OMNIMODAL DATASET DISTILLATION VIA HIGH-ORDER PROXY ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Dataset distillation compresses large-scale datasets into compact synthetic sets while preserving training performance, but existing methods are largely restricted to single-modal or bimodal settings. Extending dataset distillation to scenarios involving more than two modalities, *i.e.*, Omnimodal Dataset Distillation, remains underexplored and challenging due to increased heterogeneity and complex cross-modal interactions. In this work, we propose HoPA, a unified framework for omnimodal dataset distillation, which captures high-order cross-modal alignments via a compact proxy. By abstracting omnimodal alignment with a shared similarity structure, our method avoids the combinatorial complexity of pairwise modality modeling and enables scalable joint distillation across heterogeneous modalities. Theoretical analysis from the spectral perspective reveals the rationality of our proposed method. Extensive experiments on various benchmarks demonstrate that the proposed method achieves superior compression–performance trade-offs compared to existing competitors. For reproducibility, the source code is now available at https://anonymous.4open.science/r/HoPA-7C5E/.

1 INTRODUCTION

While the abundance of large-scale data has been a cornerstone of recent AI breakthroughs (Kaplan et al., 2020; Team et al., 2023; 2025; Achiam et al., 2023; Bahri et al., 2024), it is increasingly recognized that modern datasets often harbor a high degree of informational redundancy (Sorscher et al., 2022; Lin et al., 2024; Lee et al., 2024; Xia et al., 2024a; Zhang et al., 2025a). Dataset Distillation, which has emerged as a particularly promising direction, condenses a large-scale dataset into a compact synthetic subset that preserves the critical information required for effective model training (Wang et al., 2018; Zhao et al., 2020; Yu et al., 2025). Unlike traditional pruning techniques (Huang et al., 2024; He et al., 2024; Zhang et al., 2023a), dataset distillation typically leverages optimization to generate synthetic data, achieving extremely high data compression while maintaining competitive performance, particularly in the context of large models and massive datasets (Geng et al., 2023; Liu & Du, 2025; Li et al., 2026b).

Most existing works of dataset distillation are limited to single-modal data, including vision (Ding et al., 2025; Choi et al., 2026; Dong et al., 2025; Shao et al., 2024), text (Nguyen et al., 2025; Lu et al., 2025; Shen et al., 2025), and etc. More recently, a small number of studies have begun to extend vanilla dataset distillation to multimodal settings (Xu et al., 2024; Zhao et al., 2025; Zhang et al., 2025b), typically involving *two* modalities, *i.e.*, vision and text. These works adopt trajectory matching to optimize a compact synthetic set by minimizing the discrepancy between a student model's parameter update path and a pre-recorded teacher trajectory derived from real data. Meanwhile, cross-modal similarity is highlighted as a vital abstraction within this paradigm, transitioning dataset distillation from simple data compression to the preservation of inter-modal alignments. Despite the promising results, these studies are still confined to bimodal settings. Their specification on pairwise alignment limits the extension to complex scenarios involving more than two modalities in a unified and holistic manner. To this end, we formalize this frontier as Omnimodal Dataset Distillation, which necessitates the preservation across an arbitrary number of modalities.

Distilling datasets that involve more than two modalities is inherently challenging. As the number of modalities increases, the heterogeneity of feature spaces, statistical properties, and semantic granularity across modalities becomes more pronounced (Cicchetti et al., 2025; Liu et al., 2026), making

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

it difficult to construct a compact synthetic set that simultaneously preserves modality-specific information and cross-modal dependencies. Moreover, high-order interactions among multiple modalities introduce combinatorial complexity, which substantially complicates alignment, optimization, and consistency constraints during distillation. These factors collectively render dataset distillation beyond two modalities a non-trivial and largely underexplored problem, calling for the advent of Omnimodal Dataset Distillation techniques.

In this paper, to address this research gap, we propose a unified omnimodal dataset distillation method via **H**igh-**o**rder **P**roxy **A**lignment (**HoPA**) that explicitly targets datasets with more than two modalities. Our analysis of trajectory matching highlights the role of the inner training objective, motivating an objective that jointly models alignment among modalities within each instance and relations between instances. To this end, we construct a compact proxy that captures high-order cross-modal relationships among all modalities. This proxy is derived from multimodal similarity reconstruction by the spectral analysis. We jointly optimize the synthetic data and the similarity targets between these proxies through trajectory matching, allowing all modality branches to share a compact distilled set. Moreover, this design avoids the combinatorial explosion induced by enumerating all modality pairs, enabling efficient and scalable distillation across heterogeneous modalities while preserving both modality-specific information and global cross-modal consistency. We further provide a theoretical analysis of relational fidelity and local proxy stability to support the proposed design. Empirically, performance comparison against the state-of-the-art (SOTA) baselines across diverse omnimodal datasets and data ratios demonstrates superior efficacy and efficiency. Generalization is validated across different architectures, while ablation studies and further analysis reveal the rationality of our proposed method. Our contributions are summarized as follows:

- Conceptually, we identify and study, for the first time, the problem of omnimodal dataset distillation for data involving more than two modalities, highlighting a critical limitation of existing single-modal and bimodal dataset distillation paradigms.
- Technically, we propose HoPA, a unified omnimodal dataset distillation method that captures high-order cross-modal relationships via a compact proxy, enabling scalable and effective distillation without relying on exhaustive pairwise modality interactions. We also provide theoretical analysis for the proxy design.
- Empirically, we conduct extensive experiments on omnimodal benchmarks to demonstrate the effectiveness and scalability of the proposed method. It consistently achieves better compression-performance trade-offs than existing baselines. Detailed ablations and case studies are also provided.

2 PRELIMINARY

Omnimodal similarity. Given an input instance with multiple modalities, we denote by $E_m(\cdot)$ the encoder for modality m and by $\boldsymbol{z}_m = E_m(\boldsymbol{x}_m) \in \mathbb{R}^d$ the corresponding representation. Following common practice, all representations are ℓ_2-normalized such that $\|\boldsymbol{z}_m\|_2 = 1$. For a bimodal pair (m_1, m_2), the similarity is measured by the cosine similarity, *i.e.*, $s(\boldsymbol{x}_{m_1}, \boldsymbol{x}_{m_2}) = \langle \boldsymbol{z}_{m_1}, \boldsymbol{z}_{m_2} \rangle$, which characterizes the alignment strength between the two modalities for the same instance. The bimodal similarity can be extended to the omnimodal setting by jointly considering all modalities associated with a single instance. Let $\mathcal{M} = \{m_1, \ldots, m_k\}$ denote the modality set and $\boldsymbol{z} = [\boldsymbol{z}_{m_1}, \ldots, \boldsymbol{z}_{m_k}]^\top \in \mathbb{R}^{k \times d}$. The pairwise similarities among modalities can be compactly represented by a Gram matrix, formally represented as: $\boldsymbol{G} = \boldsymbol{z}\boldsymbol{z}^\top = [\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle]_{i,j \in \mathcal{M}}$. The Gram matrix encodes the complete correlation structure of all modalities within one instance and serves as a unified descriptor of omnimodal alignment.

Trajectory matching. A common paradigm in dataset distillation is to learn a compact synthetic set $\mathcal{D}_{\text{syn}}$ whose induced training dynamics match those induced by the real data (Liu et al., 2024; Cazenavette et al., 2022; Lee & Chung, 2024). An expert model is first trained on real data under an inner objective $\mathcal{L}_{\text{inner}}$, and store its parameter trajectory in an expert buffer. During distillation, a teacher segment is sampled from the buffer, where $\boldsymbol{\theta}_0$ denotes the start point, and $\boldsymbol{\theta}_T$ denotes the target parameter after T training steps on real data. The student model is then initialized from the same start point $\boldsymbol{\theta}_0$ and optimized on the synthetic set $\mathcal{D}_{\text{syn}}$ for t steps, yielding a student endpoint $\boldsymbol{\theta}_e^t$. The synthetic set is learned by minimizing the discrepancy between the student endpoint and the teacher target: $\mathcal{L}_{\text{traj}} = \|\boldsymbol{\theta}_e^t - \boldsymbol{\theta}_T\|_2^2 / \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_T\|_2^2$, $\mathcal{D}_{\text{syn}} = \arg\min_{\mathcal{D}_{\text{syn}}} \mathcal{L}_{\text{traj}}$. This objective encourages

the synthetic data to reproduce the local training trajectory of the real data on the student model. It provides a generic optimization method that can be instantiated with different distillation objectives.

Trajectory matching for bimodal distillation. Existing trajectory-based methods typically optimize the two modality branches jointly using the same synthetic set, while measuring the trajectory discrepancy separately for each encoder branch. For two modalities m_a and m_b, the objective can be written as $\mathcal{L}_{\text{bi-traj}} = \sum_{m\in\{m_a,m_b\}} \left(\|\boldsymbol{\theta}^t_{e,m} - \boldsymbol{\theta}_{T,m}\|_2^2 / \|\boldsymbol{\theta}_{0,m} - \boldsymbol{\theta}_{T,m}\|_2^2\right)$, where m indexes modality branches, and each term measures the normalized trajectory mismatch between student and expert endpoints for that modality.

3 METHODOLOGY

In this section, we first motivate the design of the inner objective, then we study omnimodal dataset distillation with similarity learning, where synthetic data are jointly optimized with a instance-wise similarity matrix. Optimizing this targets leaves open how the corresponding predictions should be computed jointly from multiple modalities. We address this design question with HoPA, which constructs a spectral proxy for each instance and uses proxy similarities in synthetic training for trajectory matching. We further analyze how well proxy similarities approximate pairwise similarities, how this approximation affects the synthetic training loss, and the local stability of the proxy.

3.1 INNER OBJECTIVES DESIGN FOR OMNIMODAL DATASET DISTILLATION

The inner objective of trajectory matching. Trajectory matching provides a generic optimization method for dataset distillation. Prior work has mainly focused on improving trajectory construction and matching efficiency (Du et al., 2023; Liu et al., 2024). However, the inner training objective $\mathcal{L}$ also serves as a key factor.

Denote the gradients of $\mathcal{L}$ by $\mathbf{g}_{\mathrm{T}}(\boldsymbol{\theta})$, $\mathbf{g}_{\mathrm{S}}(\boldsymbol{\theta})$ evaluated on the real data and synthetic data, respectively. We consider the gradient-descent rollout, $\boldsymbol{\theta}^a_{r+1} = \boldsymbol{\theta}^a_r - \eta \mathbf{g}_a(\boldsymbol{\theta}^a_r), a \in \{\mathrm{T}, \mathrm{S}\}$, with $\eta > 0$ and shared initialization. Assume $\mathbf{g}_T$ is locally L-Lipschitz along the student rollout, i.e., $\|\mathbf{g}_{\mathrm{T}}(\boldsymbol{\theta}^{\mathrm{S}}_r) - \mathbf{g}_{\mathrm{T}}(\boldsymbol{\theta}^{\mathrm{T}}_r)\| \le L\|\boldsymbol{\theta}^{\mathrm{S}}_r - \boldsymbol{\theta}^T_r\|, \forall r$. Then

$$\|\boldsymbol{\theta}^{\mathrm{S}}_n - \boldsymbol{\theta}^{\mathrm{T}}_n\| \le \eta \sum_{r=0}^{n-1} (1+\eta L)^{n-1-r} \Delta_r, \tag{1}$$

where $\Delta_r = \|\mathbf{g}_{\mathrm{S}}(\boldsymbol{\theta}^{\mathrm{S}}_r) - \mathbf{g}_{\mathrm{T}}(\boldsymbol{\theta}^{\mathrm{S}}_r)\|_2$ measures gradient mismatch at the same parameter state (See the full derivation in Appendix B.1). This dependence on gradient mismatch motivates designing the inner objective together with the synthetic data. For omnimodal distillation, this calls for a training signal that jointly supervises all modality branches within a shared synthetic set.

Omnimodal similarity modeling. For omnimodal distillation, we seek an inner objective that models both alignment among modalities within an instance and relations between instances. Existing distillation methods further motivate learnable supervision alongside synthetic inputs, including soft class labels (Guo et al., 2024) and image-text similarity matrix (Xu et al., 2024). Following this principle, we jointly optimize $\mathcal{D}_{\text{syn}}$ and a similarity matrix $\boldsymbol{S}_e = [s_{ij}]$, where s_{ij} serves as the similarity target for two complete instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

However, learning $\boldsymbol{S}_e$ does not specify how the corresponding similarity $\widetilde{s}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ should be computed from multiple modalities. Different modality pairs can yield different similarities for the same instance pair. The target s_{ij} can be applied to these predictions separately or compared with a joint prediction constructed from the modality representations. We therefore seek an instance-level similarity that jointly uses the modality representations of both instances.

3.2 HOPA: HIGH-ORDER PROXY ALIGNMENT FOR OMNIMODAL DATASET DISTILLATION

Motivated by the above analysis, we propose HoPA, which moves beyond pairwise constraints by treating the set of all modalities as a unified whole. To operationalize this, we introduce a compact proxy derived from the optimal approximation of the Gram matrix. This proxy can preserve the structural alignment of multimodal similarities and is compatible with trajectory matching. This al-

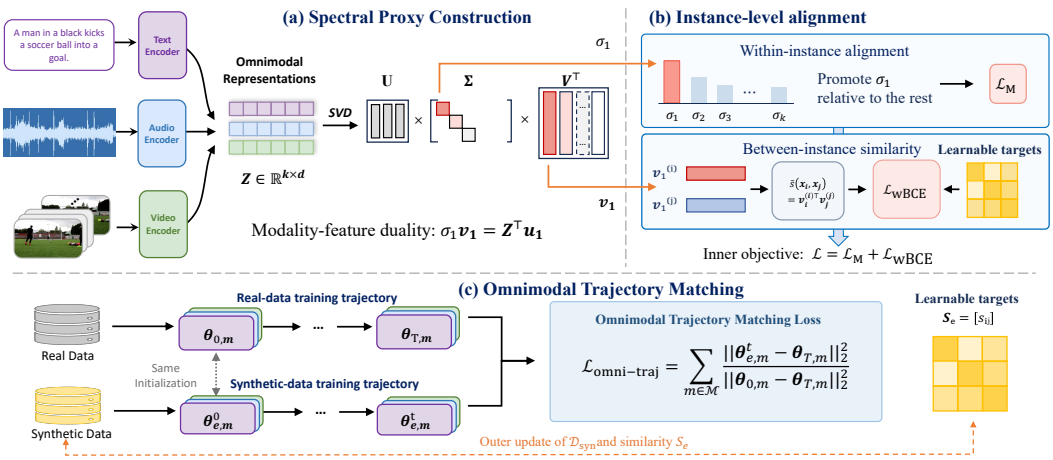

Figure 1: **The overall framework of omnimodal dataset distillation beyond bimodal data.** We decompose omnimodal representations via SVD, rather than combining pairwise modeling, to handle the increased heterogeneity among modalities (*a*). In this way, we maximize the σ_1 and meanwhile adopt the principal right singular vector $\boldsymbol{v}_1$ as a compact proxy to guide alignment (*b*). Finally, we distill the dataset into the synthetic set by matching the training trajectories. A shared similarity matrix $\boldsymbol{S}_e$ is jointly learned to preserve the multimodal alignment within the distilled data (*c*).

lows for minimal adjustments when extending bimodal trajectory matching to a multimodal context without incurring the computational cost of full matching. The full procedure is in Appendix C.4.

Multimodal similarity reconstruction. The Gram matrix $\boldsymbol{G}$ encodes the pairwise correlations between all modalities, admitting a spectral decomposition $\boldsymbol{G} = \boldsymbol{z}\boldsymbol{z}^\top = \sum_{i=1}^{k} \sigma_i^2 \boldsymbol{u}_i \boldsymbol{u}_i^\top$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ are the singular values of $\boldsymbol{z}$, and $\boldsymbol{u}_i \in \mathbb{R}^k$ are the corresponding eigenvectors, *i.e.*, the left singular vectors of $\boldsymbol{z}$.

