ImageBindDC: Compressing Multi-Modal Data with ImageBind-Based Condensation

Yue Min[1,2,3*], **Shaobo Wang**[1,4*], **Jiaze Li**[1], **Tianle Niu**[1], **Junxin Fan**[1], **Yongliang Miao**[1], **Lijin Yang**[2†], **Linfeng Zhang**[1†]
[1]EPIC Lab, Shanghai Jiao Tong University
[2]Bosch Corporate Research Asia Pacific
[3]Hong Kong University of Science and Technology
[4]Alibaba Group

Abstract

Data condensation techniques aim to synthesize a compact dataset from a larger one to enable efficient model training, yet while successful in unimodal settings, they often fail in multimodal scenarios where preserving intricate intermodal dependencies is crucial. To address this, we introduce ImageBindDC, a novel data condensation framework operating within the unified feature space of ImageBind. Our approach moves beyond conventional distribution-matching by employing a powerful Characteristic Function (CF) loss, which operates in the Fourier domain to facilitate a more precise statistical alignment via exact infinite moment matching. We design our objective to enforce three critical levels of distributional consistency: (i) uni-modal alignment, which matches the statistical properties of synthetic and real data within each modality; (ii) cross-modal alignment, which preserves pairwise semantics by matching the distributions of hybrid real-synthetic data pairs; and (iii) joint-modal alignment, which captures the complete multivariate data structure by aligning the joint distribution of real data pairs with their synthetic counterparts. Extensive experiments highlight the effectiveness of ImageBindDC: on the NYU-v2 dataset, a model trained on just 5 condensed datapoints per class achieves lossless performance comparable to one trained on the full dataset, achieving a new state-of-the-art with an 8.2% absolute improvement over the previous best method and more than 4× less condensation time.

Introduction

The remarkable success of modern AI has been largely propelled by the synergy between large-scale models and vast datasets (Brown et al. 2020; Dosovitskiy et al. 2020; Grattafiori et al. 2024; Yang et al. 2025). However, this paradigm comes at a significant cost: the computational, storage, and financial burdens associated with training on massive datasets are becoming prohibitive. Dataset Condensation (DC) has emerged as a compelling solution to this challenge (Wang et al. 2018; Sachdeva and McAuley 2023). The goal of DC is to synthesize a small, synthetic dataset that, while a fraction of the original size, preserves its essential information, enabling models to be trained to high performance with dramatically reduced resources.

Early successes in this domain have demonstrated the feasibility of condensing thousands of images into just a handful, accelerating training by orders of magnitude (Zhao, Mopuri, and Bilen 2020; Zhao and Bilen 2021; Cazenavette et al. 2022a; Guo et al. 2023; Wang et al. 2025d). While effective in uni-modal settings (*e.g.*, images-only), these traditional DC methods falter in the face of the increasingly prevalent multi-modal world. Modern applications frequently leverage rich, interconnected data from various sources, such as image, audio, text, and depth (Radford et al. 2021; Girdhar et al. 2023). The key challenge in condens-

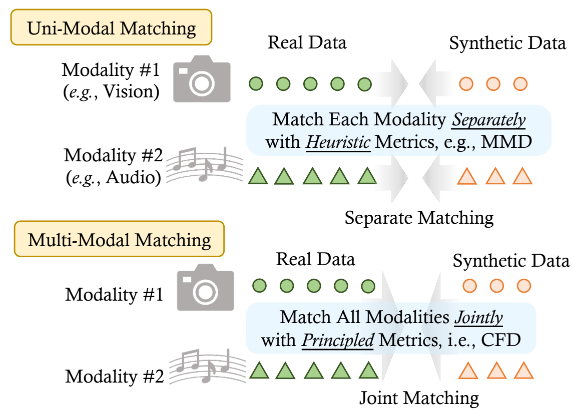

Figure 1: A Comparison of multi-modal Data Condensation Paradigms. **(Top) Separate Matching:** Conventional methods condense each modality (*e.g.*, vision, audio) independently, often using heuristic metrics like MMD. This preserves uni-modal statistics but critically fails to capture the cross-modal relationships that link the data together. **(Bottom) Joint Matching:** Our proposed framework, ImageBindDC, performs joint matching of all modalities simultaneously within a unified feature space. By using a principled metric like Characteristic Function Distance (CFD), our approach preserves the complete multi-modal data structure, ensuring the synthesized data is semantically coherent.

*These authors contributed equally.
†Corresponding authors.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

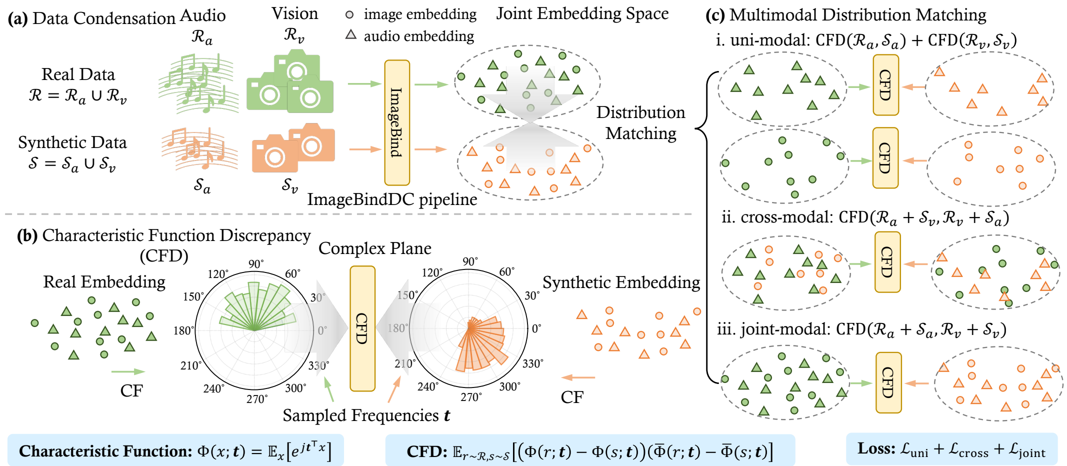

Figure 2: Overview of the **ImageBindDC** framework. We compress large multi-modal datasets (*e.g.*, vision + audio) into tiny synthetic versions while preserving their structure. (a) Condensation pipeline: Real data ($\mathcal{R}_v, \mathcal{R}_a$) and synthetic data ($\mathcal{S}_v, \mathcal{S}_a$) are encoded into a shared space using ImageBind. We optimize the synthetic data to match the real data's distribution in this space. (b) Characteristic function discrepancy (CFD): We measure distribution similarity using CFD, capturing all statistical patterns for precise alignment. (c) Three alignment goals: Our loss combines three CFD-based terms: (i) *Uni-modal*: Match vision-to-vision and audio-to-audio distributions. (ii) *Cross-modal*: Match relationships between modalities. (iii) *Joint-modal*: Match entire paired distributions (real pairs vs. synthetic pairs).

