Fragmented Text Is Insufficient for Image Representation: Fine-Grained Correspondence in Multimodal Dataset Distillation

Jingwei Fang[1], Yaxin Hou[2], Bo Han[2], Xu Zhang[2], Hui Liu[3], Junhui Hou[4], and Yuheng Jia[2,3,5]

[1] College of Software Engineering, Southeast University
[2] School of Computer Science and Engineering, Southeast University
[3] School of Computing and Information Sciences, Saint Francis University
[4] Department of Computer Science, City University of Hong Kong
[5] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education
Correspondence to: Yuheng Jia (Email: yhjia@seu.edu.cn)

Abstract. Multimodal dataset distillation (MDD) aims to synthesize a compact image-text dataset on which a visual-language model can be trained to achieve performance comparable to that trained on the original, full-scale dataset. While previous methods have achieved promising results by using trajectory matching, they fail to consider the gap between single-semantic texts and multi-semantic images. Additionally, they usually use contrastive losses to pull together paired samples and penalize non-paired samples, ignoring the textual information from unpaired samples that are semantically related to the corresponding images. To reduce the semantic gap, we introduce a multi-text fusion module that strengthens the cross-modal interaction between the image and text. In addition, we leverage uncertainty to adaptively guide the contribution of non-paired samples in the synthetic dataset, thereby improving the effective utilization of information from these samples. We then provide an information theoretic analysis of the semantic gap and the limitations of contrastive supervision. Experiments on Flickr-30K and MS-COCO show that our method consistently outperforms previous state-of-the-art MDD methods, achieving significant improvements in retrieval performance (e.g., +**6.5%** in IR@10 and +**4.6%** in TR@10 in a 500-pair setting). Code is available at https://github.com/yiqiqiandefanhua/P2DE.

Keywords: Multimodal · Dataset distillation · Information theoretic

1 Introduction

Recent large language models [3, 18, 26, 39] have achieved exceptional performance in tackling complex tasks. These successes are largely dependent on the availability of vast datasets. Collecting and processing such data are both time-consuming and computationally expensive. To mitigate this issue, Dataset Distillation (DD) [45] has emerged as a solution. DD synthesizes a compact dataset

Fig. 1: (a) The three fragmented single-sentence captions from the real dataset are fused into a single multi-semantic caption to mitigate the mismatch in cross-modal semantics distributions. (b) Yellow denotes the anchor caption; blue denotes dissimilar non-paired captions; green denotes non-paired captions that are partially similar to the anchor.

that approximates the performance of the original, thereby reducing both storage and computational costs.

Significant progress has recently been made in distillation techniques for unimodal datasets, including images [13,51,56], text [24,28–30], and videos [7,23,46], with these advances gradually extending to multimodal scenarios, where multimodal dataset distillation (MDD) [49] aims to reduce dataset size while preserving semantic and relational information. The pioneering work on MDD [49] introduces a dual-trajectory matching framework, which jointly optimizes the image and text encoders under a contrastive objective. Building upon this, LoRS [50] improves distillation efficiency by low-rank similarity mining, where a similarity matrix between synthetic image-text pairs is constructed to capture their global semantic relationships. Recently, RepBlend [54] alleviates modality collapse and enhances intra-modality diversity through representation blending and symmetric projection trajectory matching.

However, all of these methods overlook two key issues. First, in real datasets, each image is typically paired with multiple captions, whereas in synthetic datasets, each image is often paired with a single caption. This asymmetry induces an inherent semantic gap between the paired text and image modalities, which makes cross-modal alignment more challenging. As illustrated in Fig. 1 (a), a single sentence caption captures only fragmented aspects of the image. Second, existing methods [49, 50, 54] rely on contrastive losses (e.g., InfoNCE [33], WBCE [50]) to strengthen positive pairs while penalizing unpaired ones. However, this contrastive scheme treats all non-paired samples as negatives, even though some non-paired texts contain fragmented semantics of the image. In Fig. 1 (b), we can observe that there is some semantic overlap between the images and certain non-paired texts. The above limitations indicate that describing an image using only fragmented text is insufficient.

Recognizing these limitations, we propose **P**aired-and-non-**P**aired **D**escription **E**nhancement (P^{2}DE), a novel MDD method aimed at enhancing text-based descriptions of images. First, we propose an Intra-pair Text Fusion (ITF) strategy. Specifically, under image supervision, we use a large vision–language model (e.g., LLaVA [27]) to merge multiple textual descriptions of the same image into a semantically richer caption (see Fig. 1 (a)). This enriches and smooths the semantic distribution of text modalities, reducing the structural differences

between text and image modalities. Moreover, to overcome the limitation of conventional contrastive learning that treats all non-paired samples as negatives, P^{2}DE introduces an Inter-pair Contribution Controller (ICC) to adjust the learning state of the student model by modifying the information contribution from non-paired texts using conditional information entropy. It enables deeper exploration of informative from unpaired samples. By increasing the utilization of non-paired sample information, the synthetic dataset can better approximate the real dataset.