Lemma 1 (Rank-1 optimal approximation (Liu et al., 2026)). *Let $\boldsymbol{G} \in \mathbb{R}^{k\times k}$ be the Gram matrix of the omnimodal representations. The rank-1 matrix $\tilde{\boldsymbol{G}} = \sigma_1^2 \boldsymbol{u}_1 \boldsymbol{u}_1^\top$ is the unique optimal approximation of $\boldsymbol{G}$ in terms of the Frobenius norm when $\sigma_1 \neq \sigma_2$, satisfying:*

$$\tilde{\boldsymbol{G}} = \underset{\mathbf{A}\in\mathbb{R}^{k\times k}, rank(\mathbf{A})=1}{\arg\min} \|\boldsymbol{G} - \mathbf{A}\|_F. \tag{2}$$

Remark. Accordingly, the optimal rank-1 approximation under the Frobenius norm is given by $\sigma_1^2 \boldsymbol{u}_1 \boldsymbol{u}_1^\top$. When representations from different modalities are well aligned, the leading eigenvalue dominates, and the reconstruction error becomes negligible, *i.e.*, $\|\boldsymbol{G} - \sigma_1^2 \boldsymbol{u}_1 \boldsymbol{u}_1^\top\|_F \to 0$.

Hinged with the omnimodal distillation setting, we propose to extract the omnimodal alignment holistically, rather than the direct distillation from multiple pair-wise similarities. Therefore, we derive the following lemma.

Lemma 2 (Modality-feature duality). *Let $\boldsymbol{v}_1 \in \mathbb{R}^d$ be the principal right singular vector of $\boldsymbol{z}$. The relationship between the modality alignment vector $\boldsymbol{u}_1$ and the unified semantic feature $\boldsymbol{v}_1$ is given by: $\sigma_1 \boldsymbol{v}_1 = \boldsymbol{z}^\top \boldsymbol{u}_1 = \sum_{m\in\mathcal{M}} [\boldsymbol{u}_1]_m \boldsymbol{z}_m$.*

Remark. This duality implies that $\boldsymbol{u}_1$ acts as an optimal weight vector that aggregates diverse modality representations. $\boldsymbol{v}_1$ serves as a weighted aggregation of all modality representations $\{\boldsymbol{z}_m\}$. Unlike simple averaging, this spectral aggregation automatically assigns higher importance to modalities that are more consistent with the dominant semantic direction. To this end, we propose using the proxy for the dataset distillation, adopting one global scale for the multimodal similarity rather than a full Gram matrix, yielding omnimodal trajectory matching as well.

(i) Eigenvector as proxy. Rather than operating on the full Gram matrix, we adopt the leading eigenvector as a low-dimensional proxy for omnimodal similarity. And we derive the following proposition. Let $\boldsymbol{v}_1$ denote the right singular vector associated with σ_1 and the nonnegative unit

principal eigenvector $\boldsymbol{u}_1$ of $\boldsymbol{G}$. For two instances indexed by i and j, we approximate their omnimodal similarity by

$$\tilde{s}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{v}_1^{(i)\top} \boldsymbol{v}_1^{(j)}. \tag{3}$$

This approximation also compresses the original $\mathcal{O}(k^2)$ modality interactions into a single scalar similarity while preserving the dominant alignment structure. As a result, large multimodal similarity tensors are reduced to a tractable instance-level similarity matrix.

(ii) Omnimodal trajectory matching. Using the eigenvector-based proxy, we match the training trajectory induced by synthetic data to that of real data. Specifically, we construct a proxy similarity matrix $\boldsymbol{v}_1^{(i)\top} \boldsymbol{v}_1^{(j)}$ for different instances i, j and enforce consistency between the optimization dynamics driven by synthetic and real samples. This trajectory matching principle encourages that the distilled data induces gradient updates aligned with those of the original multimodal dataset. We formulate multimodal distillation as a bilevel optimization problem. Let $\boldsymbol{\theta}_T$ denote the model parameters trained on real data and $\boldsymbol{\theta}_e$ those trained on synthetic data $(\mathcal{D}_{\text{syn}}, \boldsymbol{S}_e)$, where $\boldsymbol{S}_e$ is the synthetic similarity matrix that follows the same structure as the eigenvector-based proxy. We retain the branch-wise trajectory matching used in bimodal dataset distillation, and extend it to the omnimodal setting over all modality branches.

$$\mathcal{L}_{\text{omni-traj}} = \sum_{m \in \mathcal{M}} \frac{\|\boldsymbol{\theta}_{e,m}^t - \boldsymbol{\theta}_{T,m}\|_2^2}{\|\boldsymbol{\theta}_{0,m} - \boldsymbol{\theta}_{T,m}\|_2^2}, \qquad (\mathcal{D}_{\text{syn}}, \boldsymbol{S}_e) = \arg\min_{\mathcal{D}_{\text{syn}}, \boldsymbol{S}_e} \mathcal{L}_{\text{omni-traj}}, \tag{4}$$

Model training is driven by a combined objective consisting of modality-level contrastive learning and instance-level alignment regularization:

$$\mathcal{L} = \mathcal{L}_{\text{M}} + \mathcal{L}_{\text{wBCE}}(\boldsymbol{S}_e), \qquad \mathcal{L}_{\text{M}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\sigma_1^{(i)}/\tau)}{\sum_\ell \exp(\sigma_\ell^{(i)}/\tau)}. \tag{5}$$

The weighted binary cross-entropy loss $\mathcal{L}_{\text{wBCE}}$ enforces consistency between $\boldsymbol{S}_e$ and the proxy similarities derived from $\boldsymbol{v}_1$, thereby aligning instance-level gradients across real and synthetic data. Specifically, we follow (Xu et al., 2024) to implement $\mathcal{L}_{\text{wBCE}}$, but instead use the proposed proxies:

$$\mathcal{L}_{\text{wBCE}}(\boldsymbol{S}_e) = \sum_{i,j} w_{ij}\, \ell\Big(s_{ij}, \text{sigm}\Big(\boldsymbol{v}_1^{(i)\top} \boldsymbol{v}_1^{(j)}/\tau'\Big)\Big), \tag{6}$$

where $\ell(y, \hat{y}) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$, $w_{ij} = 1/N_+$ for $s_{ij} > \beta$ and $1/N_-$ otherwise, with N_+ and N_- denoting the numbers of positive and negative pairs, respectively. β is the positive/negative threshold and set to 0.5, s_{ij} is the learnable similarity of the i-th instance and j-th instance. Through this gradient-based matching, the distilled data preserves the essential omnimodal alignment behavior of the original dataset.

3.3 Relational Fidelity and Local Stability

We analyze how replacing pairwise modality similarities with proxy changes the loss, and how the proxy responds to small representation perturbations.

Define the spectral residual $\rho_i := 1/k \sum_{\ell=2}^{k} (\sigma_\ell^{(i)})^2 = 1 - (\sigma_1^{(i)})^2/k$. The residual measures the fraction of spectral energy outside the leading direction. The spectral loss $\mathcal{L}_{\text{M}}$ attains its minimum exactly when $\rho_i = 0$, and approaching this minimum forces $\rho_i \to 0$. This connects the spectral alignment objective to the approximation and sensitivity bounds below.

Relational fidelity. We bound the error of replacing pairwise modality similarities by $\tilde{s}(\boldsymbol{x}_i, \boldsymbol{x}_j)$, then bound its effect on the loss with the same $\boldsymbol{S}_e$.

Theorem 3 (Proxy similarity approximation)**.** *Following (Cicchetti et al., 2025; Liu et al., 2026), we assume that the similarity scores between modalities are nonnegative, and therefore $\boldsymbol{G}$ is entrywise nonnegative. For any two instances i, j,*

$$\frac{1}{k^2} \sum_{m,n \in \mathcal{M}} \Big| \langle \boldsymbol{z}_m^{(i)}, \boldsymbol{z}_n^{(j)} \rangle - \tilde{s}(\boldsymbol{x}_i, \boldsymbol{x}_j) \Big| \le \sqrt{2\rho_i} + \sqrt{2\rho_j} \tag{7}$$

Its implication for the wBCE loss follows by comparing the proxy and all-pairs constructions with the same learned targets. Let $\mathcal{L}_{\text{all}}(\boldsymbol{S}_e)$ apply the same scalar loss and targets to all modality-pair similarities and average over the k^2 modality pairs. The theorem implies

$$|\mathcal{L}_{\text{all}}(\boldsymbol{S}_e) - \mathcal{L}_{\text{wBCE}}(\boldsymbol{S}_e)| \le C\sqrt{\rho_{\max}}, \qquad \rho_{\max} = \max_i \rho_i, \tag{8}$$

where C depends on the fixed targets, weights, and temperature. Full proof in Appendix B.4.

Remark. Small spectral residuals and directional agreement allow one proxy similarity to summarize the original relations with small average error. This provides a conditional reason for the shared instane relations used in Sec. 3.2.

Local stability. HoPA use the spectral proxy to approximate the relation, which brings to concerns about the spectral stability. We examine whether small changes in $\boldsymbol{z}^{(i)}$ can produce a large change in the proxy direction.

Proposition 4 (Local stability of the proxy)**.** *For an instance $\boldsymbol{x}_i$, suppose the leading singular value is simple, with $\gamma_i := (\sigma_1^{(i)})^2 - (\sigma_2^{(i)})^2 > 0$, and $\boldsymbol{v}_1^{(i)}$ is locally differentiable as a function of $\boldsymbol{z}^{(i)}$, and its differential satisfies*

$$\|\mathrm{d}\boldsymbol{v}_1^{(i)}\|_2 \le \frac{2\sigma_1^{(i)}}{\gamma_i}\|\mathrm{d}\boldsymbol{z}^{(i)}\|_F. \tag{9}$$

Remark. The bound shows that the leading direction controls the local sensitivity of the proxy. The spectral alignment objective encourages this separation by promoting the leading singular value relative to the remaining spectrum. This explains the compact representation used in synthetic training. See the full derivation in Appendix B.5.

4 Experiments

In this section, we conduct experiments to address the following research questions (RQs):

- **RQ1:** How does HoPA perform compared to existing baselines across different datasets, data ratios, and modality configurations? Does our method achieve higher efficiency?
- **RQ2:** Can our method generalize to different model architectures, maintaining superior performance with the distilled data?
- **RQ3:** How does each component, *i.e.*, pairwise *vs.* omniwise distillation and different learning objectives, affect the performance?
- **RQ4:** Is the premise supported by empirical analysis, including the rank-1 Gram matrix and efficiency on SVD?

4.1 Experimental Setups

Datasets. We evaluate our appoarch on five datasets with diverse modality configuration. MSR-VTT (Xu et al., 2016), VGGSound-S (Chen et al., 2020), and DiDeMo (Anne Hendricks et al., 2017) contain video, audio, and text, while SUN-SPOT (Mauceri et al., 2019) provides a distinct combination of image, depth, and text. We further include Ego4D-S (Grauman et al., 2022), which extends the evaluation to four modalities–video, audio, text, and IMU. Important statistics of these datasets are summarized in Appendix C.1.

Metrics. Model performance is evaluated using the top-K retrieval recall (R@K), which measures how often the correct match appears among the top-K retrieved results. Specifically, given a query from one modality, the model retrieves the top-K similar items from the other modality, and the proportion of correct matches is reported as R@K.

Baselines. We compare HoPA to random coreset selection and four dataset distillation methods: *MTT-VL* (Wu et al., 2023), *TESLA* (Cui et al., 2023), *LoRS* (Xu et al., 2024), and *RepBlend* (Zhang et al., 2025b). Since existing dataset distillation baselines operate on two modalities, we apply each baseline independently to each modailty pair respectively. All pairwise methods are adapted to

Table 1: **Experimental results on MSR-VTT.** This table shows the testing retrieval recall of models trained on the small coreset or synthetic set. "V": Video, "T": Text, and "A": Audio. We report the mean and standard deviation of the results. In each case, the best performance is in bold.