ing such data is not merely to preserve the statistical properties within each modality independently, but to maintain the intricate *cross-modal relationships* that encode semantic meaning. As shown in Figure 1, existing condensation techniques, designed with a uni-modal perspective, are ill-equipped for this task; they may create a statistically representative set of synthetic data, usually little datapoints per class (DPC), but they break the vital link that pairs a specific image with its corresponding sound.

To bridge this critical gap, we introduce **ImageBindDC**, a novel framework for multi-modal data condensation. Our core insight is to perform condensation not in the raw, disparate data spaces, but within a **unified, joint-embedding space** provided by a large-scale pretrained model like ImageBind (Girdhar et al. 2023). This allows us to directly address inter-modal dependencies. Furthermore, we move beyond conventional distribution matching techniques by employing a powerful **Characteristic Function (CF) loss** (Wang et al. 2025d). Operating in the Fourier domain, our CF-based objective enables a more precise statistical alignment by matching an infinite number of moments between the synthetic and real data distributions. As illustrated in Figure 2, our approach enforces distributional consistency at three critical levels: uni-modal, cross-modal, and joint-modal, ensuring that the synthesized data captures the complete multi-modal structure. Our main contributions are summarized as follows:

- We propose **ImageBindDC**, the first data condensation framework specifically designed to operate in a unified feature space, effectively preserving complex multi-modal data relationships.
- We design a novel, multi-faceted objective that leverages a Characteristic Function (CF) based loss to enforce three critical levels of statistical consistency: **uni-modal**, **cross-modal**, and **joint-modal alignment**, which together preserve the complete multi-modal data structure.
- Extensive experiments demonstrate the state-of-the-art performance of ImageBindDC, achieving the best results across various multimodal datasets. Notably, a model trained on just 20 synthesized image-audio pairs reached 98% of the full-dataset performance in Audio-Visual Event Localization classification, marking a 2.53% improvement. Also, on NYU-v2 Dataset, ImageBindDC achieved 8.2% absolute improvement over the method second in performance. Our method also shows exceptional computational efficiency, reducing dataset condension time by over 4.6× for 20 DPC.

Related Work

Uni-modal Data Condensation. Uni-modal condensation methods fall into two main categories. *Data Selection* (or coreset selection) identifies a representative subset of the original data based on various criteria like clustering (Alexey et al. 2016; Bautista et al. 2016), greedy utility maximization (Wei, Iyer, and Bilmes 2015; Soper 2021), gradient information (Paul, Ganguli, and Dziugaite 2021; Mirzasoleiman, Bilmes, and Leskovec 2020; Killamsetty et al. 2021; Xia et al. 2024), or other model-based metrics (Toneva et al. 2018; Wang et al. 2025a,c,b). However, these methods are fundamentally limited by being confined to the original data sources and cannot yield unseen data. In contrast, *Dataset Distillation* (Wang et al. 2018) synthesizes a small set of new, optimized data points. The goal is to match the learning dynamics of the full dataset. Dominant strategies include *Gradient Matching*, which aligns training gradients (Zhao, Mopuri, and Bilen 2020; Lee et al. 2022; Zhao and Bilen 2021; Wang et al. 2025e,f) or entire parameter trajectories (Guo et al. 2023; Cazenavette et al. 2022a), and *Distribution Matching*, which aligns feature statistics in a pretrained embedding space (Zhao and Bilen 2022; Zhao et al. 2023; Wang et al. 2025d). Our work builds upon this data synthesis paradigm, adapting it for the unique challenges of the multi-modal setting.

Multi-modal Data Condensation. The primary challenge in multi-modal data condensation is not only preserving the intra-modal statistics but, more importantly, the cross-modal semantic relationships. A few recent methods have begun to tackle this problem. Although recent advancements in data selection, such as LLM-based filtering (Chen et al. 2023; Liu et al. 2023; Xu et al. 2023), gradient-based influence estimation (Attendu and Corbeil 2023), and self-instruction generation (Kung et al. 2023), have shown promise in optimizing instruction tuning, their effectiveness can be inconsistent, with some studies indicating they often fail to consistently outperform random sampling. AVDD (Kushwaha et al. 2024) performs condensation for audio-visual data but does so by matching distributions in separate, modality-specific feature spaces, which risks misalignment and relies on now-outdated backbone architectures. To address cross-modal relationships, LoRS (Xu et al. 2024) proposes matching a pre-computed ground-truth similarity matrix between modalities, using low-rank factorization for efficiency. However, boiling down the complex, high-dimensional relationship between modalities to a single scalar similarity may be an oversimplification that fails to capture the full distributional structure. More recently, RepBlend (Zhang et al. 2025) introduced representation blending to encourage diversity and prevent modality collapse, but this remains a heuristic approach that can be difficult to balance and may not guarantee the preservation of joint-modal semantics.

