X-AVDD: Cross-Attentive Audio-Visual Dataset Distillation

Saumyaranjan Mohanty[1] **Aravind Reddy**[2] **Konda Reddy Mopuri**[1]
[1]Department of Artificial Intelligence, Indian Institute of Technology Hyderabad
[2]Centre for Responsible AI, Wadhwani School of Data Science & AI, Indian Institute of Technology Madras
ai23resch04001@iith.ac.in aravind@cerai.in krmopuri@ai.iith.ac.in

Abstract

Audio-visual dataset distillation aims to synthesize a small dataset from a large paired audio-visual dataset, while nearly preserving the training utility for neural networks. Existing approaches such as DM and AVDD face two critical limitations: 1) they are memory-intensive, making them difficult to scale to complex architectures and higher samples-per-class (SPC) settings, and 2) they often fail to preserve strong cross-modal semantic correspondence. To tackle these issues, we propose **X-AVDD**, a scalable audio-visual dataset distillation framework that matches cross-modally conditioned representations using bi-directional multi-head cross-attention. By decoupling cross-modal alignment from joint optimization, X-AVDD is substantially more memory-efficient, reducing end-to-end synthesis cost by $\sim 8\times$ GPU-hours, while producing highly aligned synthetic data. Empirically, X-AVDD significantly outperforms other audio-visual distillation methods. On VGGSound, X-AVDD improves top-1 test accuracy from $8.2\% \rightarrow 24.0\%$ at SPC=10 and $9.8\% \rightarrow 26.5\%$ at SPC=20. Furthermore, X-AVDD improves cross-architecture generalization (e.g., $25.5\% \rightarrow 35.3\%$ for a ViT target on the AVE dataset at SPC $= 50$) and successfully scales to complex architectures such as AudioCLIP, achieving 62.3% accuracy using only 5.65% of the full dataset, compared to 72.4% with full-data training, while reducing downstream training time by approximately $9\times$. The code is available at `https://github.com/DIL-IITH/XAVDD`.

1 Introduction

Modern multimodal models derive much of their capability from training on massive datasets that jointly capture complementary signals across modalities (e.g., vision, audio, text) [Qin et al., 2025]. While such datasets enable richer representations and improved robustness over single-modality learning, their scale creates a growing bottleneck: storing, curating, and repeatedly training on multimodal corpora is expensive in computation, energy, and time [Schwartz et al., 2020, Strubell et al., 2019]. Recent work has therefore increasingly emphasized *data-centric* methods to improve data quality, and data curation as a key lever for efficiently scaling multimodal models [Qin et al., 2025].

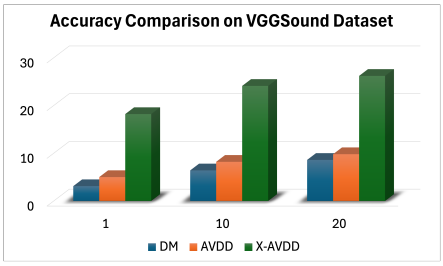

Figure 1: Accuracy on the VGGSound dataset under different samples-per-class (SPC) settings. X-AVDD consistently outperforms DM [Zhao and Bilen, 2023] and AVDD [Kushwaha et al., 2024].

Dataset distillation (DD), also known as Dataset Condensation, is an increasingly popular route to data-efficient learning. DD methods synthesize a

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

small set of labeled synthetic examples such that models trained on the synthetic data achieve performance comparable to that of models trained on the full dataset [Wang et al., 2018, Zhao et al., 2021, Sachdeva and McAuley, 2023, Liu and Du, 2025]. Distilled datasets enable rapid iteration for tasks such as neural architecture search and hyperparameter tuning [Ding et al., 2024, Such et al., 2020] and provide a compact interface for sharing training signals while maintaining privacy [Dong et al., 2022].

As DD moves beyond unimodal settings (primarily image classification), *multimodal* distillation [Wu et al., 2024] introduces a central additional challenge: the synthetic data must preserve not only modality-specific information, but also the cross-modal associations that make multimodal learning effective. In this paper, we focus on the multimodal distillation problem of *audio-visual dataset distillation* introduced in [Kushwaha et al., 2024]. Given a large-scale paired audio-visual dataset, we want to synthesize a much smaller dataset such that downstream models trained on the synthetic dataset have good performance. Following Kushwaha et al. [2024], we adopt audio-visual event recognition as the primary evaluation task, representing visual inputs using sampled video frames and audio inputs using corresponding spectrograms.

AVDD extends Distribution Matching (DM) [Zhao and Bilen, 2023] to paired audio-visual data. However, this approach faces two fundamental limitations:

Memory Inefficiency. It is highly memory-intensive, making it difficult to scale distillation to higher samples-per-class (SPC) settings and complex modern architectures.

Modality Gap. Matching modality-specific feature distributions does not explicitly preserve the interactions between audio and visual representations. As illustrated in Figure 2, synthetic datasets generated by AVDD exhibit larger class-wise l_2 distances between audio and visual feature centers than the original data, indicating a greater modality gap [Liang et al., 2022, Zhang et al., 2025].

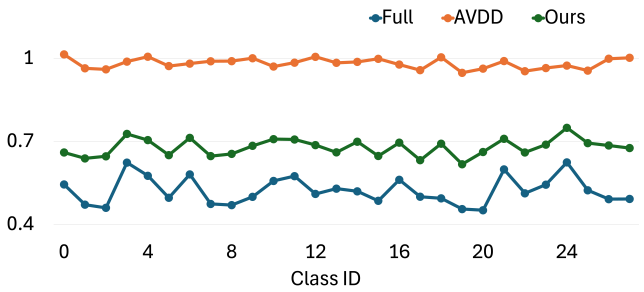

Figure 2: Modality gap (class-wise ℓ_2 distance between audio and visual class centers in the feature space) on the AVE dataset, and datasets distilled by AVDD and X-AVDD using AudioCLIP network. X-AVDD achieves a consistently lower modality gap.

To address these limitations, we propose *X-AVDD*, a *cross-attentive* audio-visual dataset distillation method designed for scalability and stronger cross-modal coupling. The key idea is to *explicitly capture inter-modality relationships during distillation* using bi-directional, multi-head cross-attention, while *decoupling cross-modal alignment from joint optimization* via a pre-trained teacher model. X-AVDD synthesizes audio and visual samples by matching cross-modally conditioned embeddings, where representations from one modality attend to information from the other. While cross-attention is well established in multimodal learning [Tsai et al., 2019, Praveen et al., 2022], our major contribution lies in integrating cross-conditioned representation matching with a decoupled synthesis strategy for audio-visual dataset distillation.

This design substantially improves efficiency: X-AVDD reduces total synthesis time on a Tesla V100 GPU from **10 hours** to **1.3 hours**. Figure 1 presents the impressive generalization performance obtained by X-AVDD compared to the baseline DM and AVDD methods.

Contributions.

- We introduce **X-AVDD**, a cross-attention-guided audio-visual dataset distillation framework that explicitly models inter-modal relationships during audio-visual dataset distillation.

- We propose a **decoupled cross-modal alignment strategy** using bi-directional cross-attention, yielding substantially improved **memory efficiency** and enabling scaling to larger SPC and modern architectures.
- We demonstrate improved classification performance, cross-modal retrieval, and cross-architecture generalization across multiple audio-visual benchmarks. Our experiments further establish the scalability of X-AVDD to higher SPC budgets and complex architectures, alongside improvements in modality alignment and synthesis efficiency.

2 Related Works

2.1 Dataset Distillation

Dataset Distillation (DD), also referred to as Dataset Condensation, aims to synthesize a compact set of labeled examples that preserves the training signal of a much larger original dataset. While early methods rely on bi-level optimization [Wang et al., 2018] (e.g., gradient matching [Zhao et al., 2021, Zhao and Bilen, 2021], trajectory matching [Cazenavette et al., 2022, Cui et al., 2023]), which are quite memory and compute intensive. To improve scalability, distribution-matching methods [Zhao and Bilen, 2023] align the statistics of real and synthetic embeddings. However, these methods often encounter memory bottlenecks as the synthetic budget or model complexity increases. We refer the reader to recent surveys for broader coverage and taxonomies of DD methods Sachdeva and McAuley [2023], Lei and Tao [2023], Liu and Du [2025].