Triplets (Ratio)	Methods	R@1 ↑							R@5 ↑							Time↓ (mins)
		V→T	T→V	V→A	A→V	T→A	A→T	Avg.	V→T	T→V	V→A	A→V	T→A	A→T	Avg.	
100 (1.11%)	Random	$29.1_{\pm0.5}$	$\mathbf{37.2}_{\pm0.0}$	$22.5_{\pm0.1}$	$19.9_{\pm0.0}$	$6.9_{\pm0.1}$	$6.0_{\pm0.1}$	20.3	$54.1_{\pm0.2}$	$60.7_{\pm0.2}$	$43.8_{\pm0.1}$	$40.0_{\pm0.1}$	$17.8_{\pm0.1}$	$17.3_{\pm0.0}$	39.0	–
	MTT-VL	$34.3_{\pm0.6}$	$36.7_{\pm0.8}$	$22.4_{\pm0.1}$	$19.1_{\pm0.1}$	$6.7_{\pm0.5}$	$5.8_{\pm0.4}$	20.8	$59.0_{\pm0.2}$	$58.9_{\pm1.2}$	$43.3_{\pm0.1}$	$39.3_{\pm0.2}$	$18.6_{\pm0.6}$	$18.4_{\pm0.7}$	39.6	75.7
	TESLA	$34.0_{\pm0.5}$	$36.5_{\pm0.8}$	$22.4_{\pm0.1}$	$19.2_{\pm0.1}$	$6.7_{\pm0.3}$	$5.8_{\pm0.1}$	20.8	$58.8_{\pm0.6}$	$58.8_{\pm1.0}$	$43.1_{\pm0.2}$	$39.2_{\pm0.1}$	$18.4_{\pm0.2}$	$17.6_{\pm0.4}$	39.3	75.4
	LoRS	$36.1_{\pm0.8}$	$36.6_{\pm0.3}$	$21.8_{\pm0.1}$	$20.5_{\pm0.6}$	$7.8_{\pm0.1}$	$\mathbf{6.8}_{\pm0.5}$	**21.6**	$60.0_{\pm0.4}$	$60.4_{\pm1.1}$	$43.3_{\pm0.5}$	$41.7_{\pm1.1}$	$19.8_{\pm0.6}$	$19.6_{\pm0.4}$	40.8	77.1
	RepBlend	$\mathbf{36.2}_{\pm0.7}$	$36.7_{\pm0.7}$	$21.7_{\pm0.4}$	$19.8_{\pm0.9}$	$7.4_{\pm0.3}$	$6.6_{\pm0.4}$	21.4	$60.7_{\pm0.6}$	$60.0_{\pm1.2}$	$43.3_{\pm0.1}$	$40.6_{\pm0.8}$	$19.7_{\pm0.5}$	$18.6_{\pm0.3}$	40.5	82.0
	HoPA	$\mathbf{36.2}_{\pm0.5}$	$35.6_{\pm0.6}$	$\mathbf{22.6}_{\pm0.9}$	$\mathbf{21.3}_{\pm0.6}$	$\mathbf{7.9}_{\pm0.3}$	$6.0_{\pm0.5}$	**21.6**	$\mathbf{62.0}_{\pm0.3}$	$\mathbf{62.2}_{\pm0.6}$	$\mathbf{44.9}_{\pm0.2}$	$\mathbf{42.5}_{\pm0.4}$	$\mathbf{21.3}_{\pm0.2}$	$\mathbf{20.2}_{\pm0.4}$	**42.2**	**42.8**
200 (2.22%)	Random	$29.3_{\pm0.2}$	$\mathbf{37.1}_{\pm0.1}$	$22.5_{\pm0.0}$	$20.0_{\pm0.0}$	$6.9_{\pm0.1}$	$6.0_{\pm0.0}$	20.3	$54.4_{\pm0.2}$	$60.8_{\pm0.1}$	$43.7_{\pm0.1}$	$40.0_{\pm0.1}$	$18.0_{\pm0.1}$	$17.3_{\pm0.1}$	39.0	–
	MTT-VL	$31.1_{\pm0.1}$	$30.1_{\pm0.9}$	$21.0_{\pm1.3}$	$20.0_{\pm0.8}$	$5.8_{\pm0.2}$	$\mathbf{7.4}_{\pm0.6}$	19.2	$55.3_{\pm0.8}$	$57.4_{\pm0.3}$	$43.7_{\pm0.5}$	$44.7_{\pm0.1}$	$18.9_{\pm1.2}$	$21.0_{\pm0.7}$	40.2	103.6
	TESLA	$30.7_{\pm0.0}$	$30.5_{\pm0.3}$	$21.6_{\pm1.0}$	$20.2_{\pm0.5}$	$5.8_{\pm0.3}$	$\mathbf{7.4}_{\pm0.8}$	19.4	$55.9_{\pm1.0}$	$55.3_{\pm1.6}$	$43.6_{\pm0.9}$	$43.7_{\pm0.7}$	$19.5_{\pm0.6}$	$20.8_{\pm0.4}$	39.8	96.4
	LoRS	$33.0_{\pm0.5}$	$32.7_{\pm0.3}$	$22.5_{\pm0.8}$	$20.8_{\pm0.3}$	$6.0_{\pm0.2}$	$7.3_{\pm0.6}$	20.4	$58.9_{\pm1.0}$	$59.4_{\pm0.5}$	$\mathbf{45.2}_{\pm1.0}$	$\mathbf{45.7}_{\pm0.5}$	$19.5_{\pm0.7}$	$\mathbf{21.3}_{\pm0.3}$	41.7	100.9
	RepBlend	$33.4_{\pm0.4}$	$33.6_{\pm0.2}$	$22.1_{\pm0.6}$	$20.7_{\pm0.8}$	$5.5_{\pm0.6}$	$7.2_{\pm0.4}$	20.4	$57.8_{\pm0.5}$	$59.6_{\pm1.0}$	$45.1_{\pm0.6}$	$45.5_{\pm0.0}$	$18.1_{\pm0.3}$	$19.6_{\pm0.9}$	41.0	97.5
	HoPA	$\mathbf{37.3}_{\pm0.1}$	$34.8_{\pm0.5}$	$\mathbf{23.4}_{\pm0.4}$	$\mathbf{21.5}_{\pm1.0}$	$\mathbf{8.0}_{\pm0.1}$	$6.2_{\pm0.1}$	**21.9**	$\mathbf{63.2}_{\pm0.1}$	$\mathbf{62.4}_{\pm0.2}$	$44.8_{\pm0.6}$	$42.8_{\pm0.5}$	$\mathbf{22.2}_{\pm0.3}$	$\mathbf{20.6}_{\pm0.7}$	**42.7**	**48.2**
500 (5.55%)	Random	$34.1_{\pm0.3}$	$\mathbf{37.3}_{\pm0.1}$	$22.1_{\pm0.2}$	$19.9_{\pm0.0}$	$6.9_{\pm0.1}$	$6.6_{\pm0.1}$	21.2	$59.7_{\pm0.5}$	$61.2_{\pm0.2}$	$43.5_{\pm0.3}$	$40.2_{\pm0.2}$	$19.0_{\pm0.1}$	$17.3_{\pm0.1}$	40.3	–
	MTT-VL	$34.8_{\pm0.3}$	$36.0_{\pm0.5}$	$21.9_{\pm0.2}$	$20.0_{\pm0.5}$	$6.7_{\pm0.1}$	$6.6_{\pm0.1}$	21.0	$60.6_{\pm0.6}$	$59.0_{\pm0.6}$	$43.9_{\pm0.9}$	$\mathbf{44.8}_{\pm0.4}$	$19.0_{\pm0.3}$	$21.0_{\pm0.2}$	41.4	118.3
	TESLA	$35.1_{\pm0.7}$	$36.3_{\pm0.7}$	$20.1_{\pm0.2}$	$20.2_{\pm0.3}$	$6.8_{\pm0.2}$	$6.4_{\pm0.2}$	20.8	$60.3_{\pm0.3}$	$59.8_{\pm0.3}$	$43.5_{\pm0.9}$	$43.2_{\pm1.1}$	$19.7_{\pm0.4}$	$20.1_{\pm0.4}$	41.1	115.7
	LoRS	$35.9_{\pm0.5}$	$36.6_{\pm0.2}$	$22.6_{\pm0.5}$	$21.2_{\pm0.4}$	$6.9_{\pm0.4}$	$\mathbf{7.3}_{\pm0.3}$	21.8	$61.5_{\pm0.5}$	$60.6_{\pm1.4}$	$45.4_{\pm0.5}$	$44.5_{\pm0.8}$	$20.4_{\pm0.1}$	$\mathbf{21.9}_{\pm0.8}$	42.4	120.4
	RepBlend	$36.1_{\pm0.4}$	$36.5_{\pm0.6}$	$22.7_{\pm0.6}$	$21.2_{\pm0.3}$	$6.9_{\pm0.2}$	$\mathbf{7.3}_{\pm0.3}$	21.8	$61.7_{\pm0.4}$	$60.3_{\pm0.8}$	$45.6_{\pm0.6}$	$44.3_{\pm0.6}$	$20.1_{\pm0.7}$	$20.9_{\pm0.4}$	42.2	122.1
	HoPA	$\mathbf{36.3}_{\pm0.2}$	$35.1_{\pm0.3}$	$\mathbf{24.2}_{\pm0.2}$	$\mathbf{21.9}_{\pm0.4}$	$\mathbf{8.6}_{\pm0.2}$	$6.3_{\pm0.2}$	**22.1**	$\mathbf{63.4}_{\pm0.6}$	$\mathbf{62.0}_{\pm0.7}$	$\mathbf{45.7}_{\pm0.3}$	$44.1_{\pm0.9}$	$\mathbf{21.5}_{\pm1.0}$	$\mathbf{20.5}_{\pm1.0}$	**42.9**	**57.6**
Full	–	$37.1_{\pm0.1}$	$36.2_{\pm0.2}$	$24.4_{\pm0.5}$	$22.8_{\pm0.1}$	$9.0_{\pm0.3}$	$6.1_{\pm0.1}$	22.6	$64.9_{\pm0.1}$	$64.7_{\pm0.4}$	$47.6_{\pm0.2}$	$46.3_{\pm0.1}$	$23.6_{\pm0.4}$	$21.6_{\pm0.1}$	44.8	–

the same three-modality setting; the adaptation and budget-matching protocols are detailed in Appendix C.2 Note that we only involve one coreset selection method. This is because we focus on data distillation in this work. Furthermore, most coreset selection methods are limited to unimodal classification tasks (Xia et al., 2023; 2024b; Zheng et al., 2022), and their performance is not competitive with that of data distillation methods (Wang et al., 2022; Cui et al., 2024; Cazenavette et al., 2025; Yu et al., 2023; Lei & Tao, 2023; Joshi et al., 2025).

Implementation details. We utilize the pretrained encoders of ImageBind (Girdhar et al., 2023) to encode the data of different modalities. Consistent with the original architecture, we retain the modality-specific linear projection heads to map features into a fixed d-dimensional embedding space. Notably, for computational efficiency, we freeze the heavy backbone encoders, and only the projection heads are optimized during both the distillation and training stages, consistently following (Zhang et al., 2025b). Instead of synthesizing raw inputs, we directly synthesize the modality embeddings. We employ LoRS (Xu et al., 2024) as our base distillation method. To generate expert trajectories, we train the network on the full real dataset for 10 epochs, repeated over 20 independent runs. During the distillation phase, synthetic data is optimized using the SGD optimizer with a momentum of 0.5. We initialize the synthetic modality embeddings by randomly sampling from real data. Detailed hyperparameters across different cases are provided in Appendix C.3. Experiments are conducted on an NVIDIA TESLA A100 GPU. Results are reported over five random seeds.

4.2 Main Results (RQ1)

Performance comparison. To assess the quality of the distilled data, we evaluate the cross-modal retrieval performance of models trained solely on the distilled sets. Results on MSR-VTT and VGGSound-S are reported in Table 1 and Table 2, respectively, while results on DiDeMo are provided Appendix D.1. Across all datasets and compression ratios, HoPA consistently outperforms pairwise dataset distillation baselines, with absolute improvements of 0.2%–1.5% in Avg. R@1 and 0.5%–1.7% in Avg. R@5. This result demonstrates the benefit of jointly distilling multimodal data rather than treating modality pairs independently. Furthermore, HoPA preserves over 95% of the full-dataset performance on both R@1 and R@5 while using only 500 distilled examples (approximately 5%-6% of the original training data) across all three datasets. These results indicate that our distilled data effectively captures both modality-specific information and cross-modal alignment cues, even under aggressive compression.

Time consumption. The total distillation time is measured on an NVIDIA TESLA A100 GPU. It is reported in the rightmost column of the tables. For bimodal baselines, the reported distillation time is the total over the three pairwise runs, while HoPA requires only one joint run. Although HoPA involves SVD-based operations, it achieves an approximately $2\times$ speedup compared to pairwise distillation baselines because all modalities are distilled *simultaneously* rather than through repeated pairwise optimization.

Table 2: **Experimental results on VGGSound-S**. This table shows the testing retrieval recall of models trained on the small coreset or synthetic set. "V": Video, "T": Text, and "A": Audio. We report the mean and standard deviation of the results. In each case, the best performance is in bold.

Triplets (Ratio)	Methods	R@1 ↑							R@5 ↑							Time↓ (mins)
		V→T	T→V	V→A	A→V	T→A	A→T	Avg.	V→T	T→V	V→A	A→V	T→A	A→T	Avg.	
100 (1.25%)	Random	$5.7_{\pm 0.5}$	$7.7_{\pm 0.0}$	$26.3_{\pm 0.0}$	$25.7_{\pm 0.0}$	$5.6_{\pm 0.0}$	$3.7_{\pm 0.3}$	12.4	$25.9_{\pm 0.3}$	$28.3_{\pm 0.1}$	$\mathbf{54.3}_{\pm 0.1}$	$52.1_{\pm 0.1}$	$20.5_{\pm 0.1}$	$18.6_{\pm 0.2}$	33.3	–
	MTT-VL	$6.2_{\pm 0.0}$	$7.6_{\pm 0.0}$	$26.1_{\pm 0.2}$	$25.8_{\pm 0.0}$	$5.9_{\pm 0.2}$	$4.3_{\pm 0.4}$	12.7	$26.1_{\pm 0.0}$	$27.5_{\pm 0.0}$	$53.7_{\pm 0.4}$	$52.5_{\pm 0.1}$	$20.4_{\pm 0.3}$	$18.8_{\pm 0.2}$	33.2	80.9
	TESLA	$5.9_{\pm 0.3}$	$7.6_{\pm 0.2}$	$26.1_{\pm 0.2}$	$25.8_{\pm 0.1}$	$6.2_{\pm 0.3}$	$3.8_{\pm 0.1}$	12.6	$25.6_{\pm 0.6}$	$27.4_{\pm 0.4}$	$53.8_{\pm 0.4}$	$52.6_{\pm 0.1}$	$22.0_{\pm 0.2}$	$20.5_{\pm 0.7}$	33.7	81.5
	LoRS	$5.4_{\pm 0.5}$	$7.8_{\pm 0.2}$	$\mathbf{27.8}_{\pm 0.3}$	$\mathbf{26.2}_{\pm 0.1}$	$6.3_{\pm 0.2}$	$4.5_{\pm 0.2}$	13.0	$26.5_{\pm 0.6}$	$28.2_{\pm 0.1}$	$53.1_{\pm 0.4}$	$52.2_{\pm 0.3}$	$22.3_{\pm 0.4}$	$21.6_{\pm 0.4}$	34.0	84.9
	RepBlend	$5.6_{\pm 0.5}$	$7.6_{\pm 0.3}$	$27.2_{\pm 0.1}$	$25.0_{\pm 0.1}$	$6.1_{\pm 0.3}$	$4.5_{\pm 0.1}$	12.7	$27.1_{\pm 0.4}$	$28.6_{\pm 0.3}$	$54.1_{\pm 0.1}$	$51.4_{\pm 0.3}$	$20.9_{\pm 0.1}$	$19.9_{\pm 0.6}$	33.7	81.9
	HoPA	$\mathbf{6.6}_{\pm 0.4}$	$\mathbf{8.3}_{\pm 0.2}$	$27.3_{\pm 0.5}$	$26.1_{\pm 0.1}$	$\mathbf{6.6}_{\pm 0.1}$	$\mathbf{4.6}_{\pm 0.2}$	**13.3**	$\mathbf{28.6}_{\pm 0.7}$	$\mathbf{30.3}_{\pm 0.7}$	$54.0_{\pm 0.2}$	$\mathbf{53.2}_{\pm 0.2}$	$\mathbf{23.2}_{\pm 0.3}$	$\mathbf{22.2}_{\pm 0.5}$	**35.3**	**45.7**
200 (2.5%)	Random	$6.0_{\pm 0.4}$	$7.7_{\pm 0.0}$	$26.6_{\pm 0.3}$	$25.6_{\pm 0.3}$	$5.8_{\pm 0.1}$	$4.0_{\pm 0.5}$	12.6	$25.5_{\pm 0.8}$	$27.9_{\pm 0.5}$	$54.4_{\pm 0.2}$	$52.0_{\pm 0.4}$	$20.6_{\pm 0.2}$	$19.0_{\pm 0.8}$	33.2	–
	MTT-VL	$5.8_{\pm 0.4}$	$8.1_{\pm 0.2}$	$26.8_{\pm 0.4}$	$25.4_{\pm 0.6}$	$6.6_{\pm 0.3}$	$4.8_{\pm 0.3}$	12.9	$26.8_{\pm 0.1}$	$29.2_{\pm 0.1}$	$52.7_{\pm 0.6}$	$\mathbf{53.3}_{\pm 0.3}$	$23.8_{\pm 0.5}$	$22.8_{\pm 0.2}$	34.8	86.3
	TESLA	$6.0_{\pm 0.4}$	$7.6_{\pm 0.1}$	$25.8_{\pm 0.3}$	$25.7_{\pm 0.1}$	$5.7_{\pm 0.1}$	$4.2_{\pm 0.5}$	12.5	$26.1_{\pm 1.2}$	$27.4_{\pm 0.9}$	$52.5_{\pm 0.1}$	$52.7_{\pm 0.2}$	$20.3_{\pm 0.1}$	$19.0_{\pm 0.7}$	33.0	85.8
	LoRS	$6.4_{\pm 0.1}$	$8.2_{\pm 0.2}$	$26.7_{\pm 0.2}$	$25.6_{\pm 0.2}$	$6.9_{\pm 0.1}$	$5.1_{\pm 0.4}$	13.2	$27.5_{\pm 0.2}$	$28.9_{\pm 0.1}$	$53.4_{\pm 0.2}$	$53.2_{\pm 0.8}$	$24.6_{\pm 0.1}$	$22.7_{\pm 0.3}$	35.1	85.5
	RepBlend	$6.3_{\pm 0.5}$	$8.1_{\pm 0.3}$	$26.9_{\pm 0.3}$	$25.7_{\pm 0.2}$	$\mathbf{7.1}_{\pm 0.3}$	$\mathbf{5.2}_{\pm 0.6}$	13.2	$28.1_{\pm 0.5}$	$29.4_{\pm 0.5}$	$54.1_{\pm 0.4}$	$53.2_{\pm 0.1}$	$\mathbf{24.8}_{\pm 0.7}$	$22.9_{\pm 0.4}$	35.4	87.2
	HoPA	$\mathbf{6.7}_{\pm 0.3}$	$\mathbf{8.5}_{\pm 0.2}$	$\mathbf{27.0}_{\pm 0.1}$	$\mathbf{26.1}_{\pm 0.3}$	$6.9_{\pm 0.1}$	$4.9_{\pm 0.4}$	**13.4**	$\mathbf{28.6}_{\pm 0.6}$	$\mathbf{30.9}_{\pm 0.6}$	$\mathbf{54.8}_{\pm 0.4}$	$53.2_{\pm 0.2}$	$23.7_{\pm 0.5}$	$\mathbf{23.1}_{\pm 0.2}$	**35.7**	**50.4**
500 (6.25%)	Random	$6.1_{\pm 0.5}$	$7.9_{\pm 0.1}$	$26.7_{\pm 0.3}$	$25.8_{\pm 0.2}$	$6.1_{\pm 0.2}$	$4.1_{\pm 0.4}$	12.8	$27.2_{\pm 0.3}$	$28.8_{\pm 0.5}$	$54.5_{\pm 0.3}$	$52.3_{\pm 0.2}$	$21.1_{\pm 0.2}$	$19.4_{\pm 0.2}$	33.9	–
	MTT-VL	$5.9_{\pm 0.3}$	$8.2_{\pm 0.4}$	$26.9_{\pm 0.2}$	$25.8_{\pm 0.4}$	$6.7_{\pm 0.5}$	$4.9_{\pm 0.1}$	13.0	$27.0_{\pm 0.6}$	$29.5_{\pm 1.2}$	$53.0_{\pm 0.8}$	$53.5_{\pm 0.3}$	$24.0_{\pm 0.1}$	$23.0_{\pm 0.7}$	35.0	123.7
	TESLA	$6.1_{\pm 0.4}$	$7.9_{\pm 0.1}$	$26.0_{\pm 0.1}$	$\mathbf{26.0}_{\pm 0.1}$	$6.4_{\pm 0.2}$	$4.6_{\pm 0.2}$	12.8	$27.1_{\pm 0.4}$	$28.0_{\pm 0.2}$	$52.8_{\pm 0.0}$	$53.0_{\pm 0.0}$	$22.9_{\pm 0.2}$	$21.2_{\pm 0.5}$	34.2	124.3
	LoRS	$6.4_{\pm 0.5}$	$8.4_{\pm 0.1}$	$27.0_{\pm 0.0}$	$25.9_{\pm 0.0}$	$\mathbf{7.2}_{\pm 0.1}$	$5.1_{\pm 0.2}$	13.3	$27.8_{\pm 0.2}$	$30.7_{\pm 0.0}$	$53.7_{\pm 0.0}$	$53.5_{\pm 0.0}$	$24.9_{\pm 0.4}$	$22.8_{\pm 0.0}$	35.6	124.8
	RepBlend	$6.4_{\pm 0.0}$	$8.4_{\pm 0.2}$	$27.0_{\pm 0.0}$	$25.9_{\pm 0.5}$	$7.1_{\pm 0.2}$	$\mathbf{5.3}_{\pm 0.8}$	13.3	$28.3_{\pm 0.1}$	$30.8_{\pm 0.3}$	$54.3_{\pm 0.3}$	$\mathbf{53.6}_{\pm 0.8}$	$\mathbf{25.0}_{\pm 0.2}$	$23.0_{\pm 0.6}$	35.8	125.5
	HoPA	$\mathbf{6.6}_{\pm 0.1}$	$\mathbf{8.6}_{\pm 0.2}$	$\mathbf{27.1}_{\pm 0.5}$	$\mathbf{26.0}_{\pm 0.1}$	$6.9_{\pm 0.0}$	$\mathbf{5.3}_{\pm 0.5}$	**13.4**	$\mathbf{29.2}_{\pm 1.0}$	$\mathbf{32.3}_{\pm 0.4}$	$\mathbf{55.0}_{\pm 0.4}$	$\mathbf{53.6}_{\pm 0.2}$	$24.8_{\pm 0.2}$	$\mathbf{23.5}_{\pm 0.4}$	**36.4**	**55.1**
Full	–	$6.8_{\pm 0.2}$	$8.6_{\pm 0.1}$	$28.0_{\pm 0.1}$	$26.6_{\pm 0.1}$	$7.5_{\pm 0.2}$	$4.8_{\pm 0.3}$	13.7	$30.8_{\pm 0.8}$	$34.6_{\pm 0.3}$	$55.7_{\pm 0.1}$	$54.0_{\pm 0.1}$	$26.8_{\pm 0.2}$	$24.5_{\pm 0.8}$	37.7	–