Methodology

Problem Formulation

Dataset Condensation aims to obtain a small, information-rich dataset $\mathcal{S}$ that acts as an efficient substitute for a large real dataset $\mathcal{R}$. Normally, $\mathcal{S}$ consists of a rather small DPC. Formally, given a real dataset $\mathcal{R} = \{(x_i, y_i)\}_{i=1}^{N}$, the goal is to generate a much smaller (synthetic) dataset $\mathcal{S} = \{(\tilde{x}_j, y_j)\}_{j=1}^{M}$ where $M \ll N$. While many works frame this as a complex bi-level optimization problem involving unrolled model training (Cazenavette et al. 2022a; Zhao and Bilen 2021), a more efficient and increasingly popular paradigm is to perform *distribution matching* in a fixed feature space. This approach avoids the expensive inner-loop model training. Let ψ be a feature extractor, which remains frozen during condensation, mapping an arbitrary input x into $e_x = \psi(x)$. The optimization problem simplifies to a single-level objective:

$$\min_{\mathcal{S}} \mathbf{D}_{x\sim\mathcal{R},\tilde{x}\sim\mathcal{S}}\big(e_x, e_{\tilde{x}}\big), \tag{1}$$

where x and $\tilde{x}$ denote the sets of feature embeddings for the real and synthetic data, and $\mathbf{D}(\cdot,\cdot)$ is a distributional distance metric. The core challenge, which we address, is to design an effective metric $\mathbf{D}$ for the multi-modal setting.

Distribution Matching

A common choice for the distance metric $\mathbf{D}$ is the Maximum Mean Discrepancy (MMD) (Zhao and Bilen 2022; Zhao et al. 2023). MMD measures the distance between two distributions, real data distribution $P_{\mathcal{R}}$ and synthetic data distribution $P_{\mathcal{S}}$, by mapping them into a empirical kernel space and computing the distance between their mean embeddings. While widely used, the effectiveness of MMD is highly dependent on the choice of the kernel, which is often heuristic (*e.g.*, a Gaussian kernel). A poorly chosen kernel may fail to capture all statistical differences between the distributions, leading to suboptimal alignment.

To address the limitations of kernel-based methods, we adopt a more principled metric founded on the Characteristic Function (CF) (Wang et al. 2025d; Li et al. 2023). The CF of a random vector $z \in \mathcal{Z}$ is the Fourier transform of its probability density function (PDF), which uniquely defines its probability distribution. It is defined as:

$$\Phi(z;t) = \mathbb{E}_{z\sim P_{\mathcal{Z}}}\left[e^{jt^\top z}\right], \tag{2}$$

where $j = \sqrt{-1}$ is the imaginary unit and $t \in \mathbb{R}^{\dim(\mathcal{Z})}$ is a frequency vector. By Lévy's Uniqueness Theorem, two distributions are identical if and only if their characteristic functions are identical. This allows us to define the *Characteristic Function Discrepancy* (CFD) as the squared L_2-distance between the CFs of the real and synthetic feature distributions, $P_{\mathcal{R}}$ and $P_{\mathcal{S}}$:

$$\begin{aligned}
&\mathrm{CFD}(x,\tilde{x}) = \big(\Phi(x;t) - \Phi(\tilde{x};t)\big)\big(\bar{\Phi}(x;t) - \bar{\Phi}(\tilde{x};t)\big) \\
&= |\Phi(x;t)|^2 + |\Phi(\tilde{x};t)|^2 \\
&- |\Phi(x;t)||\Phi(\tilde{x};t)|\,(2\cos(a_x(t) - a_{\tilde{x}}(t)) \\
&= |\Phi(x;t) - \Phi(\tilde{x};t)|^2 \\
&+ 2|\Phi(x;t)||\Phi(\tilde{x};t)|\,(1-\cos(a_x(t) - a_{\tilde{x}}(t)))\,,
\end{aligned} \tag{3}$$

where the expectation is taken over a distribution of random frequency vectors t. Empirically, we approximate the CFs and the expectation with samples. Unlike MMD, CFD does not depend on a user-defined kernel and provides a more robust framework for matching distributions by comparing all of their statistical moments implicitly in the Fourier domain (Wang et al. 2025d).

ImageBindDC Framework

We now introduce our multi-modal data condensation framework, ImageBindDC, as illustrated in Figure 2. Without loss of generality, we take *audio and vision* multi-modal data as an example to illusate ImageBindDC framework. The goal of ImageBindDC is to learn a small multi-modal synthetic dataset, $\mathcal{S} = \mathcal{S}_a \cup \mathcal{S}_v$, that efficiently represents a much larger real dataset, $\mathcal{R} = \mathcal{R}_a \cup \mathcal{R}_v$. As shown in Figure 2(a), our method leverages a pretrained ImageBind encoder, which provides a unified embedding space for different modalities. Both the real data instance (x_a, x_v), and the synthetic data instance ($\tilde{x}_a$, $\tilde{x}_v$), are passed through the ImageBindDC pipeline to obtain their respective embeddings, (e_a, e_v) and $(\tilde{e}_a, \tilde{e}_v)$, within this common space.

The core of ImageBindDC is to minimize the discrepancy between the distribution of real embeddings and synthetic embeddings for all modalities, as shown in Figure 2(b). By aligning the distributions in this shared space, ImageBindDC ensures that the condensed synthetic data captures the essential statistical characteristics of the original, large-scale multi-modal dataset.

Multi-modal Modeling In our approach, we leverage ImageBind (Girdhar et al. 2023), which is designed to map multi-modal data into a shared feature space. The core idea behind ImageBind is to create a unified embedding space where different types of data can be represented as points in the same space, facilitating effective multi-modal learning. ImageBind operates by learning a set of embeddings for each modality, such as images and audio, and binding them into a shared feature space. It ensures that each modality is represented in a way that reflects its unique characteristics while also aligning the modalities within a shared space.

Specifically, let e_a and e_v represent the embeddings for audio and image data, respectively. These embeddings are learned such that they can be projected into a feature space $\mathcal{F}$. The objective of the ImageBind is to minimize the distance between corresponding audio and image embeddings within the shared space. This is typically achieved by optimizing a loss function that ensures that similar data points from different modalities are aligned in the feature space. The embedding process for each modality is formulated as:

$$e_a = \mathcal{E}_a(x_a), \quad e_v = \mathcal{E}_v(x_v), \tag{4}$$

where $\mathcal{E}_a$ and $\mathcal{E}_v$ are the encoding functions for audio and image data, respectively, and x_a and x_v are the input audio and image data points.