Additionally, we explore the essence of these issues from an information theoretic perspective and demonstrate the validity of ITF and ICC. Information theory [16] provides a framework for quantifying the transmission information. In multimodal data, images typically have higher dimension and smoother entropy distribution, representing broad and continuous semantics. In contrast, text often exhibits more concentrated entropy distributions and semantics. This entropy discrepancy leads to different amounts of information between image and text representations. When we fuse multiple text descriptions for distillation, we increase the upper bound of mutual information, thereby enhancing cross-modal mapping. Additionally, conditional information entropy can be used to measure the uncertainty of one modality given the other. If the conditional information entropy is very low, the information from paired samples dominates the image, leaving minimal contribution from non-paired texts in optimizing the image. By using conditional information entropy to describe the learning state of an image with respect to non-paired but similar texts, we enhance the exploration of semantic information in non-paired samples, thereby achieving finer-grained alignment.

Our contributions are summarized as follows:

- Our proposed P^{2}DE achieves finer-grained alignment between image and text modalities through two synergistic components: an Intra-pair Text Fusion strategy designed to bridge the semantic gap, and an Inter-pair Contribution Controller that optimizes the non-paired samples information utilization.
- From an information-theoretic perspective, we reveal the underlying mechanisms behind the modal semantic gap and the neglect of non-paired information, thereby theoretically justifying the validity of our proposed method.
- Extensive experiments on multiple datasets demonstrate the superior performance and efficiency of our approach. Specifically, on MS-COCO [25], our method achieves notable accuracy improvements, with a 6.5% increase in IR@10 for 100 pairs and a 4.6% increase in TR@10 for 500 pairs, compared to SOTA methods.

2 Related Work

Dataset Distillation. Dataset distillation (DD) [45] condenses the essential information of a large-scale dataset into a small, synthetic dataset, enabling models to achieve comparable performance with drastically reduced data. Existing

methods can be broadly categorized by their optimization strategies, including: kernel-based [31, 32, 58], distribution matching [43, 44, 53, 55], trajectory matching [4, 5, 8, 9, 13], generative modeling [12, 40], and decoupled distillation [41, 51]. While these methods have demonstrated remarkable success in single-modal settings, extending them to multimodal learning poses a significant challenge. The core difficulty lies in achieving efficient information compression within each modality. At the same time, we must model the complex semantic associations and interactions across modalities.

Image-Text Retrieval. Image-text contrastive learning (ITC) is a foundation of multimodal learning. Early works like DeViSE [10] and Deep CCA [1] established the foundation of joint embedding frameworks. The emergence of dual-encoder models like CLIP [35] and ALIGN [17] revolutionized the field, enabling efficient zero-shot retrieval. The challenges of noisy web data motivated fusion-based pre-training frameworks, such as ALBEF [22], BLIP [21], and CoCa [52], which combine contrastive learning with language modeling to better handle data noise.

Vision-Language Dataset Distillation. MDD, primarily studied in vision-language settings, builds upon the ITC framework and aims to synthesize a compact set of image-text pairs that preserve the semantic alignment and training dynamics of large-scale vision-language datasets. MTT-VL [49] modifies the MTT [4] framework in a simple way to apply image-text pair data, independently matching the parameters of the image and text encoders. LoRS [50] effectively captures cross-modal correspondence by jointly extracting the true similarity matrix between image-text pairs, further enhancing performance. RepBlend [54] introduces representation blending to enhance intra-modality representation diversity and employs symmetric projection trajectory matching to improve cross-modal alignment. CovMatch [20] aligns the cross-covariances of real and synthetic features to improve cross-modal alignment. EDGE [57] leverages a generative distillation framework to improve the efficiency.

3 Preliminaries

Image-Text Contrastive Learning. CLIP [35] is the foundational architecture for ITC. Given a dataset of paired image-text samples $\mathcal{T} = \{(x_v^i, x_l^i)\}_{i=1}^M$, where x_v^i and x_l^i denote the i-th image and text inputs, the goal is to train an image encoder $f_v : \mathbb{R}^v \to \mathbb{R}^{d_v}$ and a text encoder $f_l : \mathbb{R}^l \to \mathbb{R}^{d_l}$, each followed by a trainable linear projection layer, $G_v \in \mathbb{R}^{z \times d_v}$ and $G_l \in \mathbb{R}^{z \times d_l}$. These projection layers map the modality-specific embeddings into a shared space $\mathbb{R}^z$ that captures semantic similarity. For a given pair (x_v, x_l), the encoders produce intermediate representations $h_v = f_v(x_v; \theta_v)$ and $h_l = f_l(x_l; \theta_l)$, which are then projected as $z_v = G_v h_v$ and $z_l = G_l h_l$. The similarity between image x_v^i and text x_l^j is computed using cosine similarity, $s_{ij} = \cos(z_v^i, z_l^j) = \langle z_v^i, z_l^j \rangle / \|z_v^i\| \|z_l^j\|$ in the shared embedding. The model is trained using the InfoNCE loss [33] with