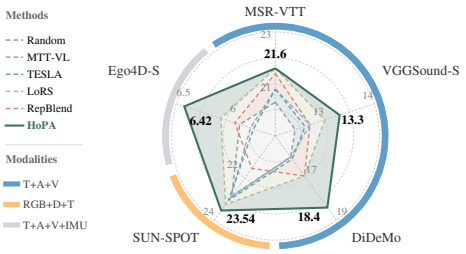

Figure 2: **Generalization to different modality compositions and cardinalities.**

Table 3: **Cross-architecture generalization results** w/ and w/o distilled settings.

Model	Setting	R@1	R@5	R@10
ImageBind	w/o Distilled	12.25	32.87	44.32
	w/ Random	12.83	33.89	45.57
	w/ Distilled	**13.15**	**35.33**	**47.34**
LanguageBind	w/o Distilled	4.91	19.94	29.56
	w/ Random	5.87	21.76	32.67
	w/ Distilled	**5.93**	**22.11**	**33.32**
OmniBind	w/o Distilled	12.92	34.88	48.03
	w/ Random	12.75	34.93	48.28
	w/ Distilled	**13.13**	**35.63**	**48.87**

Generalization across modality configurations. To examine whether HoPA remains effective under different modality combinations and with more modalities, we further evaluate HoPA on SUN-SPOT, which comprises image, depth, and text modalities, and Ego4D-S, which involves four modalities. As shown in Figure 2, HoPA achieves the highest average R@1 on both datasets, reaching 23.54 on SUN-SPOT and 6.42 on Ego4D-S. These results indicate that HoPA remains effective under diverse modality configurations.

4.3 Cross-architecture Generalization (RQ2)

Following previous works (Zhao & Bilen, 2021), we conduct a cross-architecture evaluation to examine the transferability of our distilled data. We first distill a 200-instance synthetic dataset on full ImageBind using VGGSound-S. We then fine-tune different pretrained multimodal models with the distilled data and evaluate retrieval performance on the VGGSound-S test set. For fair cross-architecture evaluation, we follow the same parameter-efficient protocol as ImageBind, where the pretrained multimodal encoders are frozen and modality-specific projection heads are optimized. As reported in Table 3, the distilled data improves performance not only on ImageBind, the source architecture used for distillation, but also on LanguageBind (Zhu et al., 2024) and OmniBind (Wang et al., 2025), which were unseen during the distillation process. This suggests that our synthetic data captures transferable supervision that generalizes across different multimodal architectures.

4.4 Ablation Study (RQ3)

Effect of each learning objective. To study the contribution of each component, we conduct an ablation study on our learning objective. As shown in Figure 3, removing either the modality-level loss $\mathcal{L}_{\text{M}}$ or instance-level loss $\mathcal{L}_{\text{wBCE}}$ consistently degrades performance, indicating that both objectives are essential. Moreover, similarity mining further improves the results on top of these objectives, and the full model achieves the best overall performance.

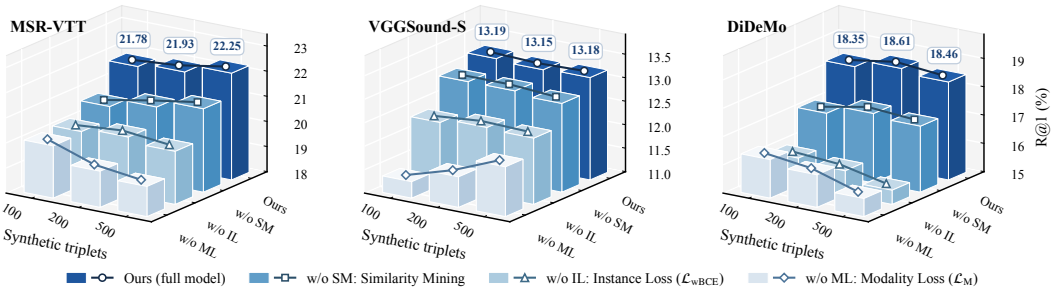

Figure 3: **Ablation study** of each learning objective across different datasets and numbers of queries.

Pairwise distillation & omniwise distillation. To verify the benefit of omniwise distillation, we compare HoPA with three pairwise variants: *3-pair*, which optimizes all three modality pairs; *T-bind*, which aligns video and audio to text; and *V-bind*, which aligns text and audio to video. As shown in Figure 4, HoPA consistently achieves the best Avg. R@1 across all three datasets and synthetic budgets. Moreover, pairwise variants tend to stagnate or degrade with more synthetic queries, especially on MSR-VTT and DiDeMo, whereas HoPA remains stable or improves, supporting the benefit of preserving joint omnimodal structure over separate pairwise alignments. Full results and details are provided in Appendix D.2.

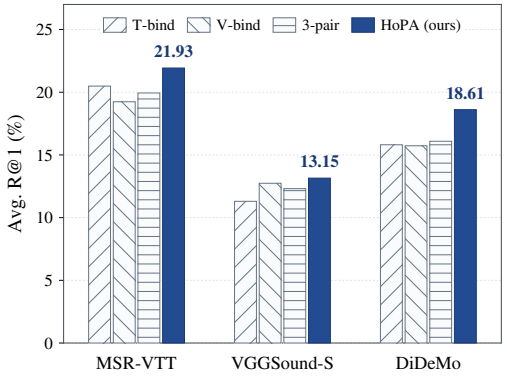

Figure 4: **Pairwise vs. omniwise distillation.** Comparison of HoPA with three pairwise variants

4.5 Further Analysis (RQ4)

Effect of proxy rank. HoPA constructs the omnimodal proxy using the rank-1 Gram matrix. We further examine whether retaining additional information is beneficial by extending the proxy from rank-1 to rank-2. Across all datasets and compression budgets, rank-2 consistently underperforms rank-1 (Appendix D.4). This indicates that the leading singular direction captures the dominant cross-modal consensus, while additional directions introduce modality-specific variations.

Efficiency analysis. A potential concern with our proposed method is the computational cost introduced by the SVD operation. We further quantify the computational overhead introduced by SVD in Appendix D.5. SVD incurs only a 3.5% overhead during expert-buffer construction, while 39.1% of the distillation time is on SVD operation due to the TESLA (Cui et al., 2023) algorithm we use in practice to reduce the memory consumption, which back-propagates through the SVD operation involving high-order derivative calculations. Despite the additional SVD cost, HoPA remains more efficient than pairwise multimodal distillation methods, as shown in our main results.

5 Conclusion

In this work, we investigated dataset distillation in omnimodal settings beyond two modalities, which is a problem that has received little prior attention. We proposed a unified omnimodal dataset distillation method, termed HoPA, that leverages a compact proxy to model high-order alignments, enabling efficient and scalable distillation without explicit pairwise modality enumeration. Our theoretical analysis also provides complementary support for the proxy design. Extensive experimental results demonstrate that our HoPA consistently achieves strong compression efficiency while preserving downstream performance across diverse datasets. These findings highlight the feasibility and promise of omnimodal dataset distillation at scale, and open new directions for data-efficient learning in complex omnimodal systems.

AI USE STATEMENT

We used an LLM solely to polish the writing and correct grammatical issues during the preparation of this submission. The LLM was not involved in any idea generation, experiment design, or analysis. All scientific contributions are entirely made by the authors.

ETHICS STATEMENT

This work studies omnimodal dataset distillation for data-efficient multimodal learning. Our experiments use existing benchmark datasets and subsets constructed from them, together with pretrained encoders. We follow the applicable licenses and terms of use associated with the datasets and models used in this study.

REPRODUCIBILITY STATEMENT

We have provided implementation details, involving illustrative algorithm descriptions in Sec. 3, Sec. 4.1, Appendix C, and pseudo-code in Algorithm 1. The source code is available at https://anonymous.4open.science/r/HoPA-7C5E/.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report, 2023. arXiv preprint arXiv:2303.08774.

Zeyuan Allen-Zhu and Yuanzhi Li. Lazysvd: Even faster svd decomposition yet without agonizing pain. In *NeurIPS*, 2016.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *CVPR*, pp. 4750–4759, 2022.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Generalizing dataset distillation via deep generative prior. In *CVPR*, pp. 3739–3748, 2023.

George Cazenavette, Antonio Torralba, and Vincent Sitzmann. Dataset distillation for pre-trained self-supervised vision models. In *NeurIPS*, 2025.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pp. 721–725, 2020.

Yilan Chen, Wei Huang, and Lily Weng. Provable and efficient dataset distillation for kernel ridge regression. In *NeurIPS*, pp. 88739–88771, 2024.

Jaehyun Choi, Jiwan Hur, Gyojin Han, Jaemyung Yu, and Junmo Kim. Prism: Video dataset condensation with progressive refinement and insertion for sparse motion. In *CVPR*, 2026.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *ICLR*, 2025.

Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *ICML*, pp. 6565–6590, 2023.

Justin Cui, Ruochen Wang, Yuanhao Xiong, and Cho-Jui Hsieh. Mitigating bias in dataset distillation. In *ICML*, 2024.

Xiao Cui, Yulei Qin, Wengang Zhou, Hongsheng Li, and Houqiang Li. Optical: Leveraging optimal transport for contribution allocation in dataset distillation. In *CVPR*, pp. 15245–15254, 2025.

Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. In *NeurIPS*, pp. 34391–34404, 2022.

Guodong Ding, Rongyu Chen, and Angela Yao. Condensing action segmentation datasets via generative network inversion. In *CVPR*, pp. 17733–17742, 2025.

Le Dong, Jinghao Bian, Jingyang Hou, Jingliang Hu, Yilei Shi, Weisheng Dong, Xiao Xiang Zhu, and Lichao Mou. High-order progressive trajectory matching for medical image dataset distillation. In *MICCAI*, pp. 273–283, 2025.

Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *CVPR*, pp. 3749–3758, 2023.

Luyang Fang, Xiaowei Yu, Jiazhang Cai, Yongkai Chen, Shushan Wu, Zhengliang Liu, Zhenyuan Yang, Haoran Lu, Xilin Gong, Yufang Liu, et al. Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions. *Artificial Intelligence Review*, 59(1):17, 2026.

Jiahui Geng, Zongxiong Chen, Yuandou Wang, Herbert Woisetschlaeger, Sonja Schimmler, Ruben Mayer, Zhiming Zhao, and Chunming Rong. A survey on dataset distillation: Approaches, applications and future directions, 2023. arXiv preprint arXiv:2305.01975.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, pp. 15180–15190, 2023.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pp. 18995–19012, 2022.

Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *CVPR*, pp. 15793–15803, 2024.

Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In *ICLR*, 2024.

Muyang He, Shuo Yang, Tiejun Huang, and Bo Zhao. Large-scale dataset pruning with dynamic uncertainty. In *CVPR*, pp. 7713–7722, 2024.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge; New York, 2nd edition, 2013. ISBN 9780521839402.

Lingxiao Huang, Jian Li, and Xuan Wu. On optimal coreset construction for euclidean (k, z)-clustering. In *STOC*, pp. 1594–1604, 2024.

Jun-Peng Jiang, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and Han-Jia Ye. Representation learning for tabular data: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6488–6508, 2026.

Peng Jin, Hao Li, Li Yuan, Shuicheng Yan, and Jie Chen. Hierarchical banzhaf interaction for general video-language representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2125–2139, 2024.

Siddharth Joshi, Jiayi Ni, and Baharan Mirzasoleiman. Dataset distillation via knowledge distillation: towards efficient self-supervised pre-training of deep networks. In *ICLR*, 2025.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. arXiv preprint arXiv:2001.08361.

Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *ICML*, pp. 11102–11118, 2022.

Hojun Lee, Suyoung Kim, Junhoo Lee, Jaeyoung Yoo, and Nojun Kwak. Coreset selection for object detection. In *CVPR*, pp. 7682–7691, 2024.

Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *ICML*, pp. 12352–12364, 2022.

Yongmin Lee and Hye Won Chung. Selmatch: Effectively scaling up dataset distillation via selection-based initialization and partial updates by trajectory matching. In *ICML*, 2024.

Shiye Lei and Dacheng Tao. A comprehensive survey of dataset distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):17–32, 2023.

Lijiang Li, Zuwei Long, Yunhang Shen, Heting Gao, Haoyu Cao, Xing Sun, Caifeng Shan, Ran He, and Chaoyou Fu. Omni-diffusion: Unified multimodal understanding and generation with masked discrete diffusion, 2026a. arXiv preprint arXiv:2603.06577.

Muquan Li, Hang Gou, Yingyi Ma, Rongzheng Wang, Ke Qin, and Tao He. Fixed anchors are not enough: Dynamic retrieval and persistent homology for dataset distillation. In *CVPR*, 2026b.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, et al. Not all tokens are what you need for pretraining. In *NeurIPS*, pp. 29029–29063, 2024.

Dai Liu, Jindong Gu, Hu Cao, Carsten Trinitis, and Martin Schulz. Dataset distillation by automatic training trajectories. In *ECCV*, pp. 334–351, 2024.

Haoyang Liu, Yijiang Li, Tiancheng Xing, Peiran Wang, Vibhu Dalal, Luwei Li, Jingrui He, and Haohan Wang. Dataset distillation via the wasserstein metric. In *ICCV*, pp. 1205–1215, 2025.

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

Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Principled multimodal representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Huimin Lu, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. Unidetox: Universal detoxification of large language models via dataset distillation. In *ICLR*, 2025.

Run Luo, Xiaobo Xia, Lu Wang, Longze Chen, Renke Shan, Jing Luo, Min Yang, and Tat-Seng Chua. Next-omni: Towards any-to-any omnimodal foundation models with discrete flow matching. In *ICLR*, 2026.