Uni-modal Alignment In the context of uni-modal distribution matching, we aim to minimize the discrepancy between the embeddings of real and synthetic data for each modality separately. This is achieved by employing the CFD in Eq. (3), which is designed to align the distribution of embeddings from real data with that of synthetic data in the shared feature space.

Specifically, for audio data, let $e_{a,\text{real}}$ and e_a denote the embeddings of real and synthetic audio data, respectively. Similarly, for image data, let $e_{v,\text{real}}$ and $e_{v,\text{syn}}$ represent the embeddings of real and synthetic images, respectively. The objective of Uni-modal Distribution Matching is to minimize the embedding discrepancy for each modality, as shown in Figure 2(c)i. Specifically, we define the CF loss for the audio modality as follows:

$$\mathcal{L}_{\text{audio}} = \text{CFD}(e_a, \tilde{e}_a). \tag{5}$$

The CF loss computes the difference between the characteristic functions of the real and synthetic audio embeddings, measured in terms of the CFD. Similarly, for image data:

$$\mathcal{L}_{\text{image}} = \text{CFD}(e_v, \tilde{e}_v). \tag{6}$$

The total loss for the Uni-modal Distribution Matching across both modalities is the sum of the individual losses:

$$\mathcal{L}_{\text{uni}} = \mathcal{L}_{\text{audio}} + \mathcal{L}_{\text{image}}. \tag{7}$$

By minimizing this loss, we ensure that the embeddings of real and synthetic data for both audio and image modalities are aligned in the feature space, facilitating better distribution matching guaranteed by CFD's principled properties.

Cross-modal Alignment In the Cross-modal Alignment (CMA), we aim to measure the alignment between the real audio and image embeddings, as well as between the synthetic audio and image embeddings. Specifically, we perform element-wise multiplication between the embeddings of real audio and image data, and similarly for synthetic audio and image embeddings. The resulting values are then used to compute the cosine similarity between the real and synthetic data embeddings.

Specifically, let e_a and e_v represent the real audio and image embeddings, respectively, while $\tilde{e}_a$ and $\tilde{e}_v$ represent the synthetic audio and image embeddings, respectively. The element-wise multiplication of the embeddings for real audio and image data is defined as:

$$e_a \odot e_v = [e_{a,1} \cdot e_{v,1}, \ldots, e_{a,N} \cdot e_{v,N}], \tag{8}$$

where $\odot$ denotes element-wise multiplication and N is the dimension of the embedding vectors. Similarly, for synthetic audio and image data, the element-wise multiplication is defined as:

$$\tilde{e}_a \odot \tilde{e}_v = [\tilde{e}_{a,1} \cdot \tilde{e}_{v\,1}, \ldots, \tilde{e}_{a,N} \cdot \tilde{e}_{v\,N}]. \tag{9}$$

Next, we compute the cosine similarity between the element-wise multiplied real and synthetic embeddings. The cosine similarity for the real embeddings is given by:

$$\rho_{\text{cross}} = \frac{\langle e_a \odot e_v, \tilde{e}_a \odot \tilde{e}_v \rangle}{\|e_a \odot e_v\|_2 \|\tilde{e}_a \odot \tilde{e}_v\|_2}, \tag{10}$$

where $\langle \cdot, \cdot \rangle$ represents the dot product and $\| \cdot \|_2$ denotes the L2 norm. The cosine similarity computes the alignment between the element-wise multiplied real and synthetic embeddings, with higher values indicating better alignment. The final Cross-modal Alignment (CMA) loss is then calculated by minimizing the cosine similarity for both real and synthetic data:

$$\mathcal{L}_{\text{cross}} = 1 - \rho_{\text{cross}}. \tag{11}$$

By minimizing this loss, the model learns to align the real and synthetic audio-image embeddings, facilitating better cross-modal matching.

Joint-Modal Alignment In the Joint-Modal Alignment (JMA) part, we focus on minimizing the gap between the average embeddings of real and synthetic data across different modalities. The key idea is to compute the mean embedding for each modality (audio and image) and then perform a matrix multiplication between the real and synthetic embeddings to calculate a cross-modal similarity score.

Specifically, let e_a, $\tilde{e}_a$, e_v, and $\tilde{e}_{v,\text{syn}}$ represent the embeddings for real audio, synthetic audio, real image, and synthetic image data, respectively. First, we compute the mean of the embeddings along each dimension. The mean embeddings for each modality (audio and image) are denoted by E_a, $\tilde{E}_a$, E_v, and $\tilde{E}_v$. Next, we reshape the average embeddings into 2D matrices for matrix multiplication, where each average embedding is transformed into a row vector. Then, we compute the cross-modal similarity score by performing a matrix multiplication between the element-wise multiplied average embeddings:

$$\rho_{\text{joint}} = E_a \odot \tilde{E}_v^\top \times E_v \odot \tilde{E}_a^\top \tag{12}$$

Dataset	DPC	Ratio %	Selection-based: Random	Herding	Forgetting	GraNd	Distillation-based: DC	DSA	MTT	DM	AVDD	**ImageBindDC**	Whole data
VGGS-10K	1	0.11	15.44±1.87	20.77±2.11	23.41±1.31	15.42±0.42	18.28±1.36	19.32±1.35	34.13±3.6	36.54±2.52	40.41±1.81	**42.66±1.48**	68.24±0.75
	10	1.13	32.01±1.64	39.89±1.64	40.78±2.04	34.95±1.52	32.10±0.84	36.61±1.04	36.79±1.97	43.85±1.75	48.08±0.92	**55.23±0.13**	
	20	2.26	45.1±2.31	50.2±0.74	52.16±0.49	49.22±1.22	-	-	51.87±1.26	49.01±2.44	48.86±1.53	**55.30±0.18**	
AVE	1	0.1	10.07±1.16	11.84±0.4	10.07±0.45	8.69±0.42	10.45±0.39	10.76±0.62	12.13±0.41	16.70±1.46	16.90±0.14	**18.08±0.52**	52.20±0.38
	10	1	13.64±0.22	21.94±0.52	20.31±0.38	19.54±0.35	22.04±1.04	20.92±1.00	23.15±0.95	26.14±1.80	32.90±0.14	**34.42±0.32**	
	20	2	26.32±1.01	33.04±0.38	31.17±0.49	29.27±0.51	-	-	-	32.57±0.97	36.67±0.49	**38.41±0.07**	

Table 1: Classification accuracy (%) for data condensation on VGGS-10K and AVE under different DPC settings. All methods use a randomly initialized ConvNet to condense, with accuracy measured by training that same ConvNet from scratch.