2.2 Attention in DD

While attention has been explored in unimodal distillation, such as DataDAM [Sajedi et al., 2023], which matches spatial attention maps, our work is the first to employ bi-directional cross-attention in a multimodal context. Unlike DataDAM's focus on unimodal spatial patterns, X-AVDD uses bi-directional multi-head cross-attention to explicitly model the temporal and semantic dependencies between audio and visual streams during synthesis.

2.3 Multimodal Dataset Distillation (MDD)

MDD extends distillation to settings that require preserving cross-modal associations. While vision-language distillation [Wu et al., 2024, Xu et al., 2024, Zhang et al., 2025, Zhao et al., 2025, Lee and Chung, 2025] has seen rapid progress, audio-visual distillation remains under-explored. A primary challenge in MDD is the "modality collapse", in which synthetic pairs lose semantic alignment [Zhang et al., 2025]. Our work is the first to address this in the audio-visual domain by using a decoupled alignment strategy to proactively minimize this gap.

2.4 Audio-Visual Event Recognition

Recognizing events in unconstrained videos requires learning aligned representations across audio and visual modalities [Brousmiche et al., 2022]. Benchmark datasets such as AVE [Tian et al., 2018] and VGGSound [Chen et al., 2020] have driven the development of complex architectures, such as AudioCLIP [Guzhov et al., 2022]. Contrastive learning and cross-modal attention have become prominent directions for learning aligned representations across modalities [Arandjelovic and Zisserman, 2017, Miech et al., 2019]. However, the computational cost of training these models on the full dataset is quite high. Our work contributes to this domain by making audio-visual training more data-efficient through distillation, enabling faster iteration and lower training cost in this setting.

2.5 Limitations of AVDD

Despite establishing the first framework for audio-visual dataset distillation, AVDD faces a fundamental scalability bottleneck due to its reliance on a non-decoupled Distribution Matching (DM) paradigm. In AVDD, the entire set of synthetic samples must be held in GPU VRAM simultaneously during the optimization loop to match the global feature distributions, leading to Out-of-Memory (OOM) failures as the SPC budget or model complexity increases. While a class-wise batching strategy could theoretically reduce memory, our empirical tests show that it drastically degrades performance (e.g.,

a drop from 58.0% to 49.5% at SPC=20), because it loses the inter-class relationship information necessary for effective distribution matching. Consequently, AVDD is largely restricted to simple ConvNet architectures and lower SPC settings, failing to generalize to modern high-parameter models like ViT [Dosovitskiy et al., 2021] or AudioCLIP [Guzhov et al., 2022]. Furthermore, AVDD involves a "buffer creation" cost (reflected in our AVDD timing breakdown in Table 15).

X-AVDD addresses these limitations by decoupling the alignment phase from the synthesis loop, enabling class-wise optimization via frozen, cross-attentive-guided networks that maintain superior accuracy while reducing peak VRAM usage and total synthesis time.

2.6 Comparison with DAVDD

A concurrent work, Decoupled Audio-Visual Dataset Distillation (DAVDD) [Li et al., 2025a] proposes a framework that uses a bank of frozen encoders pretrained on the target dataset and lightweight decouplers to obtain shared representations. The decouplers split each feature into a private representation and a common representation. Distribution matching is applied to the private features within each modality, whereas Common Intermodal Matching aligns the common features across modalities. The differences between X-AVDD and DAVDD are:

- DAVDD uses a pre-trained encoder bank; X-AVDD utilizes a cross-attention-based teacher network.
- A decoupler bank is utilized by DAVDD, while X-AVDD uses bi-directional cross-attention.
- While DAVDD uses MMD for private representation matching and common intermodal matching, X-AVDD uses Euclidean Distribution Matching among cross-attended embeddings.
- DAVDD provides results for VGGSound-10K, AVE, and Music-21, but not for the larger and more complex VGGSound Dataset, and limits itself to SPC=20. On the other hand, X-AVDD provides results on all four datasets, including the 309-class VGGSound dataset. In most settings (except for SPC=1 with AVE and the VGGSound-10K dataset), X-AVDD outperforms DAVDD.
- While DAVDD limits itself to SPC=20, X-AVDD provides results for up to SPC=50, proving its scalability. While DAVDD provides results only for ConvNet, X-AVDD scales to more complex architectures such as ResNet-18, ViT, and AudioCLIP.

3 Methodology

In this section, we describe our proposed X-AVDD in detail.

Problem formulation. Let the labeled dataset be $\mathcal{V} = \{(x_i^a, x_i^v, y_i)\}_{i=1}^N$, where $x_i^a \in \mathcal{X}^a$ and $x_i^v \in \mathcal{X}^v$ denote the audio and visual modalities of the i-th sample, respectively, and $y_i \in \mathcal{Y}$ is the corresponding ground-truth label. The visual input x_i^v is represented by the central frame of the video, while the audio input x_i^a is represented by the corresponding spectrogram.

The goal of DD is to learn a compact synthetic dataset $\mathcal{S} = \{(\tilde{x}_i^a, \tilde{x}_i^v, \tilde{y}_i)\}_{i=1}^M, M \ll N$, such that a model parameterized by $\theta^{\mathcal{S}}$, trained on $\mathcal{S}$, achieves generalization performance comparable to a model $\theta^{\mathcal{V}}$ trained on the full dataset $\mathcal{V}$.

Cross Attention. This is a fundamental attention mechanism that allows different modalities to exchange information in a multi-modal setting [Tsai et al., 2019]. Information exchange occurs through queries from one modality attending to keys and values from another modality. Cross-attention enables fine-grained semantic alignment across different input streams. Intuitively, for audio-visual dataset distillation, cross-attention can capture complementary information [Praveen et al., 2022, Zeng et al., 2023] between the two modalities and guide distillation steps towards generating synthetic samples that capture informative semantics from the real dataset.

Some recent works [Hiller et al., 2024, Si et al., 2025] have explored bidirectional cross-attention for training models on multimodal datasets. Bi-directional cross-attention is a mechanism in multimodal models in which two modalities (e.g., text and images, or audio and video) attend to each other in both directions, allowing each to refine its representation based on the other. Concretely, one attention pass lets modality A query and extract relevant information from modality B, while a

second, symmetric pass lets modality B query modality A. This mutual exchange captures richer, more aligned relationships than one-way fusion. This bidirectional interaction improves tasks such as alignment, retrieval, and generation by modeling inter-dependencies rather than treating one modality as dominant, leading to more coherent, semantically grounded multimodal representations. Figure 3 shows the block diagram of the implementation of the bi-directional cross-attention mechanism.

Formulation of X-AVDD. In this paper, we propose *X-AVDD*, a cross-attention-based decoupled dataset distillation method. Our proposed method decouples the distillation of synthetic data from the training of the teacher model on the original dataset. This decoupling ensures the distillation process can be parallelized and does not depend on having access to the original large dataset. We incorporate a bidirectional cross-attention module in the teacher network that enables the extraction of complementary information from both modalities. Our method consists of three steps.

1. Pre-training of the teacher network, which incorporates bi-directional cross-attention.
2. Computation (and storage) of real embeddings for both audio and visual modalities through the cross-attention layers.
3. Iterative learning of a synthetic dataset by matching cross-attended embeddings from real and synthetic samples.

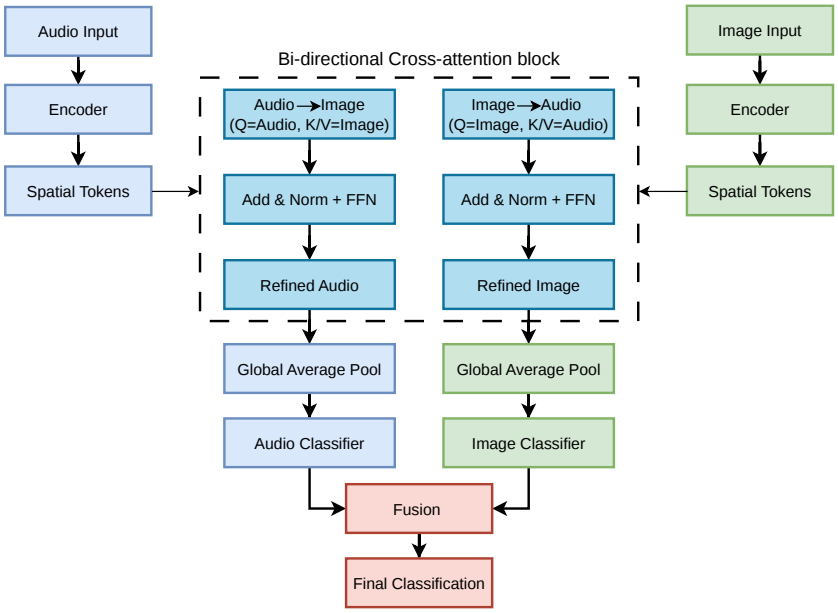

Figure 3: Overview of the bi-directional cross-attention mechanism incorporated into the teacher network for Audio-Visual event recognition.