Cecilia Mauceri, Martha Palmer, and Christoffer Heckman. Sun-spot: An rgb-d dataset with spatial referring expressions. In *ICCV Workshop*, pp. 1883–1886, 2019.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text. *arXiv preprint arXiv:2210.14395*, 2022.

Dang Nguyen, Zeman Li, Mohammadhossein Bateni, Vahab Mirrokni, Meisam Razaviyayn, and Baharan Mirzasoleiman. Synthetic text generation for training large language models via gradient matching, 2025. arXiv preprint arXiv:2502.17607.

Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression, 2020. arXiv preprint arXiv:2011.00050.

Shitong Shao, Zikai Zhou, Huanran Chen, and Zhiqiang Shen. Elucidating the design space of dataset condensation. In *NeurIPS*, pp. 99161–99201, 2024.

Cheng Shen, Yew-Soon Ong, and Joey Tianyi Zhou. Condenselm: Llms-driven text dataset condensation via reward matching. In *EMNLP*, pp. 1237–1252, 2025.

Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pp. 567–576, 2015.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *NeurIPS*, pp. 19523–19536, 2022.

Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. Dˆ 4: Dataset distillation via disentangled diffusion model. In *CVPR*, pp. 5809–5818, 2024.

Duo Su, Huyu Wu, Huanran Chen, Yiming Shi, Yuzhu Wang, Xi Ye, and Jun Zhu. Diffusion models as dataset distillation priors. In *ICLR*, 2025.

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*, pp. 9206–9216, 2020.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models, 2023. arXiv preprint arXiv:2312.11805.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report, 2025. arXiv preprint arXiv:2504.07491.

Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In *CVPR*, pp. 12196–12205, 2022.

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

Zehan Wang, Ziang Zhang, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Hengshuang Zhao, and Zhou Zhao. Omnibind: Large-scale omni multimodal representation via binding spaces. In *ICLR*, 2025.

Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation, 2023. arXiv preprint arXiv:2308.07545.

Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. Rethinking data selection at scale: Random selection is almost all you need, 2024a. arXiv preprint arXiv:2410.09335.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *ICLR*, 2023.

Xiaobo Xia, Jiale Liu, Shaokun Zhang, Qingyun Wu, Hongxin Wei, and Tongliang Liu. Refined coreset selection: Towards minimal coreset size under model performance constraints. In *ICML*, 2024b.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pp. 5288–5296, 2016.

Yue Xu, Zhilin Lin, Yusong Qiu, Cewu Lu, and Yong-Lu Li. Low-rank similarity mining for multimodal dataset distillation. In *ICML*, 2024.

Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):150–170, 2023.

Sheng-Feng Yu, Jia-Jiun Yao, and Wei-Chen Chiu. Boost self-supervised dataset distillation via parameterization, predefined augmentation, and approximation. In *CVPR*, 2025.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pp. 818–833, 2014.

Bolin Zhang, Jiahao Wang, Qianlong Du, Jiajun Zhang, Zhiying Tu, and Dianhui Chu. A survey on data selection for LLM instruction tuning. *Journal of Artificial Intelligence Research*, 83, 2025a.

Xin Zhang, Ziruo Zhang, Jiawei Du, Zuozhu Liu, and Joey Tianyi Zhou. Beyond modality collapse: Representations blending for multimodal dataset distillation. In *NeurIPS*, 2025b.

Yihua Zhang, Yimeng Zhang, Aochuan Chen, Jinghan Jia, Jiancheng Liu, Gaowen Liu, Mingyi Hong, Shiyu Chang, and Sijia Liu. Selectivity drives productivity: efficient dataset pruning for enhanced transfer learning. In *NeurIPS*, pp. 36913–36937, 2023a.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Universal multimodal representation for language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9169–9185, 2023b.

Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *ICML*, pp. 12674–12685, 2021.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *WACV*, pp. 6514–6523, 2023.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching, 2020. arXiv preprint arXiv:2006.05929.

Zhenghao Zhao, Haoxuan Wang, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. Efficient multimodal dataset distillation via generative models. In *NeurIPS*, 2025.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates, 2022. arXiv preprint arXiv:2210.15809.

Wenliang Zhong, Haoyu Tang, Qinghai Zheng, Mingzhu Xu, Yupeng Hu, and Weili Guan. Towards stable and storage-efficient dataset distillation: Matching convexified trajectory. In *CVPR*, pp. 25581–25589, 2025a.

Xinhao Zhong, Hao Fang, Bin Chen, Xulin Gu, Meikang Qiu, Shuhan Qi, and Shu-Tao Xia. Hierarchical features matter: A deep exploration of progressive parameterization method for dataset distillation. In *CVPR*, pp. 30462–30471, 2025b.

Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. In *NeurIPS*, pp. 9813–9827, 2022.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*, 2024.

Yongshuo Zong, Oisin Mac Aodha, and Timothy M Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5299–5318, 2024.

APPENDIX

A Related Work **16**

A.1 Dataset Distillation 16

A.2 Omnimodal Alignment 16

B Theoretical Analysis **16**

B.1 Proof of Eq. (1) 16

B.2 Nonnegative proxy selection 17

B.3 Proxy similarity approximation 17

B.4 WBCE loss approximation with a fixed similarity matrix 18

B.5 Local proxy stability 18

B.6 Gradient Analysis 19

C Supplementary Implementation Details **19**

C.1 Datasets 19

C.2 Baselines 20

C.3 Hyperparameter Settings 20

C.4 Algorithm Flow 21

D Additional Results **21**

D.1 Results on more datasets 21

D.2 Pairwise distillation & Omniwise distillation 22

D.3 Effect of Each Learning Objective 22

D.4 Effect of proxy rank 23

D.5 Efficiency analysis of SVD operation 23

D.6 Visualization 23

E Limitations **27**

A Related Work

In this section, we review recent literature related to this work, including dataset distillation (Appendix A.1) and omnimodal alignment (Appendix A.2).

A.1 Dataset Distillation

Dataset distillation seeks to compress a large-scale dataset into a compact synthetic set that can effectively substitute the original data for model training, while preserving comparable performance (Wang et al., 2018; Yu et al., 2023; Lei & Tao, 2023; Fang et al., 2026; Zhong et al., 2025b; Cui et al., 2025; Su et al., 2025). Early work has mainly focused on single-modal data distillation, with a particular emphasis on image data. These methods can be broadly categorized into several representative paradigms. Gradient matching methods optimize synthetic samples by aligning the training gradients induced by real and synthetic data (Zhao et al., 2020; Lee et al., 2022; Kim et al., 2022). Trajectory matching methods instead preserve the optimization dynamics of training on the original dataset by matching parameter trajectories or intermediate model states (Cazenavette et al., 2022; Guo et al., 2024; Zhong et al., 2025a). Distribution matching methods align feature-level or statistical distributions between real and distilled data (Liu et al., 2025; Cui et al., 2025; Zhao & Bilen, 2023). Meta-learning–based methods formulate the distilled set as learnable data optimized for fast adaptation or strong generalization across training episodes (Nguyen et al., 2020; Such et al., 2020; Deng & Russakovsky, 2022; Zhou et al., 2022; Chen et al., 2024). Generative methods have introduced pretrained or trainable generators, such as generative adversarial networks (GANs) and diffusion models, to parameterize the distilled data and improve cross-architecture generalization (Cazenavette et al., 2023; Gu et al., 2024; Su et al., 2024).

A.2 Omnimodal Alignment

Omnimodal alignment has emerged as a fundamental problem in multimodal learning to learn coherent and consistent representations across a diverse set of modalities. Representative works in this direction include ImageBind (Girdhar et al., 2023), LanguageBind (Zhu et al., 2024), and more recent studies on scalable and principled cross-modal representation learning (Cicchetti et al., 2025; Wang et al., 2025; Liu et al., 2026; Luo et al., 2026; Zhang et al., 2023b; Zong et al., 2024; Jiang et al., 2026; Jin et al., 2024; Li et al., 2026a). These methods typically seek to map heterogeneous inputs into a shared embedding space in which semantically related samples from different modalities are well aligned. Such an alignment provides a strong foundation for cross-modal transfer, retrieval, and joint understanding.

Different from prior omnimodal alignment methods, which are mainly concerned with representation learning from full training datasets, our work leverages omnimodal alignment as a surrogate objective for dataset distillation. The central goal is not merely to learn a shared representation space, but to preserve the cross-modal relational structure of the original omnimodal dataset using a highly compact distilled set.

B Theoretical Analysis

We use the original notation from the method. For one instance, $\boldsymbol{z} \in \mathbb{R}^{k\times d}$ has unit rows, $2 \le k \le d$, $G = \boldsymbol{z}\boldsymbol{z}^\top$, and $\boldsymbol{z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is a thin SVD. Consequently,

$$\sum_{\ell=1}^{k} \sigma_\ell^2 = k, \qquad 1 \le \sigma_1^2 \le k, \qquad 0 \le \rho = 1 - \sigma_1^2/k \le 1 - 1/k. \tag{10}$$

Following(Cicchetti et al., 2025; Liu et al., 2026), we also assume that the similarity scores between modalities are nonnegative, and therefore G is entrywise nonnegative. It does not impose a sign restriction on entries of the feature vectors in an arbitrary coordinate system.

B.1 Proof of Eq. (1)

Proof. Let $\boldsymbol{\delta}_r := \boldsymbol{\theta}_r^S - \boldsymbol{\theta}_r^T$ and $\Delta_r := \|\mathrm{g}_S(\boldsymbol{\theta}_r^S) - \mathrm{g}_T(\boldsymbol{\theta}_r^S)\|$, the one-step update gives

$$
\begin{aligned}
\|\boldsymbol{\delta}_{r+1}\|_2 &= \|\boldsymbol{\delta}_r - \eta\big(\mathbf{g}_S(\boldsymbol{\theta}_r^S) - \mathbf{g}_T(\boldsymbol{\theta}_r^T)\big)\|_2 \\
&= \|\boldsymbol{\delta}_r - \eta \underbrace{\big(\mathbf{g}_S(\boldsymbol{\theta}_r^S) - \mathbf{g}_T(\boldsymbol{\theta}_r^S)\big)}_{\text{data mismatch}} - \eta \underbrace{\big(\mathbf{g}_T(\boldsymbol{\theta}_r^S) - \mathbf{g}_T(\boldsymbol{\theta}_r^T)\big)}_{\text{Lipschitz term}} \|_2 \\
&\le \|\boldsymbol{\delta}_r\|_2 + \eta\Delta_r + \eta\|\mathbf{g}_T(\boldsymbol{\theta}_r^S) - \mathbf{g}_T(\boldsymbol{\theta}_r^T)\|_2 \\
&\le (1+\eta L)\|\boldsymbol{\delta}_r\|_2 + \eta\Delta_r, \qquad (L-\text{Lipschitz condition})
\end{aligned}
\tag{11}
$$

Since both rollouts share the same initialization, $\boldsymbol{\delta}_0 = \boldsymbol{\theta}_0^S - \boldsymbol{\theta}_0^T = \mathbf{0}$, unrolling the recursion gives

$$
\|\delta_n\| \le \eta \sum_{r=0}^{n-1} (1+\eta L)^{n-1-r}\Delta_r, \tag{12}
$$

which is the stated bound. □

B.2 Nonnegative proxy selection

Existence and projections. A nonnegative symmetric matrix has a nonnegative principal eigenvector (Horn & Johnson, 2013). Here this also follows directly from the Rayleigh quotient. If $\boldsymbol{u}$ is a unit leading eigenvector, entrywise nonnegativity gives

$$
|\boldsymbol{u}|^\top \boldsymbol{G}|\boldsymbol{u}| \ge \boldsymbol{u}^\top \boldsymbol{G}\boldsymbol{u} = \sigma_1^2. \tag{13}
$$

Since $|\boldsymbol{u}|$ is unit and σ_1^2 is the largest eigenvalue, equality holds, so $|\boldsymbol{u}|$ is a nonnegative leading eigenvector. Choose such a unit vector $\boldsymbol{u}_1$. Its paired feature direction satisfies

$$
\boldsymbol{v}_1 = \frac{\boldsymbol{z}^\top \boldsymbol{u}_1}{\sigma_1}, \qquad \|\boldsymbol{v}_1\|_2^2 = \frac{\boldsymbol{u}_1^\top \boldsymbol{G}\boldsymbol{u}_1}{\sigma_1^2} = 1, \qquad \boldsymbol{z}\boldsymbol{v}_1 = \sigma_1\boldsymbol{u}_1 \ge 0. \tag{14}
$$

Moreover, $(\boldsymbol{z}^\top\boldsymbol{z})v_1 = \sigma_1^2\boldsymbol{v}_1$, so $\boldsymbol{v}_1$ is a leading right singular vector. Lemma 2 expresses it as a nonnegative combination of the modality representations. Thus $\boldsymbol{v}_1$ lies in their generated cone, which includes its boundary; strict interior membership or ordinary convex-hull membership is not asserted.

Mean-based sign choice at a simple leading value. For $\bar{\boldsymbol{z}} = (1/k)\sum_m \boldsymbol{z}_m$, Eq. (14) gives

$$
\boldsymbol{v}_1^\top \bar{\boldsymbol{z}} = \frac{\sigma_1}{k}\mathbf{1}^\top \boldsymbol{u}_1 > 0. \tag{15}
$$

Ideal common direction. For $\boldsymbol{z} = \mathbf{1}\boldsymbol{c}^\top$ with $\|\boldsymbol{c}\|_2 = 1$, the singular values are $(\sqrt{k}, 0, \ldots, 0)$ and the rule selects $\boldsymbol{u}_1 = \mathbf{1}/\sqrt{k}$, $\boldsymbol{v}_1 = \boldsymbol{c}$. Thus, for two instances aligned at unit directions $\boldsymbol{c}_i, \boldsymbol{c}_j$,

$$
\boldsymbol{z}^{(i)} = \mathbf{1}\boldsymbol{c}_i^\top, \qquad \langle \boldsymbol{z}_m^{(i)}, \boldsymbol{z}_n^{(j)} \rangle = \boldsymbol{c}_i^\top \boldsymbol{c}_j = \widetilde{s}(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{16}
$$

This gives the exact similarity described in Sec. 3.2.

B.3 Proxy similarity approximation

Proof of Theorem 3. For one instance, let $a_m = \langle \boldsymbol{z}_m, \boldsymbol{v}_1 \rangle$. Eq. (14) and unit normalization give $a_m \in [0,1]$ and $\sum_m a_m^2 = \sigma_1^2$. The mean squared distance to the proxy therefore satisfies

$$
D := \frac{1}{k}\sum_m \|\boldsymbol{z}_m - \boldsymbol{v}_1\|_2^2 = \frac{2}{k}\sum_m (1 - a_m) \le \frac{2}{k}\sum_m (1 - a_m^2) = 2\rho. \tag{17}
$$

For two instances, substituting $\widetilde{s}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{v}_1^{(i)})^\top \boldsymbol{v}_1^{(j)}$ and averaging the inner-product difference gives

$$
\begin{aligned}
&\frac{1}{k^2}\sum_{m,n\in\mathcal{M}} \Big| \langle \boldsymbol{z}_m^{(i)}, \boldsymbol{z}_n^{(j)} \rangle - \widetilde{s}(\boldsymbol{x}_i, \boldsymbol{x}_j) \Big| \\
=&\frac{1}{k^2}\sum_{m,n\in\mathcal{M}} \Big| \langle \boldsymbol{z}_m^{(i)} - \boldsymbol{v}_1^{(i)}, \boldsymbol{z}_n^{(j)} \rangle + \langle \boldsymbol{v}_1^{(i)}, \boldsymbol{z}_n^{(j)} - \boldsymbol{v}_1^{(j)} \rangle \Big| \\
\le&\frac{1}{k}\sum_{m\in\mathcal{M}} \|\boldsymbol{z}_m^{(i)} - \boldsymbol{v}_1^{(i)}\|_2 + \frac{1}{k}\sum_{n\in\mathcal{M}} \|\boldsymbol{z}_n^{(j)} - \boldsymbol{v}_1^{(j)}\|_2 \\
\le&\sqrt{D_i} + \sqrt{D_j} \le \sqrt{2\rho_i} + \sqrt{2\rho_j}.
\end{aligned}
\tag{18}
$$

This proves Eq. (7). □

B.4 WBCE LOSS APPROXIMATION WITH A FIXED SIMILARITY MATRIX

Consider the same synthetic batch and fix the learned similarity matrix $\boldsymbol{S}_e = [s_{ij}]$, whose entries serve as the loss targets. For a similarity $r \in [-1, 1]$, define the scalar loss

$$\phi_s(r) = -s \log(p(r)) - (1 - s) \log(1 - p(r)), \qquad p(r) = \mathrm{sigmoid}(r/\tau'), \tag{19}$$

where $\tau' > 0$. Let $w_{ij} \geq 0$ be the fixed weights of the included ordered pairs. Excluded pairs may be assigned zero weight. Define

$$\mathcal{L}_{\mathrm{all}}(\boldsymbol{S}_e) = \sum_{i,j} \frac{w_{ij}}{k^2} \sum_{m,n \in \mathcal{M}} \phi_{s_{ij}}(\langle \boldsymbol{z}_m^{(i)}, \boldsymbol{z}_n^{(j)} \rangle), \tag{20}$$

$$\mathcal{L}_{\mathrm{wBCE}}(\boldsymbol{S}_e) = \sum_{i,j} w_{ij} \phi_{s_{ij}}(\widetilde{s}(\boldsymbol{x}_i, \boldsymbol{x}_j)). \tag{21}$$

Both objectives use identical representations, targets, masks, temperatures, and normalization. The first is a defined reference objective, not an identification with an experimental baseline using different targets or loss normalization.