Dataset	DPC	Ratio (%)	Selection-based: Random	Herding	Forgetting	GraNd	Distillation-based: DC	DSA	MTT	DM	AVDD	**ImageBindDC**	Whole data
VGGS-10K	1	0.11	25.24±3.25	27.14±1.58	29.34±1.49	25.33±2.03	OOM	28.87±1.50	OOM	31.55±1.34	36.10±1.67	**37.25±1.62**	60.57±0.05
	10	1.13	40.86±1.32	43.15±1.28	44.67±1.07	41.06±1.13	OOM	44.73±1.36	OOM	45.46±1.27	46.61±0.59	**48.89±0.78**	
	20	2.26	50.18±1.66	51.92±0.98	52.98±0.79	50.77±1.09	OOM	52.72±1.37	OOM	53.17±0.89	53.87±1.14	**56.11±0.98**	
AVE	1	0.1	45.11±3.17	45.77±1.75	43.23±1.88	40.62±2.08	OOM	44.46±2.16	OOM	65.35±2.06	67.32±1.55	**70.10±1.24**	76.93±0.14
	10	1	64.56±1.88	66.76±1.24	65.25±1.45	62.06±1.19	OOM	58.49±5.05	OOM	69.49±0.69	71.33±0.35	**73.67±0.31**	
	20	2	67.01±1.96	69.90±0.47	68.75±0.64	65.66±0.76	OOM	71.49±0.51	OOM	70.50±0.29	72.81±0.24	**75.34±0.27**	

Table 2: Classification accuracy (%) for data condensation methods on VGGS-10K and AVE under different DPC settings. The condensation for all methods is guided by a pretrained ImageBind model, with accuracy measured by training that same model from scratch on the condensed data. Note that OOM indicates that the method ran out of memory on 2 A100 GPUs.

DPC Ratio (%)	1 0.034	2 6.79	5 16.98	10 33.96
Random	60.60±2.43	69.22±1.87	73.85±1.42	88.38±3.05
DM	67.97±11.69	75.25±12.92	89.08±4.96	96.89±1.26
AVDD	72.22±10.03	81.30±2.84	95.92±1.63	98.62±0.45
ImageBindDC	**80.43±0.44**	**88.33±4.99**	**97.30±1.17**	**98.73±1.04**
Whole Data	98.62±0.25			

Table 3: Accuracy (%) on the NYU-v2 dataset for a depth-text classification task. The text modality is derived from the scene name (*e.g.*, “bathroom”). For this experiment, a randomly initialized ImageBind model guides the condensation, and performance is evaluated by training the same model architecture from scratch on the condensed data.

where $\times$ denotes the matrix multiplication operation. This step computes the similarity between the real and synthetic embeddings, considering the alignment between both modalities. Finally, the JMA loss is calculated as the mean of the joint-modal gap, which is given by:

$$\mathcal{L}_{\text{joint}} = 1 - \rho_{\text{joint}}, \tag{13}$$

This loss encourages the model to minimize the gap between real and synthetic multi-modal embeddings, thereby enhancing the alignment between the two modalities.

Put All Components Together The final loss function is obtained by combining the above mentioned three distinct losses: $\mathcal{L}_{\text{uni}}$, $\mathcal{L}_{\text{cross}}$, and $\mathcal{L}_{\text{joint}}$, each corresponding to different levels of alignment between the real and synthetic data distributions. These losses are scaled by respective hyperparameters λ_{uni}, λ_{cross}, and λ_{joint}, which control their contribution to the overall objective. The final loss is defined as:

$$\mathcal{L} = \lambda_{\text{uni}}\mathcal{L}_{\text{uni}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}} + \lambda_{\text{joint}}\mathcal{L}_{\text{joint}}. \tag{14}$$

The detailed pseudocode of the overall framework including input and output is provided in Appendix B.

Metric	A2T R@1	A2T R@5	A2T R@10	T2A R@1	T2A R@5	T2A R@10
DM	0.0268	0.1014	0.1761	0.0306	0.1148	0. 1684
AVDD	0.0316	0.1024	0.1866	0.0402	0.133	0.2048
ImageBindDC	**0.0362**	**0.1297**	**0.1995**	**0.0464**	**0.1567**	**0.2281**
Whole Data	0.0526	0.1627	0.2364	0.0565	0.1674	0.2488

Table 4: Audio-text retrieval performance (Recall@K) on the Clotho dataset. All methods distill features from the pretrained ImageBind space at a setting of 20 DPC. Performance is evaluated by training an ImageBind model, which comprises a frozen backbone and a trainable linear head.

Experiments

Experimental Settings

Audio-visual Tasks For audio-visual tasks, following (Zhao and Bilen 2023; Kushwaha et al. 2024; Cazenavette et al. 2022b), we evaluated all distillation methods using two widely used audio-visual datasets: VGGS-10K and AVE. VGGSound (Chen et al. 2020) is a large-scale audio-visual dataset containing approximately 200k YouTube videos across 309 classes. VGGS-10K(Simonyan and Zisserman 2014) is a subset of VGGSound. AVE (Tian et al. 2018) consists of 4,143 video clips spanning 28 event categories.

Text-image Tasks We utilized NYU-v2 (Nathan Silberman and Fergus 2012) dataset. The text modality contains the name of the scene from which the corresponding image was captured. This dataset includes RGB and depth images.

Audio-text Tasks Experiments were conducted on Clotho (Drossos, Lipping, and Virtanen 2020), a large-scale audio captioning dataset designed for training and evaluating models that generate captions for audio clips.

