18.21	29.32	22.47	32.07	76.82	49.44	60.16	52.02	9.68	15.01	11.77	15.78
HiP [16]	32.03	29.83	30.89	38.46	71.59	43.43	54.06	48.07	33.78	17.59	23.13	20.67
Zorro [66]	29.84	39.46	33.98	43.63	87.82	48.46	62.45	57.10	34.15	28.55	31.10	30.31
AVCA [52]	15.30	32.20	20.75	32.64	60.00	44.93	51.39	44.93	24.47	29.88	26.91	30.76
AV-DIFF	33.17	39.46	36.04	44.79	86.51	59.94	70.82	65.73	39.25	31.06	34.68	32.89