For any finite real s, differentiation gives $\phi_s'(r) = [p(r) - s]/\tau'$,

$$|\phi_s'(r)| \leq \frac{|s|(1 - p(r)) + |1 - s| p(r)}{\tau'} \leq \frac{\kappa_s}{\tau'}, \qquad \kappa_s := \max\{1, |s|, |1 - s|\}. \tag{22}$$

The triangle inequality, this Lipschitz bound, and Theorem 3 yield

$$|\mathcal{L}_{\mathrm{all}} - \mathcal{L}_{\mathrm{wBCE}}| \leq \sum_{i,j} \frac{w_{ij} \kappa_{s_{ij}}}{\tau'} (\sqrt{2\rho_i} + \sqrt{2\rho_j}) \leq \underbrace{\frac{2\sqrt{2}}{\tau'} \sum_{i,j} w_{ij} \kappa_{s_{ij}}}_{C} \sqrt{\rho_{\max}}. \tag{23}$$

This proves Eq. (8) for ordinary BCE and the stabilized form. For targets in $[0, 1]$, $\kappa_s = 1$; outside that interval the same algebra holds without treating s as a probability. The constant is finite at a fixed batch and finite targets; it is not asserted to be uniform over unconstrained updates of $\boldsymbol{S}_e$.

B.5 LOCAL PROXY STABILITY

Proof of Proposition 4. For one instance, set $\boldsymbol{C}_z = \boldsymbol{z}^\top \boldsymbol{z}$, $\lambda_1 = \sigma_1^2$, and $\gamma = \sigma_1^2 - \sigma_2^2$. The positive spectral gap implies that the leading eigenvalue of C_z is simple. With a locally consistent sign choice, its unit eigenvector $\boldsymbol{v}_1$ is locally differentiable. Differentiating $C_z \boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1$ and $\boldsymbol{v}_1^\top \boldsymbol{v}_1 = 1$ gives

$$(\lambda_1 \boldsymbol{I} - \boldsymbol{C}_z)\, \mathrm{d}\boldsymbol{v}_1 = (\mathrm{d}\boldsymbol{C}_z - \mathrm{d}\lambda_1 \boldsymbol{I})\boldsymbol{v}_1, \qquad \boldsymbol{v}_1^\top \mathrm{d}\boldsymbol{v}_1 = 0. \tag{24}$$

In a complete orthonormal feature-space eigenbasis, let

$$\boldsymbol{R} = (\lambda_1 \boldsymbol{I} - \boldsymbol{C}_z)^\dagger = \sum_{r=2}^{d} \frac{\boldsymbol{v}_r \boldsymbol{v}_r^\top}{\lambda_1 - \lambda_r}. \tag{25}$$

It satisfies $\boldsymbol{R}\boldsymbol{v}_1 = 0$ and $\|\boldsymbol{R}\|_{\mathrm{op}} = 1/\gamma$. The sum includes feature nullspace directions, with $\lambda_r = 0$ for $r > k$. Hence

$$\mathrm{d}\boldsymbol{v}_1 = \boldsymbol{R}\left[\boldsymbol{z}^\top \mathrm{d}\boldsymbol{z} + (\mathrm{d}\boldsymbol{z})^\top \boldsymbol{z}\right] \boldsymbol{v}_1, \qquad \|\mathrm{d}\boldsymbol{v}_1\|_2 \leq \frac{2\sigma_1}{\gamma} \|\mathrm{d}\mathbf{z}\|_F. \tag{26}$$

Using $\sigma_1 \leq \sqrt{k}$ and the gap bound proves Eq. (9). The mean-oriented rule remains defined in a neighborhood of the base point, so this is an ambient derivative; its bound also applies to tangent perturbations of unit-row representations. □

B.6 Gradient Analysis

To provide theoretical insight into how our proposed objective drives omnimodal alignment and semantic relation, we analyze the gradient of the total loss $\mathcal{L} = \mathcal{L}_{\mathrm{M}} + \mathcal{L}_{\mathrm{wBCE}}$ with respect to the multimodal representation matrix $\boldsymbol{z}$. Our analysis reveals that the optimization process consists of two synergistic forces: an *Within-instance centripetal force* driven by singular value maximization, and an *Between-instance relational supervision* regulated by the proxy.

Within-instance alignment. The modality-level loss $\mathcal{L}_{\mathrm{M}}$ aligns different modalities of the same instance. By differentiating the loss to the representation matrix $\boldsymbol{z}$, we obtain the gradient direction:

$$\frac{\partial \mathcal{L}_{\mathrm{M}}}{\partial \boldsymbol{z}} = \frac{1}{\tau}\left[(p_1 - 1)\boldsymbol{u}_1\boldsymbol{v}_1^\top + \sum_{j=2}^{k} p_j \boldsymbol{u}_j \boldsymbol{v}_j^\top\right], \tag{27}$$

where p_j is the softmax probability of the j-th singular value. This result reveals a self-adaptive alignment mechanism: since $(p_1 - 1) < 0$, the first term generates a negative gradient that acts as a centripetal force, pulling all modality embeddings aligned with the principal direction $\boldsymbol{v}_1$. Meanwhile, the second term ($p_j > 0$ for $j \geq 2$) acts as a penalty, suppressing energy in all other directions. This theoretically guarantees that minimizing $\mathcal{L}_{\mathrm{M}}$ drives the Gram matrix towards rank-1, achieving full omnimodal alignment.

Between-instance relational supervision. Write $\boldsymbol{v}_i := \boldsymbol{v}_1^{(i)}$ for the proxy of instance i, let $p_{ij} = \mathrm{sigmoid}(\boldsymbol{v}_i^\top \boldsymbol{v}_j/\tau')$, $a_{ij} = w_{ij}(p_{ij} - s_{ij})$.

$$\frac{\partial \mathcal{L}_{\mathrm{wBCE}}}{\partial \boldsymbol{v}_i} = \frac{1}{\tau'}\sum_j (a_{ij} + a_{ji})\boldsymbol{v}_j. \tag{28}$$

For distinct instances, the coefficient $a_{ij} + a_{ji}$ determines the pair's pressure to decrease or increase its shared similairty. A positive coefficient favours decreasing the proxy similarity, whereas a negative coefficient favours increasing it. The supervision therefore corrects the similarity prediction according to the learned relation targets.

C Supplementary Implementation Details

C.1 Datasets

In the main paper, we evaluate our method on MSR-VTT, VGGSound-S, and DiDeMo. The details of the datasets are summarized below.

- MSR-VTT (Xu et al., 2016) is a benchmark from Microsoft Research for video captioning. It contains 10,000 web video clips and about 200,000 clip-sentence pairs spanning 20 categories, with roughly 20 captions per clip. We use the split MSR-VTT 1kA following Imagebind (Girdhar et al., 2023).
- VGGSound-S (Chen et al., 2020) is a curated subset of the original VGGSound dataset. VGGSound is an audio–visual benchmark created by the University of Oxford's Visual Geometry Group. It consists of more than 200,000 ten-second YouTube clips, each depicting a sound source that is visible in the video and labeled with one of 310+ sound-event classes. The full collection spans over 550 hours of audio-visual content. To construct VGGSound-S, we randomly sampled 8,000 examples from the official training set and 2,000 examples from the validation set, yielding a smaller dataset that preserves the diversity of the original while enabling more efficient multi-modal retrieval experiments.
- DiDeMo (Anne Hendricks et al., 2017) is a collection of more than 10,000 unedited personal videos scraped from Flickr, each roughly 25–30s long with an accompanying audio track. Each video comes with multiple natural-language descriptions (about 3–5 per clip) written by annotators. These descriptions were validated so that they uniquely match the described clip. In total, the dataset offers more than 40,000 (description, video) pairs and spans a wide range of real-world content such as pets, concerts, and sports. We use the official split, which contains 8,395 training videos, 1,065 validation videos, and 1,004 test videos.

Table 4: **Hyperparameter settings for buffer.**

Dataset	MSR-VTT	VGGSound-S	DiDeMo
epochs	10	10	10
num_experts	20	20	20
batch_size	128	128	128
lr_teacher	0.01	0.01	0.01
τ	0.1	0.1	0.1
τ'	0.2	0.2	0.2

- SUN-SPOT (Mauceri et al., 2019) is a public spatial referring-expression dataset built on RGB-D observations from SUN RGB-D (Song et al., 2015). We treat each annotated target object as an instance, crop its aligned RGB and depth observations using the provided object box, and associate it with the referring expression describing that object. After retaining complete RGB–depth–text instances, our split contains 2,460/249/410 training/validation/test objects.
- Ego4D-S (Grauman et al., 2022) is a four-modal subset of Ego4D, a large-scale egocentric video dataset capturing daily activities in diverse real-world environments. We construct Ego4D-S using the video-level split released by IMU2CLIP (Moon et al., 2022). Each instance is defined by a ten-second window centered on a timestamped narration, associating the corresponding video, audio, text, and six-axis IMU observations. We retain instances with valid observations in all four modalities and remove duplicate cleaned narrations within each split. The resulting subset contains 11,742 training, 203 validation, and 790 test instances.

C.2 Baselines

Our main experiments involve three or more modalities. Since the compared dataset distillation baselines are designed for bimodal data, each baseline is executed independently $k(k-1)/2$ times for the k-modality dataset. For a budget of N, each bimodal baseline synthesizes N samples for each modality pair, while HoPA synthesizes N multimodal triplets in a single joint run. These directional results are then reported together in the main tables. The details of the baselines are showed below.

- MTT-VL (Wu et al., 2023) extends Matching Training Trajectories (MTT) from unimodal classification to vision-language dataset distillation. It jointly distills paired multimodal samples under a contrastive learning objective and matches the optimization trajectories of the modality-specific branches.
- TESLA (Cui et al., 2023) is a memory-efficient trajectory matching method built upon MTT. It decomposes the gradient computation through the unrolled student optimization so that trajectory matching can be performed with constant memory with respect to the number of unrolled steps. We adapt this memory-efficient trajectory matching to MTT-VL and optimize the corresponding cross-modal objective.
- LoRS (Xu et al., 2024) is specifically designed for multimodal dataset distillation and preserves cross-modal correspondence. In addition to synthesizing paired multimodal samples, LoRS jointly leans a soft cross-modal similarity matrix that captures the correspondence between synthetic instances.
- RepBlend (Zhang et al., 2025b) addresses the modality collapse problem in multimodal dataset distillation, where distilled representations become overly concentrated within each modality while the distribution gap between modalities increases. It introduces representation blending to weaken overly dominant cross-modal supervision and preserve intra-modal diversity.

C.3 Hyperparameter Settings

The hyperparameter settings are summarized in Table 4 and Table 5.

Table 5: **Hyperparameter settings for dataset distillation.** We separate globally shared hyperparameters from dataset- and size-specific settings.

Shared Settings	
Hyperparameter	Value
lr_data	100
lr_lr	1×10^{-4}
lr_teacher	0.01
mini_batch_size	25
τ	0.1
τ'	0.2
syn_steps	16
expert_epochs	2

Dataset- and Size-specific Settings									
	MSR-VTT			VGGSound-S			DiDeMo		
Hyperparameter	100	200	500	100	200	500	100	200	500
lr_sim	10.0	10.0	100.0	10.0	10.0	100.0	10.0	10.0	100.0
sim_rank	10	20	20	10	20	40	10	20	20
sim_alpha	1.0	0.5	0.01	1.0	1.0	0.1	1.0	0.5	0.05
iterations	2,000	2,000	3,000	2,000	2,000	3,000	2,000	2,000	3,000
max_start_epoch	5	5	7	5	5	7	5	5	7

Algorithm 1 The technical flow of the HoPA framework

Require: Inputs:

The real dataset $\mathcal{D} = \{(\boldsymbol{x}_i^{m_1}, \boldsymbol{x}_i^{m_2}, \ldots, \boldsymbol{x}_i^{m_k})\}_{i=1}^{N}$ with $k = |\mathcal{M}|$ modalities per instance.
Expert trajectory buffer $\mathcal{B}$.
Modality-specific encoders $\{E_m\}_{m \in \mathcal{M}}$.
Student rollout steps t, segment length T, learning rates η, α, the number of total iterations $Iter$.

Ensure: Synthetic set $\mathcal{D}_{\text{syn}}$ and learnable synthetic similarity matrix $\boldsymbol{S}_e$.

Initialize synthetic data $\mathcal{D}_{\text{syn}}$ and learnable parameters $\boldsymbol{\phi}_S$ for matrix $\boldsymbol{S}_e$;
for $it = 1$ to `Iter` **do**
  Sample a trajectory segment $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_T) \sim \mathcal{B}$;
  Initialize student parameters $\boldsymbol{\theta}_0^{\text{S}} \leftarrow \boldsymbol{\theta}_0$;
  // Simulate Training on Synthetic Data
  for $r = 1$ to t **do**
    For each synthetic instance i, compute multimodal representations $\boldsymbol{z}^{(i)} = [E_m(\boldsymbol{x}_{\text{syn},i}^m; \boldsymbol{\theta}_{r-1,m}^{\text{S}})]_{m \in \mathcal{M}}^{\top}$;
    For each instance i, compute SVD: $\boldsymbol{z}^{(i)} = \sum_{l=1}^{k} \sigma_l^{(i)} \boldsymbol{u}_l^{(i)} \boldsymbol{v}_l^{(i)\top}$, and extract $\sigma_1^{(i)}$, $\boldsymbol{v}_1^{(i)}$;
    // Compute Combined Objective (Eq. (5))
    $\mathcal{L}_{\text{M}} \leftarrow -\frac{1}{|\mathcal{D}_{\text{syn}}|} \sum_i \log \frac{\exp(\sigma_1^{(i)}/\tau)}{\sum_{l=1}^{k} \exp(\sigma_l^{(i)}/\tau)}$;
    Construct proxy similarity matrix with $\tilde{s}_{ij} = \boldsymbol{v}_1^{(i)\top} \boldsymbol{v}_1^{(j)}$;
    $\mathcal{L}_{\text{wBCE}} \leftarrow \mathcal{L}_{\text{wBCE}}(\boldsymbol{S}_e)$;
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{M}} + \mathcal{L}_{\text{wBCE}}$;
    $\boldsymbol{\theta}_r^{\text{S}} \leftarrow \boldsymbol{\theta}_{r-1}^{\text{S}} - \eta \nabla_{\boldsymbol{\theta}_{r-1}^{\text{S}}} \mathcal{L}$;
  end for
  // Trajectory Matching (Eq. (4))
  $\mathcal{L}_{\text{omni-traj}} \leftarrow \sum_{m \in \mathcal{M}} \frac{\|\boldsymbol{\theta}_{t,m}^{\text{S}} - \boldsymbol{\theta}_{T,m}\|_2^2}{\|\boldsymbol{\theta}_{0,m} - \boldsymbol{\theta}_{T,m}\|_2^2}$;
  Update $\mathcal{D}_{\text{syn}}$ and $\boldsymbol{\phi}_S$ using gradient descent with learning rate α on $\mathcal{L}_{\text{omni-traj}}$;
end for
Return: $\mathcal{D}_{\text{syn}}$ and $\boldsymbol{S}_e$.