Condensation Ratio (%)	1 DPC 0.1		10 DPC 1	
Model	ConvNet	ImageBind	ConvNet	ImageBind
Random	10.07±1.16	45.11±3.17	13.64±0.22	64.56±1.88
MTT	OOM	OOM	OOM	OOM
DM	7.10±1.36	60.51±1.12	11.61±0.88	69.26±1.67
AVDD	4.85±1.11	67.32±1.55	12.44±1.29	71.33±0.35
ImageBindDC	**12.69±1.23**	**70.10±1.24**	**16.74±0.18**	**73.67±0.31**
Whole	52.20±0.38	76.93±0.14	52.20±0.38	76.93±0.14

Table 5: Cross-architecture accuracy (%) on the AVE dataset. A pretrained ImageBind model is used to guide the distillation of synthetic datasets, which are then evaluated by training different architectures from scratch.

Metric	Time (s)			GPU mem. (GB)		
DPC	1	10	20	1	10	20
DSA	208.48	508.36	712.01	9.13	9.36	14.15
DM	140.30	611.15	707.21	8.96	9.34	14.24
MTT	OOM	OOM	OOM	OOM	OOM	OOM
DC	OOM	OOM	OOM	OOM	OOM	OOM
AVDD	158.11	507.08	700.10	8.96	9.34	14.24
ImageBindDC	**57.46**	**89.40**	**123.74**	**5.60**	**8.14**	**13.39**

Table 6: Computational efficiency on VGGS-10K. "OOM" means out-of-memory on a single A100 GPU.

Baselines For selection-based methods, we utilized Random, Herding (Welling 2009), Forgetting (Toneva et al. 2018), GraNd (Paul, Ganguli, and Dziugaite 2021). For distillation-based methods, we included DC (Zhao, Mopuri, and Bilen 2020), DSA (Zhao and Bilen 2021), MTT (Cazenavette et al. 2022a), DM (Zhao and Bilen 2022) and AVDD (Kushwaha et al. 2024). More details of experimental settings are provided in Appendix A.

Models We employed two distinct architectures for encoding the inputs: ImageBind and a ConvNet architecture. ImageBind (Girdhar et al. 2023) is used to map both audio and image data into a shared feature space.

Implementation Details Experiments were all conducted on 4 NVIDIA A100 GPUs. For AVE and VGGS-10K, we used a learning rate of 0.2 and an SGD optimizer with a momentum of 0.5. NYU-v2, we used a learning rate of 0.001 for both the backbone and the classifier, with the same optimizer. To ensure fairness, all experiments were conducted for 3 times. More details are included in Appendix A.

Main Results

Comparison with Baselines Our ImageBindDC approach consistently outperforms other methods across all datasets and settings. As shown in Table 1 and 2, ImageBindDC achieves the highest classification accuracy on both VGGS-10K and AVE datasets, under different DPC settings, when using both ConvNet and ImageBind models for distillation. For instance, on VGGS-10K with 10 DPC and using the ImageBind model, ImageBindDC achieves an accuracy of 55.23%, a significant improvement over the next best method, AVDD (48.08%). Similarly, on the NYU-v2 dataset shown in Table 3, ImageBindDC demonstrates superior performance, achieving 98.73% accuracy with 10 DPC, while the baseline DM method reaches 96.89%. Furthermore, as shown in Table 4 ImageBindDC also have outstanding performance on audio-text retrieval.

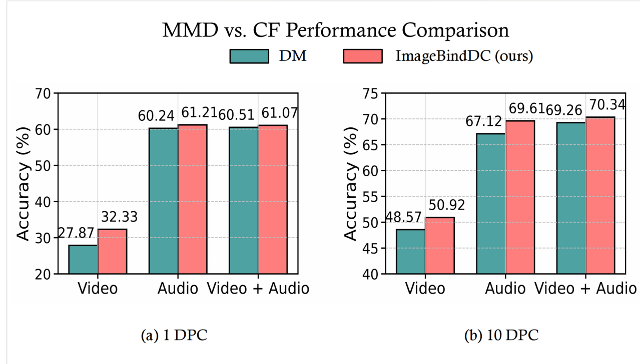

Figure 3: DM (MMD) vs. ImageBindDC (CF) Performance Comparison. The figure illustrates the accuracy of the MMD and CF methods under (a) 1 DPC and (b) 10 DPC.

Cross-Architecture Performance We tested the performance of the data distilled by our methods on unseen models. As shown in Table 5, a pretrained ImageBind model is used to guide the distillation of synthetic datasets, which are then evaluated by training different architectures from scratch, *e.g.*, Convnets and ImageBind. ImageBindDC consistently outperforms other methods. For example, with 10 DPC, ImageBindDC achieves 14.42% accuracy on ConvNet and 73.67% on ImageBind, largely surpassing previous SOTA (12.44% on ConvNet and 71.33% on ImageBind).

Ablation Study

Effectiveness of Characteristic Function Discrepancy (CFD) To validate our choice of a Characteristic Function Discrepancy (CFD), we compared its performance against the Maximum Mean Discrepancy (MMD) loss used by the DM baseline. Results in Figure 3 reveal a consistent advantage for our CF-based approach across all tested configurations. For uni-modal audio task with 1 DPC, ImageBindDC achieves 32.33% accuracy, a significant lead over the 27.87% from DM. This trend continues in the 10 DPC setting, where for the combined Video-Audio task, our method scores 70.34% compared to DM's 69.26%. While the improvements vary, the consistent superiority of our method in every scenario confirms that the CF loss provides a more robust and precise alignment, making it a better-suited metric for the complexities of data condensation.