In our approach, we use randomly sampled audio and frame pairs from real data to initialize the distillation process (similar to Cazenavette et al. [2022] and Du et al. [2024]). For the experiments, the same set of initialization samples is used for both the baseline and our proposed method. We use a Euclidean Distribution Matching (EDM) loss function that encourages synthetic embeddings to cluster near real ones in Euclidean space by minimizing the pairwise Euclidean distances between attended embeddings from real and synthetic samples. Pairwise Euclidean distance between two matrices $\mathbf{A}$ and $\mathbf{B}$ of dimensions (M, d) and (N, d) respectively, is calculated as:

$$D_{\text{euc}}(\mathbf{A}, \mathbf{B}) = \sum_{j=1}^{N} \sum_{i=1}^{M} \|\mathbf{A}[i, :] - \mathbf{B}[j, :]\|_2 \tag{1}$$

Let $\{x_{ar}^C, x_{vr}^C\}$ and $\{x_{as}^C, x_{vs}^C\}$ be the real and synthetic audio-image sample pairs belonging to class C, respectively. Let h_a and h_v be audio and visual encoders, and τ_a and τ_v be the tokenization operators for audio and visual modalities, respectively. The spatial tokens for real and synthetic

samples are expressed as:

$$\mathbf{A}_{\mathbf{r}}^{\mathbf{C}} = \tau_a(h_a(x_{ar}^C)) \qquad \mathbf{V}_{\mathbf{r}}^{\mathbf{C}} = \tau_v(h_v(x_{vr}^C)) \tag{2}$$

$$\mathbf{A}_{\mathbf{s}}^{\mathbf{C}} = \tau_a(h_a(x_{as}^C)) \qquad \mathbf{V}_{\mathbf{s}}^{\mathbf{C}} = \tau_v(h_v(x_{vs}^C)) \tag{3}$$

Let $\mathcal{T}_a$ and $\mathcal{T}_v$ be the audio $\rightarrow$ image and image $\rightarrow$ audio branches of the bi-directional cross-attention block, respectively. The output of the Global Average Pool (GAP) is the mean of the cross-attended tokens for the corresponding real data, and these are stored and used during the distillation process. Let g_a and g_v be the GAP operators for the audio and visual branches, respectively. The attended embeddings for real data can be represented as:

$$\mathbf{E}_{r(A\rightarrow V)}^C = g_a(\mathcal{T}_a(\mathbf{A}_{\mathbf{r}}^{\mathbf{C}}, \mathbf{V}_{\mathbf{r}}^{\mathbf{C}})) \qquad \mathbf{E}_{r(V\rightarrow A)}^C = g_v(\mathcal{T}_v(\mathbf{A}_{\mathbf{r}}^{\mathbf{C}}, \mathbf{V}_{\mathbf{r}}^{\mathbf{C}})) \tag{4}$$

The attended embeddings for the synthetic data are given by:

$$\mathbf{E}_{s(V\rightarrow A)}^C = g_v(\mathcal{T}_v(\mathbf{A}_{\mathbf{s}}^{\mathbf{C}}, \mathbf{V}_{\mathbf{r}}^{\mathbf{C}})) \qquad \mathbf{E}_{s(A\rightarrow V)}^C = g_a(\mathcal{T}_a(\mathbf{A}_{\mathbf{r}}^{\mathbf{C}}, \mathbf{V}_{\mathbf{s}}^{\mathbf{C}})) \tag{5}$$

Note that the audio-guided image embeddings $\mathbf{E}_{s(A\rightarrow V)}^C$ are computed by using the synthetic image as query and the real audio as key and value. Similarly, the image-guided audio embeddings $\mathbf{E}_{s(V\rightarrow A)}^C$ are computed using synthetic audio as the query and real image as key and value. As stated earlier, in order to encourage synthetic embeddings to cluster near real ones, we perform Euclidean distribution matching (EDM), defined as:

$$l_{A\rightarrow V}^C = D_{\text{euc}}(\mathbf{E}_{r(A\rightarrow V)}^C, \mathbf{E}_{s(A\rightarrow V)}^C) \tag{6}$$

$$l_{V\rightarrow A}^C = D_{\text{euc}}(\mathbf{E}_{r(V\rightarrow A)}^C, \mathbf{E}_{s(V\rightarrow A)}^C) \tag{7}$$

The complete objective for synthesizing the distilled dataset over the frozen teacher is (note that the student network is not part of this and will only be trained from scratch using the distilled data post-synthesis):

$$l^C = \lambda_{ce} l_{ce}^C + \lambda_{av} l_{(A\rightarrow V)}^C + \lambda_{va} l_{(V\rightarrow A)}^C \tag{8}$$

where l_{ce}^c is the standard cross-entropy loss of the classification task, and $\lambda_{ce}, \lambda_{av}$, and λ_{va} are corresponding weights of the loss functions.

4 Experiments and Results

Datasets and protocol. We evaluate our method on four benchmark audio-visual datasets: AVE [Tian et al., 2018], MUSIC-21 [Zhao et al., 2019], VGGSound [Chen et al., 2020], and VGGSound-10K (10-class subset released with AVDD [Kushwaha et al., 2024]). We use AVDD's preprocessing and splits for fair comparison, and report mean and standard deviation of test accuracy over five seeds (0–4). See Appendix D for dataset and preprocessing details.

Backbone architecture. Following AVDD, we first employ a ConvNet architecture [Zhao et al., 2021] for both audio and video encoders. Additionally, we employed more complex architectures, such as ResNet-18 [He et al., 2016], ViT [Dosovitskiy et al., 2021], and AudioCLIP [Guzhov et al., 2022], to thoroughly evaluate X-AVDD's applicability.

Cross-architecture comparison. To validate the effectiveness of the synthetic dataset, we train randomly initialized models, which are different from the backbone architecture utilized to generate the synthetic images. We used ResNet-18 to generate synthetic images, and VGG-11 [Simonyan and Zisserman, 2015] and ViT [Dosovitskiy et al., 2021] models for cross-architecture comparison.

Baselines. For the ConvNet architecture, we compared AVDD-reported results with those obtained by our proposed method. For the other architectures, AVDD results in an OOM issue due to hardware limitations. We instead designed and implemented a new multimodal adaptation of SRe^{2}L [Yin et al., 2023], as a baseline for distillation, and compared it against X-AVDD. Details of the multimodal adaptation of SRe^{2}L are provided in Appendix C. Hyperparameter details are provided in Appendix E.

4.1 Results for ConvNet, ResNet-18 and ViT architectures

Table 1 provides test set accuracies obtained by training ConvNet, ResNet-18, and ViT architectures from scratch using the distilled dataset for various Samples Per Class (SPC) settings. We can see that X-AVDD significantly outperforms AVDD across almost all SPC settings. In the ConvNet architecture, at extremely low SPC, the extra cross-modal constraints can over-regularize, whereas benefits emerge as soon as synthetic diversity increases (SPC $\geq$ 10). There is also a particularly significant jump across all SPCs for VGGSound, which is by far the largest dataset used in audio visual dataset distillation.

Table 1: Comparison of accuracy (%) on synthetic datasets for various architectures with the ensemble fusion method. For the ConvNet architecture, comparisons are carried out against DM and AVDD, while for ResNet-18 and ViT, comparisons are carried out against a multimodal implementation of SRe^{2}L. SPC denotes samples per class. Bold indicates best performance within each architecture group. - indicates Out-of-Memory (OOM).