temperature $\tau > 0$:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(s_{ii}/\tau)}{\sum_{j\neq i}\exp(s_{ij}/\tau)} + \log\frac{\exp(s_{ii}/\tau)}{\sum_{j\neq i}\exp(s_{ji}/\tau)}\right].$$

Bi-Trajectory Matching. MDD builds upon a dual-trajectory matching mechanism, which is inspired by the concept of symmetric trajectory matching in MTT-VL [49]. This method aligns the student and teacher models and enforces consistency in the parameter evolution trajectories of the image and text encoders. Formally, given the expert model parameters $\theta^*_{V,t}$ and $\theta^*_{T,t}$ for the image and text encoders at step t, and the corresponding student parameters $\hat{\theta}_{V,t}$ and $\hat{\theta}_{T,t}$, the dual-trajectory matching loss is defined as:

$$\ell_{\text{trajectory}} = \frac{\|\hat{\theta}_{V,t+\hat{N}} - \theta^*_{V,t+M}\|_2^2}{\|\theta^*_{V,t} - \theta^*_{V,t+M}\|_2^2} + \frac{\|\hat{\theta}_{T,t+\hat{N}} - \theta^*_{T,t+M}\|_2^2}{\|\theta^*_{T,t} - \theta^*_{T,t+M}\|_2^2}, \tag{1}$$

where, M and $\hat{N}$ denote the number of parameter update steps for the expert and student models, respectively. The numerator measures the squared Euclidean deviation between the student's updated parameters and the expert's corresponding parameters after M steps, reflecting how closely the student follows the expert's optimization path. The denominator normalizes this distance by the expert's own parameter displacement $\|\theta^*_{I,t} - \theta^*_{V,t+M}\|_2^2$, ensuring scale invariance across different training stages.

Fundamentals of Information Theory. We introduce the foundational concepts of information theory, which will be later applied to the image and text modalities.

The entropy $H(X)$ describes the uncertainty of a random variable X, which is defined as

$$H(X) = -\sum_{x} p(x)\log p(x). \tag{2}$$

Higher entropy implies greater uncertainty and a broader distribution of outcomes, whereas lower entropy corresponds to more certainty and concentration.

The mutual information (MI) between two random variables X and T quantifies the reduction in uncertainty of one given the other, defined as:

$$I(X;T) = H(X) + H(T) - H(X,T), \tag{3}$$

where $H(X,T)$ is the joint entropy of X and T.

The conditional entropy $H(X|T)$ quantifies the remaining uncertainty of X given T, where a lower value implies that X is more predictable conditioned on T. It is defined as:

$$H(X|T) = H(X,T) - H(T). \tag{4}$$

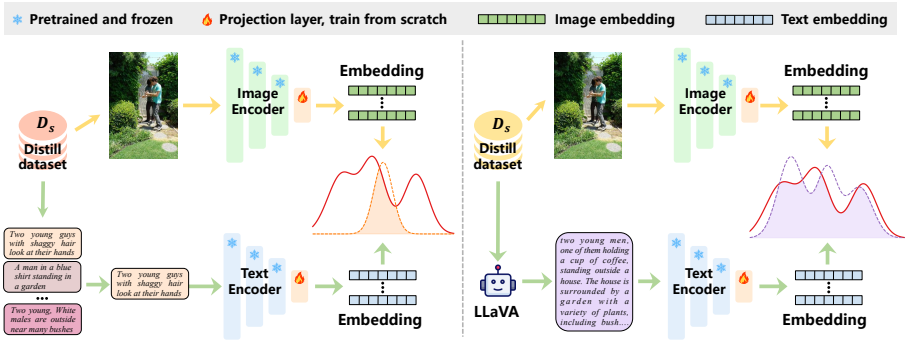

Fig. 2: Left: Existing distillation methods select a single text from the multiple texts in the real dataset and align it with the image in the feature space. **Right**: By leveraging LLaVA [27] to fuse text, the model can represent multiple semantic fragments, facilitating smoother alignment with the image's multi-semantic representation.

4 Proposed Method

In this section, we analyze the limitations of existing MDD methods and introduce our simple yet effective solution, P^2DE, to address these challenges. Specifically, we present two orthogonal modules in separate subsections: Intra-pair Text Fusion, which strengthens semantic consistency within paired samples, and the Inter-pair Contribution Controller, which aims to maximize information capture from outside the paired samples. Furthermore, from an information-theoretic perspective we elucidate the nature of the observed phenomena and validate the effectiveness of our method through empirical analyses. The pseudo-code is detailed in Appendix H.