Impact of Alignment Components To dissect the contribution of each alignment component in our proposed loss, we conducted further experiments, which are presented in Figure 4. The findings reveal that while preserving *uni-modal* statistics provides a strong baseline, it is the combination of all three objectives that un-

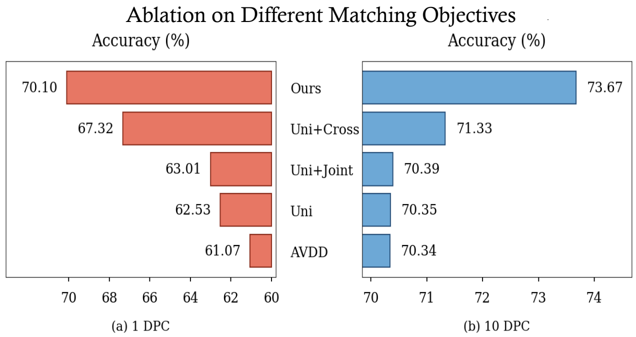

Figure 4: Ablation on Different Matching Objectives. This figure illustrates the contribution of uni-modal, joint-modal, and cross-modal matching objectives to overall accuracy. Results are presented for both (a) 1 DPC and (b) 10 DPC.

locks the full potential of the model. For instance, at 10 DPC, the *uni-modal* loss alone achieves a respectable 70.34% accuracy. Adding either the *joint-modal* or *cross-modal* objectives individually yields gains compared with the baseline. When all three components are combined, performance jumps significantly to 73.67%, an absolute improvement of 3.33%. This highlights a critical insight: *alignment objectives are not merely additive but work in synergy*. The *uni-modal* loss preserves the integrity of each modality, while the *cross-modal* and *joint-modal* losses enforce the relational structure between them. Both are essential for creating a high-quality condensed dataset.

Computational Efficiency Analysis We evaluated the computational efficiency of ImageBindDC against other baselines, with results presented in Table 6. Our method is not only faster than competing condensation techniques but also dramatically reduces the resources required. For instance, at the 1 DPC setting, ImageBindDC is over 2.4× faster than DM (57.46s vs. 140.3s) and reduces GPU memory usage by a significant 37.5% (5.6 GB vs. 8.96 GB). More critically, when compared to a single training epoch on the full dataset, the benefits are even more pronounced. Even at 20 DPC, our condensation process is over 3.4× faster than a single full-data epoch (123.74s vs. 419.9s) while slashing memory requirements by over 75% (from 55.29 GB to 13.39 GB). This remarkable efficiency makes training large models feasible on resource-constrained systems.

Discussion

Visualization of distilled data Figure 5 provides a qualitative assessment of our method by visually comparing image samples from ImageBindDC, the AVDD baseline, and the original ground truth on the NYU-v2 dataset. The distilled images from ImageBindDC clearly demonstrate superior visual coherence and semantic integrity, with key features of scenes like 'bathroom' and 'bedroom' remaining easily recognizable. This stands in stark contrast to the baseline's samples, which often degrade into distorted or abstract amalgamations that fail to capture the defining characteristics of the original data. This qualitative advantage is a

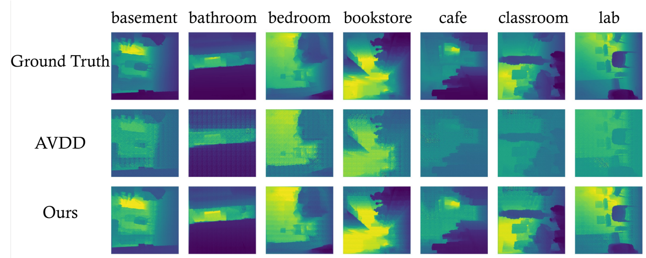

Figure 5: Qualitative comparison of distilled image samples on NYU-v2 dataset.

direct result of our core methodology; by performing condensation within a unified, joint-embedding space, ImageBindDC effectively preserves the intricate cross-modal relationships between visual data and their corresponding textual or depth information. This process prevents the semantic decoupling that plagues methods that condense modalities independently, yielding a synthetic dataset that is not only compact but also rich in high-fidelity features.

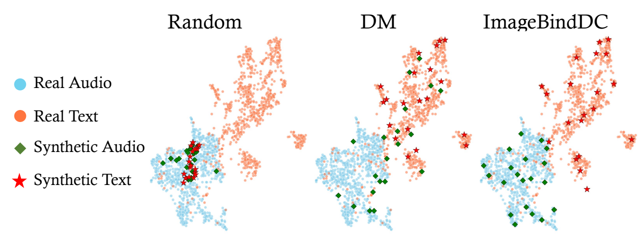

Figure 6: Umap of real and synthetic embeddings on Clotho dataset with audio and text modalities.

Visualization of synthetic embeddings Figure 6 shows 2D UMAP projections of Clotho embeddings from real data and synthetic data generated by three methods: Random, DM, and ImageBindDC. The Random baseline produces a small, concentrated cluster that fails to reflect the diversity of the real distribution, while DM achieves a broader spread but still stays largely separated from real embeddings. In contrast, ImageBindDC yields synthetic audio–text embeddings that are well aligned with the real embedding clouds, indicating a much closer match to the true data distribution.

Conclusion

In this work, we introduced ImageBindDC, a novel framework that significantly advances multi-modal data condensation. Our approach overcomes the limitations of existing methods by operating in a unified feature space and employing a powerful Characteristic Function Discrepancy for more precise distribution matching. We designed a multi-faceted objective to enforce uni-modal, cross-modal, and joint-modal consistency, ensuring that the intricate relationships between modalities are preserved. Experiments confirm the effectiveness of ImageBindDC, demonstrating its ability to generate high-quality synthetic data.

Acknowledgments

This work was supported by Alibaba Group through Alibaba Innovative Research Program, and Shanghai Science and Technology Program (Grant No. 25ZR1402278).

References

Alexey, D.; Fischer, P.; Tobias, J.; Springenberg, M. R.; and Brox, T. 2016. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE TPAMI*, 38(9): 1734–1747.

Attendu, J.-M.; and Corbeil, J.-P. 2023. Nlu on data diets: Dynamic data subset selection for nlp classification tasks. *arXiv preprint arXiv:2306.03208*.

Bautista, M. A.; Sanakoyeu, A.; Tikhoncheva, E.; and Ommer, B. 2016. Cliquecnn: Deep unsupervised exemplar learning. *Advances in Neural Information Processing Systems*, 29.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cazenavette, G.; Wang, T.; Torralba, A.; Efros, A. A.; and Zhu, J.-Y. 2022a. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4750–4759.