			ConvNet			**ResNet-18**		**ViT**	
Dataset	SPC	Ratio (%)	DM	AVDD	X-AVDD	SRe^{2}L	X-AVDD	SRe^{2}L	X-AVDD
VGGS-10K	1	0.11	36.5 ± 2.5	$\mathbf{40.4 \pm 1.8}$	37.6 ± 0.8	21.7 ± 0.6	$\mathbf{30.0 \pm 1.5}$	16.7 ± 0.9	$\mathbf{20.2 \pm 0.5}$
	10	1.13	43.8 ± 1.7	55.0 ± 1.7	$\mathbf{57.2 \pm 0.9}$	46.9 ± 2.0	$\mathbf{54.5 \pm 0.7}$	25.4 ± 0.5	$\mathbf{30.0 \pm 0.7}$
	20	2.26	49.0 ± 2.4	58.0 ± 1.7	$\mathbf{62.7 \pm 0.5}$	52.7 ± 0.5	$\mathbf{61.5 \pm 0.8}$	26.1 ± 1.8	$\mathbf{35.9 \pm 0.8}$
	50	5.65	-	-	$\mathbf{66.1 \pm 0.3}$	60.4 ± 0.9	$\mathbf{66.7 \pm 0.9}$	30.2 ± 1.2	$\mathbf{42.1 \pm 1.2}$
	Full			74.7		77.4		62.1	
AVE	1	0.1	21.7 ± 1.5	$\mathbf{23.0 \pm 1.4}$	22.0 ± 0.4	18.5 ± 1.0	$\mathbf{20.3 \pm 0.8}$	11.2 ± 0.8	$\mathbf{12.4 \pm 0.7}$
	10	1.0	28.1 ± 1.8	36.8 ± 0.9	$\mathbf{43.5 \pm 0.2}$	41.2 ± 0.5	$\mathbf{46.9 \pm 0.7}$	24.5 ± 0.7	$\mathbf{26.7 \pm 0.1}$
	20	2.0	32.6 ± 1.0	40.1 ± 1.0	$\mathbf{50.2 \pm 0.8}$	45.9 ± 0.3	$\mathbf{50.3 \pm 0.3}$	25.3 ± 2.3	$\mathbf{31.1 \pm 0.8}$
	50	5.0	-	-	$\mathbf{56.2 \pm 0.6}$	49.1 ± 0.3	$\mathbf{56.3 \pm 0.6}$	27.8 ± 1.3	$\mathbf{33.0 \pm 0.4}$
	Full			58.5		59.2		47.9	
MUSIC-21	1	0.014	38.3 ± 1.3	44.0 ± 2.2	$\mathbf{47.9 \pm 0.7}$	21.6 ± 0.8	$\mathbf{30.8 \pm 2.1}$	13.6 ± 1.3	$\mathbf{21.0 \pm 0.5}$
	10	0.14	54.8 ± 1.4	68.1 ± 1.0	$\mathbf{74.5 \pm 0.5}$	58.7 ± 1.4	$\mathbf{70.9 \pm 0.3}$	37.4 ± 0.3	$\mathbf{44.5 \pm 0.1}$
	20	0.28	61.1 ± 1.3	70.3 ± 0.7	$\mathbf{77.6 \pm 0.9}$	61.2 ± 2.9	$\mathbf{74.1 \pm 0.4}$	41.6 ± 0.9	$\mathbf{48.0 \pm 1.4}$
	50	0.70	-	-	$\mathbf{79.7 \pm 0.3}$	71.9 ± 0.6	$\mathbf{78.4 \pm 0.3}$	52.2 ± 1.2	$\mathbf{56.7 \pm 0.8}$
	Full			80.5		81.4		71.3	
VGGSound	1	0.18	3.1 ± 0.2	5.0 ± 0.3	$\mathbf{18.3 \pm 0.3}$	7.9 ± 0.9	$\mathbf{10.9 \pm 0.2}$		
	10	1.87	6.4 ± 0.3	8.2 ± 0.2	$\mathbf{24.0 \pm 0.1}$	17.3 ± 1.4	$\mathbf{22.4 \pm 0.5}$		
	20	3.74	8.6 ± 0.1	9.8 ± 0.3	$\mathbf{26.5 \pm 1.1}$	21.1 ± 0.3	$\mathbf{25.1 \pm 0.1}$		
	50	9.35	-	-	$\mathbf{28.2 \pm 0.7}$	22.2 ± 0.5	$\mathbf{28.4 \pm 0.9}$		
	Full			30.1		32.1			

4.2 Results for AudioCLIP architecture

AudioCLIP [Guzhov et al., 2022] extends CLIP to the Audio modality along with image and text, incorporating ESResNeXt audio-model into the CLIP framework. Table 2 provides a comparative analysis of accuracy values obtained on the test set of VGGSound-10K and AVE datasets with synthetic images generated by our proposed method. With a fraction of the time the model takes to train on the full dataset, distilled images can achieve comparable generalization accuracy, underscoring the role of distillation in training an audio-visual dataset with SOTA architectures.

4.3 Cross architecture analysis

To study the generalization capability of the generated synthetic dataset, we carried out a cross-architecture study with two different architectures compared to the base model through which the synthetic dataset is generated. We utilized VGG-11 [Simonyan and Zisserman, 2015], and transformer-based architecture, ViT [Dosovitskiy et al., 2021]. The results are listed in Table 3. Our proposed method significantly outperforms the baseline across various datasets and SPC settings. Significant improvement over the baseline on the MUSIC-21 dataset with ViT as the target architecture provides further proof of the efficacy of our proposed method.

4.4 Transfer learning with frozen backbone of SOTA image and audio encoder models

In this experiment, we investigate whether distilled datasets generated using simpler models such as ConvNet and ResNet-18 are sufficient to train lightweight classifiers on top of frozen, state-of-the-art image and audio encoder models. We used two complementary setups for this experiment. The first setup uses the audio and visual backbones of ImageBind [Girdhar et al., 2023], which are designed to

Table 2: Comparison of test accuracy obtained by synthetic data generated by X-AVDD on VGGSound-10K and AVE datasets with AudioCLIP architecture against baseline methods. The ratio is the percentage of the condensed dataset compared to the full dataset. X-AVDD outperforms both baselines at every SPC, and at SPC=50 it reaches 86% (VGGSound-10K) and 95% (AVE) of the full-data accuracy, while AVDD runs out of memory. Training-time comparisons are in Table 17. **Bold** denotes the best distilled-data result. OOM indicates out-of-memory.

SPC	**VGGSound-10K**				**AVE**			
	Ratio (%)	**AVDD**	**SRe^{2}L**	**X-AVDD**	**Ratio (%)**	**AVDD**	**SRe^{2}L**	**X-AVDD**
1	0.11	10.3 ± 0.9	14.5 ± 1.1	$\mathbf{24.8 \pm 0.6}$	0.1	25.4 ± 1.3	32.1 ± 0.8	$\mathbf{44.0 \pm 1.2}$
10	1.13	16.7 ± 1.2	23.4 ± 0.8	$\mathbf{33.7 \pm 0.3}$	1.0	44.5 ± 1.8	60.5 ± 1.6	$\mathbf{74.5 \pm 0.9}$
20	2.26	24.5 ± 0.7	29.7 ± 1.3	$\mathbf{46.8 \pm 0.9}$	2.0	53.6 ± 0.9	65.3 ± 1.1	$\mathbf{76.8 \pm 1.1}$
50	5.65	OOM	50.3 ± 0.9	$\mathbf{62.3 \pm 0.8}$	5.0	OOM	70.9 ± 1.7	$\mathbf{81.3 \pm 0.9}$
Full	100		72.4 ± 0.7		100		85.6 ± 1.1	

Table 3: Comparative analysis of cross-architecture generalization performance of the distilled dataset generated using the ResNet-18 architecture for both audio and video modalities. We experimented with two target architectures: VGG11 and ViT. For both, the synthetic dataset generated by X-AVDD achieves higher generalization accuracy than that generated by the SRe^{2}L adaptation. Results are provided for the Ensemble fusion method.