C.4 Algorithm Flow

For a better understanding of the technical details of our method, we provide an algorithm flow in Algorithm 1.

D Additional Results

D.1 Results on more datasets

Table 6 reports the complete DiDeMo results. HoPA achieves the highest Avg. R@1 and Avg. R@5 at all three synthetic budgets. With 500 synthetic triplets, it reaches 18.5 Avg. R@1 and 37.8 Avg. R@5, retaining 96.4% and 95.2% of the corresponding full-data performance, respectively.

Table 6: **Experimental results on DiDeMo.** This table shows the testing retrieval recall of models trained on the small coreset or synthetic set. "V": Video, "T": Text, and "A": Audio. We report the mean and standard deviation of the results. In each case, the best performance is in bold.

Triplets (Ratio)	Methods	R@1 ↑							R@5 ↑							Time (mins) ↓
		V→T	T→V	V→A	A→V	T→A	A→T	Avg.	V→T	T→V	V→A	A→V	T→A	A→T	Avg.	
100 (1.19%)	Random	$27.4_{\pm 0.1}$	$27.7_{\pm 0.1}$	$17.4_{\pm 0.1}$	$15.5_{\pm 0.1}$	$4.4_{\pm 0.1}$	$3.8_{\pm 0.1}$	16.0	$50.3_{\pm 0.1}$	$54.1_{\pm 0.1}$	$36.6_{\pm 0.1}$	$36.2_{\pm 0.1}$	$13.8_{\pm 0.1}$	$13.7_{\pm 0.0}$	34.1	–
	MTT-VL	$26.6_{\pm 0.1}$	$26.0_{\pm 0.4}$	$15.8_{\pm 0.6}$	$18.8_{\pm 0.5}$	$4.4_{\pm 0.3}$	$4.4_{\pm 0.4}$	16.0	$51.5_{\pm 0.4}$	$50.7_{\pm 1.5}$	$37.8_{\pm 0.3}$	$41.6_{\pm 0.5}$	$13.4_{\pm 0.8}$	$14.0_{\pm 1.0}$	34.8	77.3
	TESLA	$26.6_{\pm 0.1}$	$26.0_{\pm 0.4}$	$16.2_{\pm 0.1}$	$\mathbf{18.9}_{\pm 0.6}$	$4.4_{\pm 0.4}$	$4.2_{\pm 0.3}$	16.1	$51.5_{\pm 0.4}$	$50.7_{\pm 1.5}$	$37.9_{\pm 0.3}$	$41.8_{\pm 0.5}$	$13.5_{\pm 1.1}$	$14.6_{\pm 0.1}$	35.0	78.1
	LoRS	$28.0_{\pm 0.5}$	$28.4_{\pm 0.5}$	$17.9_{\pm 0.2}$	$18.0_{\pm 0.8}$	$4.3_{\pm 0.1}$	$4.7_{\pm 0.4}$	16.9	$52.7_{\pm 0.3}$	$54.1_{\pm 0.7}$	$40.2_{\pm 1.2}$	$\mathbf{42.2}_{\pm 0.6}$	$13.3_{\pm 0.5}$	$13.4_{\pm 1.0}$	36.0	79.4
	RepBlend	$27.7_{\pm 0.5}$	$28.5_{\pm 0.7}$	$18.0_{\pm 0.2}$	$18.2_{\pm 0.9}$	$4.3_{\pm 0.1}$	$4.7_{\pm 0.5}$	16.9	$52.9_{\pm 0.0}$	$54.3_{\pm 0.8}$	$\mathbf{40.8}_{\pm 0.7}$	$41.9_{\pm 0.1}$	$13.4_{\pm 0.6}$	$13.5_{\pm 1.3}$	36.1	78.2
	HoPA	$\mathbf{31.5}_{\pm 0.2}$	$\mathbf{31.2}_{\pm 1.1}$	$\mathbf{18.9}_{\pm 0.6}$	$18.2_{\pm 0.7}$	$\mathbf{5.3}_{\pm 0.5}$	$\mathbf{5.0}_{\pm 0.4}$	**18.4**	$\mathbf{56.6}_{\pm 0.7}$	$\mathbf{56.0}_{\pm 0.9}$	$39.9_{\pm 0.9}$	$40.0_{\pm 1.1}$	$\mathbf{15.1}_{\pm 0.5}$	$\mathbf{15.6}_{\pm 0.7}$	**37.2**	**41.9**
200 (2.38%)	Random	$28.3_{\pm 0.2}$	$28.7_{\pm 0.2}$	$17.3_{\pm 0.1}$	$15.4_{\pm 0.1}$	$4.4_{\pm 0.1}$	$4.0_{\pm 0.1}$	16.4	$51.4_{\pm 0.5}$	$54.9_{\pm 0.1}$	$36.7_{\pm 0.1}$	$35.9_{\pm 0.0}$	$14.2_{\pm 0.1}$	$13.7_{\pm 0.2}$	34.5	–
	MTT-VL	$27.9_{\pm 0.5}$	$27.5_{\pm 0.8}$	$15.5_{\pm 0.3}$	$16.8_{\pm 0.4}$	$4.6_{\pm 0.2}$	$4.2_{\pm 0.3}$	16.1	$53.2_{\pm 0.6}$	$53.0_{\pm 1.0}$	$36.5_{\pm 0.2}$	$38.4_{\pm 0.4}$	$14.1_{\pm 0.5}$	$13.8_{\pm 0.3}$	34.8	95.0
	TESLA	$27.8_{\pm 0.6}$	$27.3_{\pm 0.3}$	$15.9_{\pm 0.2}$	$16.9_{\pm 0.5}$	$4.6_{\pm 0.3}$	$4.1_{\pm 0.2}$	16.1	$53.5_{\pm 0.4}$	$53.1_{\pm 0.8}$	$36.6_{\pm 0.3}$	$38.6_{\pm 0.6}$	$14.0_{\pm 0.4}$	$14.1_{\pm 0.1}$	35.0	97.4
	LoRS	$29.2_{\pm 0.4}$	$28.6_{\pm 0.6}$	$18.0_{\pm 0.3}$	$16.1_{\pm 0.5}$	$4.8_{\pm 0.2}$	$5.0_{\pm 0.3}$	17.0	$54.8_{\pm 0.5}$	$55.2_{\pm 0.6}$	$40.5_{\pm 0.5}$	$\mathbf{42.0}_{\pm 0.3}$	$14.5_{\pm 0.4}$	$13.9_{\pm 0.6}$	36.8	96.3
	RepBlend	$29.0_{\pm 0.5}$	$28.5_{\pm 0.2}$	$18.0_{\pm 0.3}$	$16.0_{\pm 0.2}$	$4.7_{\pm 0.1}$	$5.0_{\pm 0.1}$	16.9	$55.0_{\pm 0.2}$	$55.1_{\pm 0.5}$	$\mathbf{41.2}_{\pm 0.4}$	$41.8_{\pm 0.2}$	$14.6_{\pm 0.2}$	$14.0_{\pm 0.4}$	37.0	97.4
	HoPA	$\mathbf{31.8}_{\pm 0.9}$	$\mathbf{31.7}_{\pm 0.2}$	$\mathbf{19.6}_{\pm 0.8}$	$\mathbf{18.1}_{\pm 0.5}$	$\mathbf{5.4}_{\pm 0.2}$	$\mathbf{5.1}_{\pm 0.2}$	**18.6**	$\mathbf{57.2}_{\pm 0.6}$	$\mathbf{56.9}_{\pm 0.6}$	$40.1_{\pm 1.2}$	$40.7_{\pm 0.2}$	$\mathbf{15.4}_{\pm 0.7}$	$\mathbf{15.3}_{\pm 0.1}$	**37.6**	**45.7**
500 (5.96%)	Random	$29.9_{\pm 0.1}$	$29.0_{\pm 0.3}$	$17.4_{\pm 0.1}$	$15.6_{\pm 0.1}$	$4.6_{\pm 0.1}$	$3.8_{\pm 0.1}$	16.7	$54.9_{\pm 0.2}$	$56.5_{\pm 0.5}$	$37.0_{\pm 0.2}$	$35.6_{\pm 0.4}$	$14.3_{\pm 0.1}$	$14.1_{\pm 0.1}$	35.4	–
	MTT-VL	$29.5_{\pm 1.0}$	$29.4_{\pm 1.1}$	$15.7_{\pm 0.1}$	$15.4_{\pm 0.1}$	$4.8_{\pm 0.4}$	$4.1_{\pm 0.2}$	16.5	$56.3_{\pm 0.8}$	$56.4_{\pm 0.1}$	$35.8_{\pm 0.1}$	$36.3_{\pm 0.1}$	$14.9_{\pm 0.3}$	$13.8_{\pm 0.1}$	35.6	120.2
	TESLA	$29.1_{\pm 1.1}$	$28.7_{\pm 0.2}$	$15.7_{\pm 0.1}$	$15.4_{\pm 0.1}$	$4.8_{\pm 0.5}$	$4.1_{\pm 0.2}$	16.3	$55.9_{\pm 0.3}$	$56.5_{\pm 0.0}$	$35.8_{\pm 0.1}$	$36.3_{\pm 0.1}$	$14.7_{\pm 0.0}$	$13.9_{\pm 0.0}$	35.5	118.9
	LoRS	$30.3_{\pm 0.6}$	$28.8_{\pm 0.4}$	$18.1_{\pm 0.4}$	$16.3_{\pm 0.2}$	$\mathbf{5.0}_{\pm 0.3}$	$\mathbf{5.3}_{\pm 0.2}$	17.3	$56.6_{\pm 0.9}$	$56.4_{\pm 0.5}$	$41.0_{\pm 0.3}$	$\mathbf{41.9}_{\pm 0.4}$	$\mathbf{15.3}_{\pm 0.3}$	$14.3_{\pm 0.4}$	37.6	122.3
	RepBlend	$30.0_{\pm 0.6}$	$28.5_{\pm 0.1}$	$18.0_{\pm 0.4}$	$16.2_{\pm 0.0}$	$4.9_{\pm 0.3}$	$5.2_{\pm 0.0}$	17.1	$56.1_{\pm 0.3}$	$56.1_{\pm 0.4}$	$\mathbf{41.9}_{\pm 0.4}$	$41.6_{\pm 0.0}$	$15.2_{\pm 0.1}$	$14.5_{\pm 0.2}$	37.6	122.8
	HoPA	$\mathbf{31.3}_{\pm 1.1}$	$\mathbf{31.4}_{\pm 0.3}$	$\mathbf{19.8}_{\pm 1.3}$	$\mathbf{18.6}_{\pm 0.1}$	$4.8_{\pm 0.1}$	$4.9_{\pm 0.0}$	**18.5**	$\mathbf{57.3}_{\pm 2.9}$	$\mathbf{56.8}_{\pm 1.4}$	$41.2_{\pm 0.4}$	$41.4_{\pm 1.2}$	$15.1_{\pm 0.4}$	$\mathbf{14.7}_{\pm 0.1}$	**37.8**	**53.6**
Full	–	$32.7_{\pm 0.4}$	$31.9_{\pm 0.2}$	$20.7_{\pm 0.5}$	$19.5_{\pm 0.2}$	$5.6_{\pm 0.1}$	$4.8_{\pm 0.1}$	19.2	$58.6_{\pm 0.3}$	$58.7_{\pm 0.3}$	$43.2_{\pm 0.2}$	$44.5_{\pm 0.3}$	$16.4_{\pm 0.2}$	$16.9_{\pm 0.3}$	39.7	–

Table 7: **Experimental results on SUN-SPOT.** This table shows the testing retrieval recall of models trained on the small coreset or synthetic set at $Q = 50$ (2.03%). "R": RGB image, "T": Text, and "D": Depth. We report the mean and standard deviation of the results.

Method	R→T	T→R	R→D	D→R	T→D	D→T	Avg. ↑
Random	$33.60_{\pm 0.46}$	$\mathbf{34.56}_{\pm 0.51}$	$27.31_{\pm 1.06}$	$23.29_{\pm 1.45}$	$9.21_{\pm 1.49}$	$9.10_{\pm 1.41}$	$22.85_{\pm 0.46}$
MTT-VL	$30.92_{\pm 2.51}$	$27.97_{\pm 0.93}$	$27.98_{\pm 0.84}$	$31.33_{\pm 0.70}$	$9.48_{\pm 0.36}$	$9.91_{\pm 0.23}$	$22.93_{\pm 0.50}$
TESLA	$29.32_{\pm 1.75}$	$30.37_{\pm 0.42}$	$29.72_{\pm 0.70}$	$29.85_{\pm 1.52}$	$\mathbf{9.77}_{\pm 0.10}$	$9.24_{\pm 0.70}$	$23.05_{\pm 0.14}$
LoRS	$28.92_{\pm 1.06}$	$29.87_{\pm 0.48}$	$\mathbf{29.99}_{\pm 0.23}$	$\mathbf{31.46}_{\pm 0.61}$	$8.82_{\pm 0.10}$	$\mathbf{10.44}_{\pm 0.70}$	$23.25_{\pm 0.24}$
RepBlend	$29.32_{\pm 0.70}$	$28.98_{\pm 0.17}$	$24.90_{\pm 1.06}$	$27.71_{\pm 1.06}$	$8.93_{\pm 0.59}$	$9.37_{\pm 1.01}$	$21.54_{\pm 0.29}$
HoPA	$\mathbf{33.87}_{\pm 1.43}$	$33.67_{\pm 1.03}$	$27.95_{\pm 1.49}$	$27.79_{\pm 0.35}$	$9.08_{\pm 0.36}$	$8.86_{\pm 0.44}$	$\mathbf{23.54}_{\pm 0.11}$

Table 7 reports the full results on SUN-SPOT dataset. Under a $Q = 50$ (2.03%) budget, HoPA achives an average R@1 of 23.53%, compared with 23.25% for LoRS, the strongest competing baseline. These results show that HoPA remains effective on different modalities.

Table 8 reports the full results on Ego4D subset. Under a $Q = 100$ (0.85%) budget, HoPA achives an average R@1 of 6.42%, compared with 6.05%, the strongest competing baseline. These results show that HoPA remains effective with more modalities.

D.2 Pairwise distillation & Omniwise distillation

We compare HoPA with three pairwise variants.

3-pair optimizes all three pairwise alignments:

$$\mathcal{L}_{\text{3-pair}} = \mathcal{L}_{v,t} + \mathcal{L}_{v,a} + \mathcal{L}_{a,t}.$$

T-bind uses text as the binding modality:

$$\mathcal{L}_{\text{T-bind}} = \mathcal{L}_{v,t} + \mathcal{L}_{a,t},$$

while *V-bind* uses video as the binding modality:

$$\mathcal{L}_{\text{V-bind}} = \mathcal{L}_{v,t} + \mathcal{L}_{v,a}.$$

Here, $\mathcal{L}_{\alpha,\beta}$ denotes the pairwise wBCE objective between modalities α and β, following (Xu et al., 2024). The complete Avg. R@1 and Avg. R@5 results are reported in Table 9.

D.3 Effect of Each Learning Objective

Figure 5 reports the complete R@1, R@5, and R@10 results for the objective ablations. Removing either $\mathcal{L}_{\text{M}}$ or $\mathcal{L}_{\text{wBCE}}$ consistently degrades retrieval performance, confirming that within-instance spectral alignment and inter-instance discrimination are complementary. The modality-level loss has

Table 8: **Experimental results on Ego4D-S.** This table shows the retrieval R@1 (%) of models trained on the small coreset or synthetic set at $Q = 100$ (0.85%). "V": Video, "T": Text, "A": Audio, and "I": IMU. We report the mean and standard deviation of the results.