Cazenavette, G.; Wang, T.; Torralba, A.; Efros, A. A.; and Zhu, J.-Y. 2022b. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4750–4759.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725. IEEE.

Chen, L.; Li, S.; Yan, J.; Wang, H.; Gunaratna, K.; Yadav, V.; Tang, Z.; Srinivasan, V.; Zhou, T.; Huang, H.; et al. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Drossos, K.; Lipping, S.; and Virtanen, T. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 736–740. IEEE.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15180–15190.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guo, Z.; Wang, K.; Cazenavette, G.; Li, H.; Zhang, K.; and You, Y. 2023. Towards lossless dataset distillation via difficulty-aligned trajectory matching. *arXiv preprint arXiv:2310.05773*.

Killamsetty, K.; Durga, S.; Ramakrishnan, G.; De, A.; and Iyer, R. 2021. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, 5464–5474. PMLR.

Kung, P.-N.; Yin, F.; Wu, D.; Chang, K.-W.; and Peng, N. 2023. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. *arXiv preprint arXiv:2311.00288*.

Kushwaha, S. S.; Vasireddy, S. S. N.; Wang, K.; and Tian, Y. 2024. Audio-Visual Dataset Distillation. *Transactions on Machine Learning Research*.

Lee, S.; Chun, S.; Jung, S.; Yun, S.; and Yoon, S. 2022. Dataset condensation with contrastive signals. In *International Conference on Machine Learning*, 12352–12364. PMLR.

Li, S.; Zhang, J.; Li, Y.; Xu, M.; Deng, X.; and Li, L. 2023. Neural Characteristic Function Learning for Conditional Image Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7204–7214.

Liu, W.; Zeng, W.; He, K.; Jiang, Y.; and He, J. 2023. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Mirzasoleiman, B.; Bilmes, J.; and Leskovec, J. 2020. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, 6950–6960. PMLR.

Nathan Silberman, P. K., Derek Hoiem; and Fergus, R. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*.

Paul, M.; Ganguli, S.; and Dziugaite, G. K. 2021. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34: 20596–20607.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Sachdeva, N.; and McAuley, J. 2023. Data distillation: A survey. *arXiv preprint arXiv:2301.04272*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Soper, D. S. 2021. Greed is good: rapid hyperparameter optimization and model selection using greedy k-fold cross validation. *Electronics*, 10(16): 1973.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, 247–263.

Toneva, M.; Sordoni, A.; Combes, R. T. d.; Trischler, A.; Bengio, Y.; and Gordon, G. J. 2018. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*.

Wang, S.; Jin, X.; Wang, Z.; Wang, J.; Zhang, J.; Li, K.; Wen, Z.; Li, Z.; He, C.; Hu, X.; and Zhang, L. 2025a. Data Whisperer: Efficient Data Selection for Task-Specific LLM Fine-Tuning via Few-Shot In-Context Learning. *Annual Meeting of the Association for Computational Linguistics*.

Wang, S.; Miao, Y.; Liu, Y.; Liao, N.; Zhang, L.; et al. 2025b. CircuitSeer: Mining High-Quality Data by Probing Mathematical Reasoning Circuits in LLMs. *arXiv preprint arXiv:2510.18470*.

Wang, S.; Wang, J.; Zhang, J.; Wang, C.; Min, Y.; Wen, Z.; Huang, F.; Jiang, H.; Lin, J.; Liu, D.; et al. 2025c. Winning the pruning gamble: A unified approach to joint sample and token pruning for efficient supervised fine-tuning. *arXiv preprint arXiv:2509.23873*.

Wang, S.; Yang, Y.; Liu, Z.; Sun, C.; Hu, X.; He, C.; and Zhang, L. 2025d. Dataset distillation with neural characteristic function: A minmax perspective. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25570–25580.

Wang, S.; Yang, Y.; Wang, Q.; Li, K.; Zhang, L.; and Yan, J. 2025e. Not All Samples Should Be Utilized Equally: Towards Understanding and Improving Dataset Distillation. *Synthetic Data for Computer Vision Workshop at CVPR*.

Wang, S.; Yang, Y.; Zhang, S.; Sun, C.; Li, W.; Hu, X.; and Zhang, L. 2025f. DRUPI: Dataset Reduction Using Privileged Information. In *The Future of Machine Learning Data Practices and Repositories at ICLR 2025*.

Wang, T.; Zhu, J.-Y.; Torralba, A.; and Efros, A. A. 2018. Dataset distillation. *arXiv preprint arXiv:1811.10959*.

Wei, K.; Iyer, R.; and Bilmes, J. 2015. Submodularity in data subset selection and active learning. In *International conference on machine learning*, 1954–1963. PMLR.

Welling, M. 2009. Herding dynamical weights to learn. In *Proceedings of the 26th annual international conference on machine learning*, 1121–1128.

Xia, M.; Malladi, S.; Gururangan, S.; Arora, S.; and Chen, D. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Xu, Y.; Lin, Z.; Qiu, Y.; Lu, C.; and Li, Y.-L. 2024. Low-rank similarity mining for multimodal dataset distillation. *arXiv preprint arXiv:2406.03793*.

Xu, Y.; Yao, Y.; Huang, Y.; Qi, M.; Wang, M.; Gu, B.; and Sundaresan, N. 2023. Rethinking the instruction quality: Lift is what you need. *arXiv preprint arXiv:2312.11508*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhang, X.; Zhang, Z.; Du, J.; Liu, Z.; and Zhou, J. T. 2025. Beyond Modality Collapse: Representations Blending for Multimodal Dataset Distillation. *arXiv preprint arXiv:2505.14705*.

Zhao, B.; and Bilen, H. 2021. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, 12674–12685. PMLR.

Zhao, B.; and Bilen, H. 2022. Dataset Condensation with Distribution Matching. arXiv:2110.04181.

Zhao, B.; and Bilen, H. 2023. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6514–6523.

Zhao, B.; Mopuri, K. R.; and Bilen, H. 2020. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*.

Zhao, G.; Li, G.; Qin, Y.; and Yu, Y. 2023. Improved distribution matching for dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7856–7865.