		VGGSound-10K		AVE		Music-21	
Target	SPC	SRe^{2}L	X-AVDD	SRe^{2}L	X-AVDD	SRe^{2}L	X-AVDD
VGG-11	1	14.6 ± 3.7	$\mathbf{20.4 \pm 2.3}$	6.2 ± 2.5	$\mathbf{11.4 \pm 0.7}$	18.2 ± 1.9	$\mathbf{22.3 \pm 1.2}$
	10	34.8 ± 1.3	$\mathbf{41.3 \pm 2.2}$	21.8 ± 0.4	$\mathbf{29.9 \pm 0.7}$	39.7 ± 1.6	$\mathbf{47.3 \pm 2.6}$
	20	44.0 ± 2.4	$\mathbf{51.5 \pm 0.8}$	29.1 ± 0.3	$\mathbf{35.6 \pm 0.6}$	42.8 ± 1.1	$\mathbf{58.7 \pm 1.9}$
	50	53.2 ± 0.2	$\mathbf{59.4 \pm 0.2}$	35.0 ± 0.7	$\mathbf{42.4 \pm 0.3}$	48.5 ± 0.5	$\mathbf{64.3 \pm 1.1}$
ViT	1	13.9 ± 1.1	$\mathbf{23.3 \pm 0.9}$	12.6 ± 0.8	$\mathbf{13.9 \pm 0.1}$	10.0 ± 0.5	$\mathbf{14.4 \pm 1.8}$
	10	30.0 ± 2.9	$\mathbf{39.9 \pm 0.8}$	21.0 ± 0.4	$\mathbf{27.2 \pm 1.1}$	26.7 ± 2.0	$\mathbf{44.5 \pm 0.8}$
	20	38.6 ± 0.6	$\mathbf{46.2 \pm 0.7}$	23.1 ± 0.3	$\mathbf{30.4 \pm 0.7}$	28.8 ± 0.7	$\mathbf{45.1 \pm 0.9}$
	50	44.3 ± 1.1	$\mathbf{51.3 \pm 1.4}$	25.5 ± 2.4	$\mathbf{35.3 \pm 0.5}$	29.0 ± 0.9	$\mathbf{56.5 \pm 1.3}$

map multimodal data onto a shared feature space. The second setup leverages Audio Spectrogram Transformer (AST) [Gong et al., 2021] as the audio encoder and SigLIP2 [Tschannen et al., 2025] as the visual encoder. We chose AST because of its strong performance on music/environmental audio domains on MAEB [Assadi et al., 2026], and chose SigLIP2 [Tschannen et al., 2025] because the SigLIP family is among the strongest open image embedding models on MIEB [Xiao et al., 2025]. MAEB and MIEB are integrated with the very popular MTEB benchmark [Muennighoff et al., 2023]. Small projection heads and linear classifiers are then trained on the extracted, frozen features using only the distilled synthetic data.

We compare the performance of data distilled by AVDD and X-AVDD using ConvNet, and the data distilled by SRe^{2}L and X-AVDD using ResNet-18. Table 4 tabulates these results for the VGGSound-10K dataset. With only 50 samples per class, the dataset distilled by X-AVDD using ResNet-18 achieves performance closer to training with the full dataset, highlighting the capability of X-AVDD and its adaptation to powerful pretrained foundation models for downstream audio-visual classification tasks.

Table 4: Performance analysis of distilled dataset for transfer learning with frozen backbone of SOTA models. X-AVDD (C) and X-AVDD (R) denote the synthetic datasets generated by ConvNet and ResNet-18, respectively. With only 50 samples per class, synthetic data generated by the ResNet-18 architecture and X-AVDD achieves performance comparable to that of training on the full dataset.

	ImageBind				AST + SigLIP2			
SPC	AVDD	X-AVDD (C)	SRe^{2}L	X-AVDD (R)	AVDD	X-AVDD (C)	SRe^{2}L	X-AVDD (R)
1	20.1 ± 0.4	25.7 ± 0.8	25.2 ± 0.6	$\mathbf{28.3 \pm 0.4}$	23.1 ± 1.3	24.3 ± 0.3	18.1 ± 1.2	$\mathbf{37.3 \pm 0.5}$
10	26.4 ± 0.9	40.9 ± 0.6	40.1 ± 0.8	$\mathbf{58.4 \pm 0.5}$	36.1 ± 1.0	42.2 ± 0.7	49.8 ± 0.7	$\mathbf{58.7 \pm 0.9}$
20	33.2 ± 0.5	43.1 ± 0.4	43.7 ± 0.3	$\mathbf{64.3 \pm 0.5}$	37.7 ± 1.3	49.1 ± 1.6	51.2 ± 1.2	$\mathbf{65.6 \pm 0.6}$
50	-	48.3 ± 0.3	56.7 ± 0.6	$\mathbf{67.1 \pm 0.3}$	-	51.7 ± 1.1	56.5 ± 0.7	$\mathbf{73.5 \pm 0.3}$
Full		79.8				81.3		

4.5 Computational Cost comparison

For a fair comparison, we ran both AVDD and X-AVDD on a single 32 GB Tesla V100 GPU for SPC=20 on ConvNet architecture. AVDD must contain all synthetic images during the optimization stage, whereas X-AVDD can generate images batch-wise, drastically reducing memory requirements. Table 5 compares GPU-Hour consumption for generating synthetic datasets by AVDD and X-AVDD. This significant reduction in memory consumption enables X-AVDD to scale to higher SPC settings and larger architectures.

Table 5: GPU-Hour comparison between AVDD and our proposed method for various datasets.

Method	VGGSound-10K	AVE	MUSIC-21	VGGSound
AVDD	165	471	363	5342
X-AVDD	23	51	45	628

4.6 Performance on Audio-Visual Retrieval

Audio-visual retrieval assesses the preservation of semantic correspondence between audio and visual modalities. It evaluates whether representations learned from distilled data can identify semantically corresponding samples across modalities. Audio-to-visual (A $\rightarrow$ V) retrieval measures the ability to identify relevant visual content given an audio query, whereas visual-to-audio (V $\rightarrow$ A) retrieval evaluates the reverse relationship. Performance is measured through Recall@K metrics, where K $\in \{1, 5\}$. Similar to the setup used by AVDD, we measure the class-level retrieval performance. Results are presented in Table 6.

Table 6: Comparative analysis of X-AVDD against the baseline methods for performance on the audio-visual retrieval task. Similar to AVDD, SPC=20 is chosen. Our method outperforms AVDD across all four benchmark datasets.

Direction	**Method**	**VGGSound-10K**		**AVE**		**MUSIC-21**		**VGGSound**	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
A $\rightarrow$ V	DM	8.6	47.3	6.9	32.1	16.8	37.1	0.8	3.9
	AVDD	19.3	59.3	13.1	35.0	31.7	53.0	0.9	4.1
	X-AVDD	**24.5**	**69.2**	**18.3**	**41.1**	**34.5**	**57.6**	**1.3**	**5.6**
V $\rightarrow$ A	DM	11.3	44.0	**11.0**	29.5	7.3	33.6	0.7	4.2
	AVDD	27.3	59.3	6.4	34.5	27.3	42.5	0.8	4.2
	X-AVDD	**33.9**	**66.7**	**11.0**	**40.8**	**31.2**	**45.6**	**1.4**	**5.4**

4.7 Ablation Studies

4.7.1 Impact of α on test accuracy

α is the weight assigned to the cross-entropy loss from hard labels during the training of the student network with the synthetic dataset, where the total loss is a convex combination of cross-entropy loss from hard labels and KL divergence loss from soft labels. Figure 4 shows that the combination of both hard labels and soft labels results in higher generalization accuracy while training a student model from scratch using the distilled synthetic dataset.

4.7.2 Directionality ablation

We ablate the two cross-modal EDM terms by setting the respective loss weights to zero in Eq. 8. Using either the video$\rightarrow$audio term ($\lambda_{av} = 0$) or audio$\rightarrow$video term ($\lambda_{va} = 0$) in a standalone fashion consistently underperforms the full bi-directional objective across SPC regimes. This suggests that bidirectional cross-modal constraints are complementary and important for synthesizing informative paired audio-visual data. Table 7 enumerates these observations on VGGSound-10K dataset.

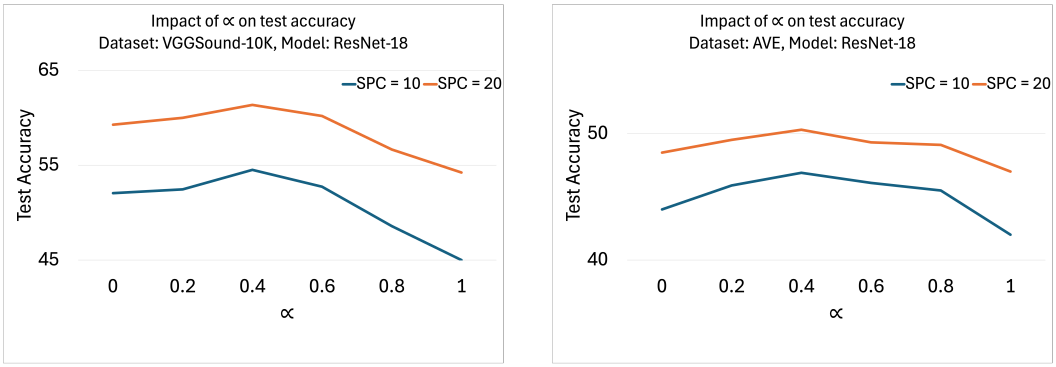

Figure 4: Impact of α on test accuracy. The left panel shows results for the VGGSound-10K dataset and the ResNet-18 architecture, while the right panel shows results for AVE and ResNet-18.