Method	V→T	T→V	V→A	A→V	V→I	I→V	T→A	A→T	T→I	I→T	A→I	I→A	Avg. ↑
Random	$13.63_{\pm 0.57}$	$16.75_{\pm 0.99}$	$5.75_{\pm 0.57}$	$6.90_{\pm 0.99}$	$5.25_{\pm 0.28}$	$4.43_{\pm 0.49}$	$4.60_{\pm 0.75}$	$6.57_{\pm 0.75}$	$0.66_{\pm 0.28}$	$2.46_{\pm 0.99}$	$1.48_{\pm 0.00}$	$1.81_{\pm 0.28}$	$5.86_{\pm 0.26}$
MTT-VL	$12.15_{\pm 1.24}$	$11.33_{\pm 3.23}$	$6.08_{\pm 0.75}$	$8.37_{\pm 1.78}$	$4.60_{\pm 1.03}$	$6.90_{\pm 0.49}$	$5.42_{\pm 0.85}$	$4.93_{\pm 0.49}$	$2.30_{\pm 0.75}$	$3.45_{\pm 0.99}$	$1.31_{\pm 1.03}$	$2.13_{\pm 1.14}$	$5.75_{\pm 0.11}$
TESLA	$11.99_{\pm 1.03}$	$11.66_{\pm 3.21}$	$6.08_{\pm 0.75}$	$7.88_{\pm 1.30}$	$4.60_{\pm 1.03}$	$6.73_{\pm 0.75}$	$5.75_{\pm 0.28}$	$4.76_{\pm 0.28}$	$2.30_{\pm 0.75}$	$3.45_{\pm 0.99}$	$1.31_{\pm 1.03}$	$2.13_{\pm 1.14}$	$5.72_{\pm 0.17}$
LoRS	$11.49_{\pm 2.48}$	$9.52_{\pm 3.01}$	$6.90_{\pm 0.99}$	$9.36_{\pm 0.85}$	$7.72_{\pm 0.75}$	$7.22_{\pm 0.75}$	$5.75_{\pm 1.03}$	$5.58_{\pm 1.24}$	$1.81_{\pm 0.57}$	$2.79_{\pm 0.75}$	$1.97_{\pm 0.49}$	$2.46_{\pm 0.49}$	$6.05_{\pm 0.35}$
RepBlend	$12.15_{\pm 2.22}$	$9.69_{\pm 3.56}$	$6.40_{\pm 0.85}$	$8.70_{\pm 1.03}$	$7.22_{\pm 0.57}$	$7.22_{\pm 0.28}$	$4.93_{\pm 1.30}$	$5.25_{\pm 1.03}$	$1.64_{\pm 0.75}$	$2.79_{\pm 0.75}$	$2.30_{\pm 0.28}$	$2.63_{\pm 0.57}$	$5.91_{\pm 0.21}$
HoPA	$\mathbf{14.29}_{\pm 1.30}$	$\mathbf{16.75}_{\pm 0.49}$	$6.40_{\pm 0.49}$	$7.06_{\pm 0.75}$	$5.91_{\pm 1.48}$	$5.91_{\pm 0.00}$	$5.25_{\pm 1.03}$	$\mathbf{7.55}_{\pm 0.75}$	$0.99_{\pm 0.00}$	$3.28_{\pm 1.50}$	$1.64_{\pm 0.28}$	$1.97_{\pm 0.00}$	$\mathbf{6.42}_{\pm 0.12}$

Table 9: **Empirical comparisons of pairwise distillation and omniwise distillation.** We compare our omniwise distillation strategy with three pairwise variants on MSR-VTT, VGGSound-S, and DiDeMo. In each case, the best performance is shown in bold. Results are averaged over three random seeds.

Datasets	Queries	Avg. R@1 ↑				Avg. R@5 ↑			
		HoPA	3-pair	T-bind	V-bind	**HoPA**	3-pair	T-bind	V-bind
MSR-VTT	100	**21.78**	20.21	20.42	21.05	**42.43**	42.08	41.45	41.97
	200	**21.93**	19.94	20.49	19.24	**42.73**	42.40	41.35	42.28
	500	**22.25**	18.11	20.19	19.85	**42.96**	41.60	38.65	41.86
VGGSound-S	100	**13.19**	12.78	12.10	12.92	34.89	**35.20**	34.01	35.00
	200	**13.15**	12.31	11.30	12.74	35.33	34.90	33.76	**35.51**
	500	**13.18**	12.50	10.74	12.43	**35.85**	35.22	32.93	34.78
DiDeMo	100	**18.35**	16.30	16.04	16.98	**37.21**	36.02	35.08	36.83
	200	**18.61**	16.09	15.81	15.73	**37.64**	36.22	33.83	35.24
	500	**18.46**	14.92	15.81	15.81	**37.76**	34.36	33.83	33.83

a particularly large effect on MSR-VTT and VGGSound-S, while the instance-level loss is especially important on DiDeMo. Similarity mining provides smaller but consistent gains on top of the two core objectives.

D.4 Effect of proxy rank

A core premise of HoPA is that the perfect omnimodal alignment corresponds to a rank-1 Gram matrix (*a.k.a.*, *Rank-1*). To verify this empirically, we investigate whether preserving more information from SVD decomposition improves distillation performance. Specifically, we modify the proxy definition to include the top-2 singular vectors instead of just the leading one (*a.k.a.*, *Rank-2*). The results are reported in Table 10. We observe a consistent performance drop when increasing the proxy rank from 1 to 2 across all datasets. This indicates that the leading singular direction ($\mathbf{v}_1$) captures the consensus across modalities, while the subsequent directions capture orthogonal components, typically corresponding to modality-specific details or noises.

D.5 Efficiency analysis of SVD operation

A potential concern with our proposed method is the computational cost introduced by the SVD operation. To provide a quantitative assessment, we benchmark the training and distillation time of HoPA against a baseline where SVD is replaced by computationally negligible tensor manipulations (*i.e.*, identity mapping and mean pooling) while maintaining the same gradient graph structure. The experiments are conducted on the MSR-VTT dataset with the same hyperparameters used in our main results. As shown in Table 11, the inclusion of SVD introduces a negligible latency (+3.5%) during the buffer training phase. In the distillation phase, the overhead increases to 39.1%. This is attributed to the TESLA (Cui et al., 2023) algorithm we use in practice to reduce the memory consumption, which back-propagates through the SVD operation involving high-order derivative calculations.

D.6 Visualization

Figure 6 visualize distilled omnimodal samples on VGGSound-S dataset using Imagebind, including a comparison with original instance and additional synthetic examples. We provide qualitative visu-

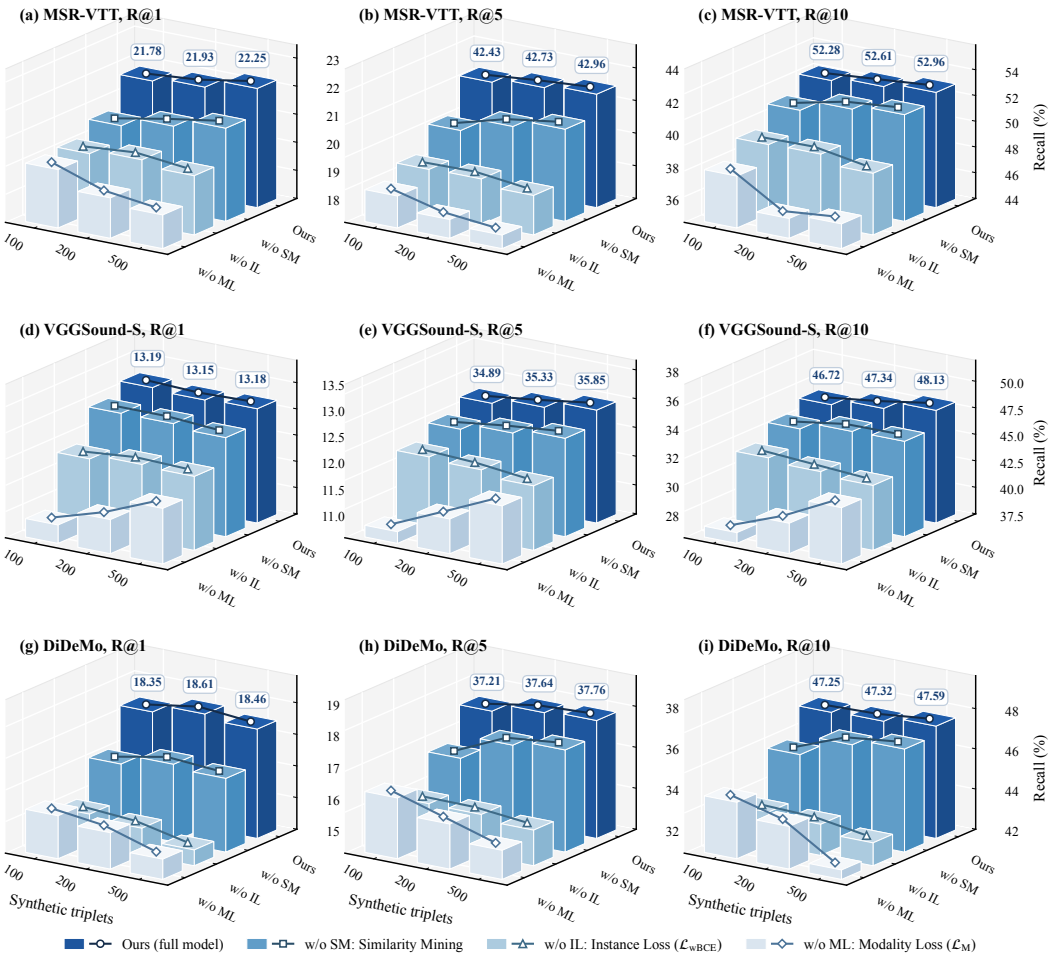

Figure 5: **Complete ablation study results.** We report Recall@1, Recall@5, and Recall@10 (%) on MSR-VTT, VGGSound-S, and DiDeMo across varying synthetic sizes. Results are averaged over three random seeds.

Table 10: **Effect of the proxy rank on distillation performance.** We compare *Rank-1* (Ours) against *Rank-2* across different metrics. In each case, the best result is shown in bold. As can be seen, *Rank-1* consistently outperforms *Rank-2*. Results are averaged over three random seeds.

Datasets	Triplets	R@1 ↑		R@5 ↑		R@10 ↑	
		Rank-1	*Rank-2*	***Rank-1***	*Rank-2*	***Rank-1***	*Rank-2*
MSR-VTT	100	**21.78**	20.80	**42.43**	39.99	**52.28**	49.77
	200	**21.93**	21.09	**42.73**	39.81	**52.61**	49.37
	500	**22.25**	20.48	**42.96**	39.78	**52.96**	49.57
VGGSound-S	100	**13.19**	12.18	**34.89**	32.85	**46.72**	44.13
	200	**13.15**	12.25	**35.33**	32.58	**47.34**	43.72
	500	**13.18**	12.44	**35.85**	32.54	**48.13**	43.34
DiDeMo	100	**18.35**	16.67	**37.21**	35.13	**47.25**	44.54
	200	**18.61**	17.01	**37.64**	35.52	**47.32**	45.21
	500	**18.46**	16.80	**37.76**	35.25	**47.59**	44.36

Table 11: **Computational efficiency analysis** by comparing the runtime with and without the SVD on MSR-VTT. The distillation time is measured per 100 iterations.

Stage	**w/o SVD**	**w/ SVD**	**SVD Ratio**
Buffer Training	10.85	11.25	3.5%
Distillation	53.28	87.48	39.1%

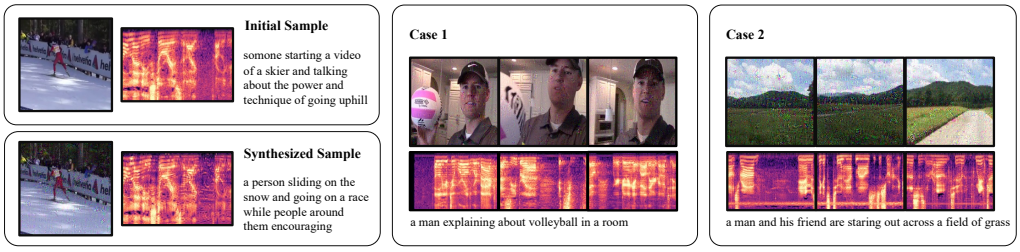

Figure 6: **The visualization of distilled omnimodal data on ImageBind with the VGGSound-S dataset.** Comparison between samples before and after distillation, shown with one representative video frame, one audio map, and the corresponding text (*left*). Illustrative cases, each visualized with three video frames, three audio maps, and the corresponding text description (*right*).

alization for omnimodal dataset distillation on ImageBind using the VGGSound-S dataset. Figure 6 shows both a representative comparison between an original instance and its distilled counterpart on the left, and additional distilled cases with richer visualization on the right. For clarity, the comparison view uses one representative video frame and one audio map, while each case visualization contains three video frames and three audio maps. Similar to prior observations in dataset distillation, the synthetic visual pattern shows a DeepDream-style appearance (Zeiler & Fergus, 2014), with enhanced high-frequency details and stylized textures. Since the text modality is represented as embeddings, we visualize each distilled text embedding by retrieving its nearest sentence from the training set, following prior work (Xu et al., 2024; Zhang et al., 2025b). The examples suggest that HoPA preserves the core event semantics and cross-modal correspondence across video, audio, and language in a compact synthetic form. More examples are presented in Appendix D.6.

We provide visualizations of $N = 100$ distilled video-audio-text instances from the VGGSound-S dataset in Figure 7 before and after distillation. Following (Wu et al., 2023), the texts we present are retrieved by the closest matching captions in the train set to the distilled text embeddings.

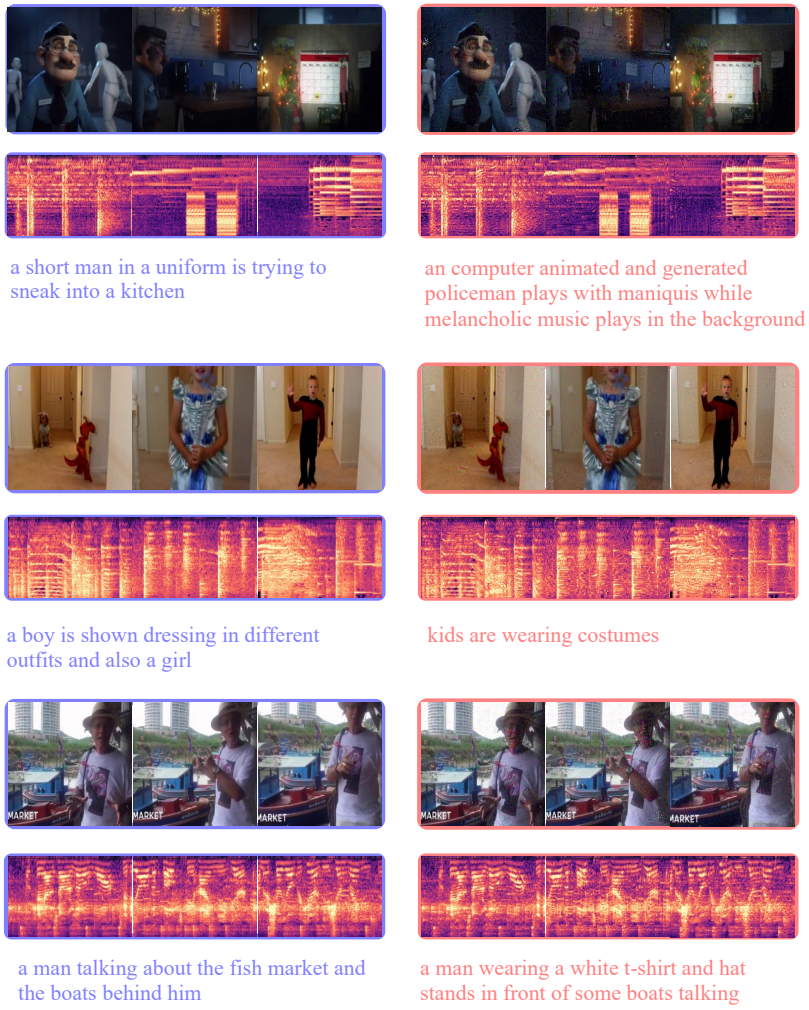

Figure 7: Examples of distilled instances from the VGGSound-S dataset with $N = 100$ triplets.

E LIMITATIONS

Although the proposed method is more efficient than existing multimodal baselines by avoiding exhaustive pairwise modality interactions, it still incurs non-negligible computational overhead due to the use of SVD for constructing the omnimodal proxy. In particular, SVD can become a bottleneck when scaling to very large datasets or a high number of modalities. Further improving the efficiency of the low-rank approximation (Allen-Zhu & Li, 2016), for example, through incremental, randomized, or approximate decomposition techniques, is an important direction for future work.