Table 7: Directionality ablation on VGGSound-10K. We toggle the EDM loss weights in Eq. 8.

Method	**SPC=1**	**SPC=10**	**SPC=20**	**SPC=50**
X-AVDD (both)	30.0 ± 1.5	54.5 ± 0.7	61.5 ± 0.8	66.7 ± 0.9
Only V→A	25.4 ± 1.3	50.7 ± 0.1	57.2 ± 0.6	63.1 ± 0.1
Only A→V	25.9 ± 1.1	53.4 ± 0.6	58.5 ± 1.2	62.0 ± 0.9

Table 8: IOR values for various SPC settings for different datasets.

Dataset	SPC=1	SPC=10	SPC=20
VGGSound-10K	9.4	6.2	7.2
Music-21	22.3	11.9	9.0
AVE	11.9	14.5	10.1

4.7.3 Improvement Over Random

DD-Ranking [Li et al., 2025b] introduced a metric called Improvement Over Random (IOR) to carry out a fair evaluation of the impact of a dataset distillation method. IOR is computed by: $\text{IOR} = \text{Acc}_{\text{syn-any}} - \text{Acc}_{\text{rdm-any}}$, where $\text{Acc}_{\text{syn-any}}$ and $\text{Acc}_{\text{rdm-any}}$ are test accuracies of a model trained on the synthetic dataset and a randomly selected subset from the original dataset, respectively. Table 8 enumerates the IOR value for each dataset across each SPC setting. Our proposed method achieves significant improvement over random selection across multiple datasets.

4.7.4 Impact of soft label on generalization accuracy

Tables 9 & 10 provide a comparative analysis between the accuracy obtained on training a student network with only hard labels vs. with a combination of hard and soft labels. We can see that training the student network with a combination of hard and soft labels improves test accuracy for both the SRe^{2}L and our proposed method.

Table 9: Impact of soft label on generalization accuracy (VGGSound-10K, ResNet-18).

SPC	Ratio (%)	SRe^{2}L (HL)	SRe^{2}L (SL)	Ours (HL)	Ours (SL)
1	0.11	16.5 ± 0.1	21.7 ± 0.6	24.5 ± 1.5	30.0 ± 1.5
10	1.13	31.4 ± 1.3	46.9 ± 2.0	44.9 ± 0.2	54.5 ± 0.7
20	2.26	36.3 ± 2.1	52.7 ± 0.5	49.3 ± 0.9	61.5 ± 0.8
50	5.65	41.8 ± 0.7	60.4 ± 0.9	52.9 ± 0.3	66.7 ± 0.9

Table 10: Impact of soft label on generalization accuracy (MUSIC-21, ResNet-18).

SPC	Ratio (%)	SRe^{2}L (HL)	Ours (HL)	SRe^{2}L (SL)	Ours (SL)
1	0.014	14.1 ± 2.7	20.5 ± 1.0	21.6 ± 0.8	30.8 ± 2.1
10	0.14	37.1 ± 2.8	51.6 ± 1.1	58.7 ± 1.4	70.9 ± 0.3
20	0.28	43.4 ± 0.5	56.1 ± 0.1	61.2 ± 2.9	74.1 ± 0.4
50	0.70	52.1 ± 0.6	61.6 ± 0.4	71.9 ± 0.6	78.4 ± 0.3

5 Conclusion

In this work, we introduced X-AVDD, the first scalable audio-visual dataset distillation approach that explicitly uses bi-directional cross-attention between the two modalities to jointly synthesize synthetic audio-visual data. By distilling data through cross-attended embeddings rather than independently matched modality features, X-AVDD preserves semantic alignment between audio and visual streams while substantially reducing memory and computational overhead. Through extensive experiments, we demonstrated the superior performance of X-AVDD across four benchmark audio-visual datasets. We also illustrated the effectiveness of our method by cross-architecture comparison with VGG-11 and ViT. The distilled data generated by X-AVDD shows impressive performance in fine-tuning frozen backbones of SOTA models, further highlighting the efficacy of our proposed methodology.

Acknowledgments and Disclosures

Aravind Reddy acknowledges a Google gift grant to IIT Madras for supporting his conference registration and travel.

References

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. doi: 10.1109/ICCV.2017.73.

Adnan El Assadi, Isaac Chung, Chenghao Xiao, Roman Solomatin, Animesh Jha, Rahul Chand, Silky Singh, Kaitlyn Wang, Ali Sartaz Khan, Marc Moussa Nasser, et al. Maeb: Massive audio embedding benchmark. *arXiv preprint arXiv:2602.16008*, 2026.

Mathilde Brousmiche, Jean Rouat, and Stéphane Dupont. Multimodal attentive fusion network for audio-visual event recognition. *Information Fusion*, 2022. doi: 10.1016/j.inffus.2022.03.001.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *CVPR*, 2022. doi: 10.1109/CVPR52688.2022.01045.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *ICASSP*, 2020. doi: 10.1109/ICASSP40776.2020.9053174.

Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *ICML*, 2023. URL https://proceedings.mlr.press/v202/cui23e.html.

Mucong Ding, Yuancheng Xu, Tahseen Rabbani, Xiaoyu Liu, Brian Gravelle, Teresa Ranadive, Tai-Ching Tuan, and Furong Huang. Calibrated dataset condensation for faster hyperparameter search, 2024. URL https://arxiv.org/abs/2405.17535.

Tian Dong, Bo Zhao, and Lingjuan Lyu. Privacy for free: How does dataset condensation help privacy? In *ICML*, 2022. URL https://proceedings.mlr.press/v162/dong22c.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Jiawei Du, Xin Zhang, Juncheng Hu, Wenxin Huang, and Joey Tianyi Zhou. Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment. In *NeurIPS*, 2024. URL https://openreview.net/forum?id=uwSaDHLlYc.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *TMLR*, 2023. URL https://openreview.net/forum?id=g97OHbQyk1.

Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Zeyad Al-Halah, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Proc. Interspeech*, 2021.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending CLIP to image, text and audio. In *ICASSP*, 2022. doi: 10.1109/ICASSP43922.2022.9747631.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. doi: 10.1109/CVPR.2016.90.

Markus Hiller, Krista A. Ehinger, and Tom Drummond. Perceiving longer sequences with bi-directional cross-attention transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Saksham Singh Kushwaha, Siva Sai Nagender Vasireddy, Kai Wang, and Yapeng Tian. Audio-visual dataset distillation. *TMLR*, 2024. URL https://openreview.net/forum?id=IJlbuSrXmk.

Yongmin Lee and Hye Won Chung. CovMatch: Cross-covariance guided multimodal dataset distillation with trainable text encoder. In *NeurIPS*, 2025. URL https://openreview.net/forum?id=2dpiR9fqUk.

Shiye Lei and Dacheng Tao. A comprehensive survey of dataset distillation. *TPAMI*, 2023. doi: 10.1109/TPAMI.2023.3322540.

Wenyuan Li, Guang Li, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Decoupled audio-visual dataset distillation, 2025a. URL https://arxiv.org/abs/2511.17890.

Zekai Li, Xinhao Zhong, Samir Khaki, Zhiyuan Liang, Yuhao Zhou, Mingjia Shi, Ziqiao Wang, Xuanlei Zhao, Wangbo Zhao, Ziheng Qin, et al. Dd-ranking: Rethinking the evaluation of dataset distillation. *arXiv:2505.13300*, 2025b.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022. URL https://openreview.net/forum?id=S7Evzt9uit3.

Ping Liu and Jiawei Du. The evolution of dataset distillation: Toward scalable and generalizable solutions. *arXiv:2502.05673*, 2025.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. doi: 10.1109/ICCV.2019.00272.

Saumyaranjan Mohanty, Aravind Reddy, and Konda Reddy Mopuri. DiRe: Diversity-promoting regularization for dataset condensation. In *WACV*, 2026. doi: 10.1109/WACV61042.2026.00276.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, 2023.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *NeurIPS*, 2019. URL https://papers.nips.cc/paper/8717-when-does-label-smoothing-help.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *ICML*, 2020. URL https://proceedings.mlr.press/v119/naeem20a.html.

R Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L. Koerich, Simon Bacon, Patrick Cardinal, and Eric Granger. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *CVPR*, 2022. doi: 10.1109/CVPRW56347.2022.00278.

Zhen Qin, Daoyuan Chen, Wenhao Zhang, Liuyi Yao, Yilun Huang, Bolin Ding, Yaliang Li, and Shuiguang Deng. The synergy between data and multi-modal large language models: A survey from co-development perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (10):8415–8434, 2025. doi: 10.1109/TPAMI.2025.3576835.

Noveen Sachdeva and Julian McAuley. Data distillation: A survey. *TMLR*, 2023. URL https://openreview.net/forum?id=lmXMXP74TO.

Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z Liu, Yuri A Lawryshyn, and Konstantinos N Plataniotis. DataDAM: Efficient dataset distillation with attention matching. In *ICCV*, 2023. doi: 10.1109/ICCV51070.2023.01568.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green AI. *Communications of the ACM*, 2020. doi: 10.1145/3381831.

X. Si, S. Zhang, Z. Yang, J. Yu, and D. Ming. A bidirectional cross-modal transformer representation learning model for eeg-fnirs multimodal affective bci. *Expert Systems with Applications*, 266: 126081, 2025.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. URL https://arxiv.org/abs/1409.1556.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *ACL*, 2019. URL https://aclanthology.org/P19-1355/.

Felipe Petroski Such, Aditya Rawal, Joel Lehman, Kenneth Stanley, and Jeffrey Clune. Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data. In *ICML*, 2020. URL https://proceedings.mlr.press/v119/such20a.html.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. doi: 10.1007/978-3-030-01216-8_16.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, 2019. doi: 10.18653/v1/P19-1656.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv:1811.10959*, 2018.

Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation. *TMLR*, 2024. URL https://openreview.net/forum?id=6L6cD65pot.

Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. Mieb: Massive image embedding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22187–22198, 2025.

Yue Xu, Zhilin Lin, Yusong Qiu, Cewu Lu, and Yong-Lu Li. Low-rank similarity mining for multimodal dataset distillation. In *ICML*, 2024. URL https://openreview.net/forum?id=mY93trX2Qz.

Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at ImageNet scale from a new perspective. In *NeurIPS*, 2023. URL https://openreview.net/forum?id=5Fgdk3hZpb.

Xin Zeng, Wanjun Zhang, Huake Wang, and Xiaoli Bian. Learning with balanced criss-cross attention for cross-modality crowd counting. In *International Conference on Information Technology and Computer Communications*, 2023. doi: 10.1145/3606843.3606852.

Xin Zhang, Ziruo Zhang, Jiawei Du, Zuozhu Liu, and Joey Tianyi Zhou. Beyond modality collapse: Representation blending for multimodal dataset distillation. In *NeurIPS*, 2025. URL https://openreview.net/forum?id=Xpi6YxNSiI.

Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *ICML*, 2021. URL https://proceedings.mlr.press/v139/zhao21a.html.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *WACV*, 2023. doi: 10.1109/WACV56688.2023.00645.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *ICLR*, 2021. URL https://openreview.net/forum?id=mSAKhLYLSsl.

Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019. doi: 10.1109/ICCV.2019.00182.

Zhenghao Zhao, Haoxuan Wang, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. Efficient multimodal dataset distillation via generative models. In *NeurIPS*, 2025. URL https://openreview.net/forum?id=Eu2Tqcvxih.

A Diversity Comparison

Following Mohanty et al. [2026], we use three different metrics for a comprehensive comparative analysis of the diversity of the generated synthetic images:

1. **Vendi Score Friedman and Dieng [2023]**, which is defined as the exponential of the Shannon entropy of the eigenvalues of a similarity matrix. A higher Vendi Score indicates greater diversity in the dataset.
2. **Cosine Similarity.** We compute intra-class pairwise cosine similarity among the embeddings of the synthetic dataset and use the mean as a measure of diversity. For a pair of embeddings $\{x, y\}$, their cosine similarity is defined as:

$$s_{cos}(x, y) = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|} \tag{9}$$

   A lower intra-class cosine similarity indicates higher diversity.

3. **Coverage Naeem et al. [2020]** measures the fraction of real samples whose neighbourhoods contain at least one synthetic sample:

$$\text{coverage} = \frac{1}{N} \sum_{i=1}^{N} 1_{\left\{\exists\, j \text{ s.t. } Y_j \in B\left(X_i, \text{NND}_k(X_i)\right)\right\}} \tag{10}$$

   where, X is the real dataset, Y is the synthetic dataset, NND_k represents k^{th} Nearest Neighbor Distance, and $B(X, r)$ represents a ball of radius r around the data point X. The distance computations here are carried out in the feature space. A higher coverage indicates greater diversity among the synthetic images.

Table 11: Comparative analysis of diversity metrics. Synthetic data generated by X-AVDD performs better across all three diversity metrics as compared to the baseline multimodal adaptation of SRe^{2}L.

Dataset	Method	Vendi ↑	Similarity ↓	Coverage ↑
VGGSound-10k	SRe^{2}L	1.26	0.89	0.05
	X-AVDD	**1.65**	**0.87**	**0.15**
Music-21	SRe^{2}L	2.13	0.81	0.03
	X-AVDD	**3.04**	**0.77**	**0.13**
VGGSound	SRe^{2}L	1.23	0.81	0.05
	X-AVDD	**1.71**	**0.77**	**0.11**

B Computing Infrastructure Information

Information regarding computing infrastructure for carrying out the experiments is provided in Table 12. An NVIDIA A100 80GB GPU is used only for testing AVDD with higher SPC settings.

Table 12: Details of the computing infrastructure.

Description	Value
GPU	32 GB Tesla V100
OS	Ubuntu 18.04.1 LTS
Python	3.11.7
PyTorch	2.4.1
NumPy	1.24.3

C Implementation of multimodal SRe^{2}L

Yin et al. [2023] introduced SRe^{2}L, a lightweight tripartite learning paradigm that can scale to larger datasets such as ImageNet-1K. The three stages of SRe^{2}L are:

- **S**queeze, in which a model is trained on the original training dataset, and the Batch-Norm (BN) layer statistics are stored.
- **Re**cover, in which the BN statistics are used to synthesize the condensed dataset.
- **Re**label, in which the synthetic dataset is assigned soft labels through the model trained on the original dataset.

In this paper, we have adapted this methodology as a baseline for distilling audio-visual datasets. In the **S**queeze stage, a teacher network is trained on the full training dataset. The BN statistics stored in various BatchNorm layers in these networks are used during the synthesis stage.

The synthesis stage (Recover) of SRe^{2}L consists of synthesizing the target data starting from independent random Gaussian noise. However, the random nature of Gaussian noise and its lack of semantic information are bottlenecks for effective information distillation. Hence, in our approach, we use real samples for initialization, as do some other dataset distillation methods [Cazenavette et al., 2022, Du et al., 2024]. For a fair comparison, the same initialization samples are used for this baseline and our proposed X-AVDD method.

D Dataset details

We use the same benchmark datasets and the corresponding preprocessing/splits from AVDD Kushwaha et al. [2024].

- **AVE** [Tian et al., 2018]: The Audio-Visual Event (AVE) dataset contains 4,143 unconstrained 10-second YouTube videos spanning 28 event categories with temporally annotated audio-visual event boundaries at second-level granularity (each video contains at least one 2-second audio-visual event). Following AVDD, we segment each clip into non-overlapping 1-second audio-visual windows aligned with the annotations, resulting in train/val/test splits of 27,726 / 3,288 / 3,305 samples.
- **MUSIC-21** [Zhao et al., 2019]: MUSIC-21 is an expanded musical performance dataset containing 1,365 untrimmed videos of musical solos and duets; the original paper reports a split of 1,065 training videos and 300 test videos, with preprocessing into video shots (5,861 shots in total). Following AVDD, we use their processed split for the *solo* subset, with train/val/test splits of 146,908 / 7,103 / 42,440 samples.
- **VGGSound** [Chen et al., 2020]: VGGSound is a large-scale audio-visual dataset of short YouTube clips curated to ensure audio-visual correspondence. We use the preprocessed train/test split provided by AVDD [Kushwaha et al., 2024], consisting of approximately 165k / 13k train/test samples across 309 classes.
- **VGGSound-10K**: We additionally use the 10-class subset of VGGSound released with AVDD [Kushwaha et al., 2024], which contains 8,808 / 444 train/test samples (using AVDD's preprocessing and split).

E Hyperparameter details

Table 13 provides the hyperparameters used during pre-training of the teacher network, audio-guided video network, and video-guided audio network. Table 14 tabulates the hyperparameters used during synthesis.

F Timing comparison

For a fair comparison, we ran both AVDD and our proposed X-AVDD on a single 32 GB Tesla V100 GPU. It may be noted that the training of the teacher network is a one-time process. The synthesis stage of our X-AVDD takes a minuscule amount of time as compared to AVDD.

Table 13: Hyperparameter details for pre-training.

Hyperparameter	Value
Epochs	100
Optimizer	Adam
Base learning rate	0.001
Learning rate schedule	cosine decay
Augmentation	RandomResizedCrop
Batch size	128

Table 14: Hyperparameter details for synthesis.

Hyperparameter	Value
Iteration	1000
Optimizer	Adam
Optimizer momentum	$\beta_1, \beta_2 = 0.5, 0.9$
Base learning rate	0.25
Learning rate schedule	cosine decay
Augmentation	RandomResizedCrop
Attn. multiplier	1.0

Table 15: Timing comparison of AVDD and our X-AVDD on VGGSound-10k dataset for synthesis of SPC=20. Kindly note that, while the synthesis stage in AVDD is a recurring process, in our proposed method, training the teacher network and storage of the embeddings are one-time requirements, and only the synthesis stage is the recurring step for generating synthetic samples across various SPC settings.

AVDD		X-AVDD	
Steps	Time	Steps	Time
Buffer creation	1 hr	teacher training & embedding storage	65 mins.
synthesis	9 hrs.	synthesis	12 mins.
Total	10 hrs.	Total	1.3 hrs.

G Modality gap on VGGSound-10K

Figure 5 shows the significant increase in modality gap by synthetic data generated through AVDD on VGGSound-10K dataset. Our proposed method is able to achieve consistently lower modality gap across all classes.

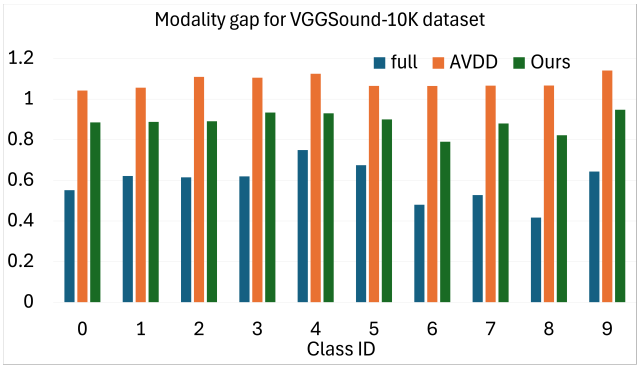

Figure 5: Comparison of modality gap observed in the real dataset, synthetic data generated by AVDD, and our proposed method. As shown, our proposed method achieves a lower modality gap consistently across all classes in the VGGSound-10K dataset.

H Impact of temperature on soft label generation

The temperature parameter τ is vital in smoothening the soft labels Müller et al. [2019]. The impact of τ on test accuracy for the VGGSound-10K dataset with the ResNet-18 architecture for SPC=10 and SPC=20 is depicted in Figure 6.

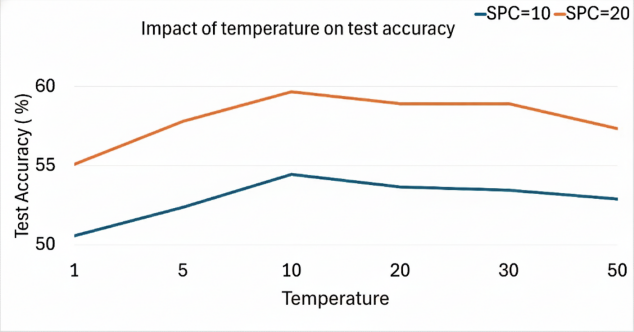

Figure 6: Impact of temperature on test accuracy. Results are provided for VGGSound-10K with ResNet-18 for SPC=10 and SPC=20. We can see that accuracy increases with temperature till it reaches a maximum and then decreases, signifying the importance of the temperature parameter during soft label generation.

I Results with 'concat' fusion method

The main paper contains results with the 'ensemble' approach of audio-visual fusion. Table 16 compares results obtained by the 'concat' fusion method with a synthetic dataset generated by the SRe^{2}L baseline and X-AVDD methods. X-AVDD consistently outperforms the baseline across all four datasets and various SPC settings.

Table 16: Test set accuracy comparison for various datasets with 'concat' fusion method.

Dataset	SPC	SRe^{2}L	X-AVDD
VGGSound-10K	1	20.8 ± 1.5	$\mathbf{26.8 \pm 0.7}$
	10	47.2 ± 0.7	$\mathbf{55.4 \pm 0.7}$
	20	49.6 ± 0.2	$\mathbf{65.1 \pm 0.2}$
	50	64.4 ± 1.2	$\mathbf{69.5 \pm 0.3}$
AVE	1	18.0 ± 0.1	$\mathbf{21.2 \pm 0.4}$
	10	41.5 ± 0.6	$\mathbf{51.6 \pm 0.4}$
	20	46.0 ± 0.7	$\mathbf{56.6 \pm 0.2}$
	50	49.2 ± 0.4	$\mathbf{61.3 \pm 0.2}$
Music-21	1	29.8 ± 1.1	$\mathbf{34.0 \pm 1.6}$
	10	66.7 ± 2.4	$\mathbf{70.8 \pm 1.2}$
	20	67.1 ± 0.5	$\mathbf{73.5 \pm 0.7}$
	50	72.1 ± 0.9	$\mathbf{79.0 \pm 0.3}$
VGGSound	1	8.1 ± 1.2	$\mathbf{11.4 \pm 0.9}$
	10	17.9 ± 0.8	$\mathbf{24.9 \pm 0.3}$
	20	22.3 ± 0.1	$\mathbf{28.3 \pm 0.2}$
	50	22.4 ± 0.6	$\mathbf{31.1 \pm 0.1}$

J Timing comparison with AudioCLIP architecture

While we compare the performance of our method with the AudioCLIP architecture against baseline methods in Table 2, Table 17 shows the time required to train a student network using distilled data vs. training on the full dataset. As can be seen, with a fraction of the total time required for the full dataset, synthetic data generated by our proposed method achieves test accuracy comparable to that of the full dataset.

K Limitations

Our work focuses on audio-visual classification and retrieval as the downstream tasks. Other tasks such as temporal localization and other modalities would be taken up as future work.

Table 17: Test accuracy and student training time with the AudioCLIP architecture on VGGSound-10K and AVE. Ratio is the size of the distilled dataset as a percentage of the full dataset. At SPC=50, training on X-AVDD data takes about $9\times$ less time on VGGSound-10K (2.74 vs. 25.30 hours) and about $11\times$ less time on AVE (6.43 vs. 72.12 hours), while reaching 86% and 95% of the full-data accuracy, respectively.

	VGGSound-10K			AVE		
SPC	Ratio (%)	Accuracy (%)	Train time (Hrs.)	Ratio (%)	Accuracy (%)	Train time (Hrs.)
1	0.11	24.8 ± 0.6	0.28	0.1	44.0 ± 1.2	0.43
10	1.13	33.7 ± 0.3	0.56	1.0	74.5 ± 0.9	1.31
20	2.26	46.8 ± 0.9	1.16	2.0	76.8 ± 1.1	2.66
50	5.65	62.3 ± 0.8	2.74	5.0	81.3 ± 0.9	6.43
Full	100	72.4 ± 0.7	25.30	100	85.6 ± 1.1	72.12

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contribution and scope. The paper introduces a bi-directional cross-attention-based decoupled distillation method for audio-visual datasets. The claims are validated through extensive experimentation presented in the main body of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the approach in the Appendix K.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed not to penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results with detailed algorithm pseudocode. The paper also provides a link to the GitHub code base for ready reference.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides a link to an anonymized version of the codebase for ready reference. All datasets used in the experiments are open-source benchmark datasets and are properly cited in the main paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the details necessary to reproduce the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have conducted experiments rigorously with multiple iterations for each settings and have reported mean and standard deviations for each experimental trial.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details are provided in appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research aligns with NeurIPS code of ethics and relevant guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our research does not have any direct societal impacts to discuss.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not involve the release of data and models with high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper provides proper citations and references to original creators.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The methodology introduced is well-documented, and the accompanying codebase is also provided.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowd-sourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve any research with human subjects and hence does not require any IRB approvals.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The research has not used any LLM.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.