4.1 Intra-pair Text Fusion

In MDD, dual trajectory matching [49] is commonly used to jointly optimize the image and text modalities. It aligns the synthetic data with the real data by minimizing the discrepancy between the parameter update trajectories of the expert and student encoders over the next M steps.

However, MDD methods such as MTT-VL [49], LORS [50], and RepBlend [54] share a common limitation: they fail to adequately account for the intrinsic characteristics of multimodal datasets. In multimodal datasets such as Flickr-30k [34] and MS-COCO [25], each image is accompanied by multiple descriptive captions that portray the scene from different perspectives. These descriptions collectively form a comprehensive semantic representation. The full set of image–text pairs is used to jointly train the expert trajectory. However, as shown in Fig. 2 (left), when training the student trajectory, this multi-view information is often compressed to a single caption for each image. This compression introduces semantic mismatch during student model training and further complicates the alignment between the student and expert trajectories, posing a significant challenge for effective cross-modal alignment.

A single image can convey complex and diverse concepts more effectively than a mere verbal description. To mitigate this limitation, we enrich the text modality by expanding its semantic coverage. As shown in Fig. 2 (right), our method leverages a large vision–language model (LLaVA [27]) to fuse multiple textual descriptions corresponding to the same image. Specifically, we use LLaVA in inference-only mode and perform Image-Text Fusion (ITF) for each image I with its associated captions $T = \{t_i\}$. To ensure that the generated text representation remains faithful to the visual content, we use the original image as a supervisory signal during fusion, enforcing alignment through explicit visual conditioning and strict prompt constraints. A representative prompt is as follows: Prompt example: *"Fuse these captions into ONE caption for the SAME image. Rules: Be faithful to the image. Do NOT invent details that are not visible. Merge complementary details; remove redundancy; write one compact paragraph."* This design prevents semantic drift and hallucination while encouraging complementary information integration across captions. Beyond standard decoding, ITF is an offline preprocessing step that introduces no trainable parameters or fusion-specific tunable hyperparameters. It only changes the input textual content: the fused captions are encoded by the same text encoder and projected into the same shared embedding space as the original captions. Therefore, the text feature dimensionality and the distillation architecture remain unchanged.

For example, the raw captions may describe different perspectives of the image (e.g., "a dog sitting", "a person walking a dog", "two people skateboarding"). After large-model fusion, the resulting text contain all of this information and further includes additional fine-grained details, yielding a more comprehensive description. The fused text and the image are encoded by their respective encoders and subsequently aligned within a shared embedding space. By simulating the training process of a real dataset, we obtain more precise trajectory matching Eq. (1). Intra-pair Text Fusion achieves a more fine-grained mode alignment and improves the accuracy of MMD.

4.2 Inter-pair Contribution Controller

Current MDD methods generally rely on contrastive losses, whose core mechanism is to pull paired samples closer while indiscriminately treating all unpaired samples as negatives and pushing them away. Our observations in Fig. 4 (RepBlend) confirm this: all unpaired samples are given nearly equal contributions. However, this approach has a fundamental flaw, as it overlooks the fact that many unpaired texts contain valuable semantic information relevant to the paired image, which should be utilized but is ignored.

In Fig. 1(b), the image-text pair on the left serves as the anchor pair and the four pairs on the right are considered negative samples with respect to this anchor. The blue text at the top describes content that is semantically distant from the anchor image and should ideally be pushed infinitely away from the anchor. However, the images at the bottom share strong semantic similarities with the anchor, with their textual descriptions containing terms such as "dog" and "running", which also apply to the anchor image. In real large scale datasets,

the impact of this shared information is almost negligible due to the high redundancy of the data itself. However, in synthetic datasets with a very small capacity, every piece of semantic information is crucial, making maximizing information utilization the core challenge.

To effectively leverage the overlooked semantic information in unpaired texts, we propose an Inter-pair Contribution Controller (ICC) that dynamically adjusts the contribution of unpaired texts to the anchor images within the synthetic dataset. The controller uses conditional information entropy (defined in Eq. (4)) as the core metric to measure the uncertainty of the text modality given the image modality. Specifically, ICC converts the cross-modal similarity s_{ij} into an image-conditioned probability p_{ij} and uses the resulting entropy to softly quantify the contribution of each text. Thus, partially relevant non-paired texts can provide non-negligible learning signals, whereas semantically distant texts contribute little. By monitoring the changes in conditional information entropy in the student model, ICC dynamically adjusts the learning process and reassigns the contribution weights of paired and unpaired texts. The computation details are provided in Appendix D.

Specifically, the way we adjust the learning process of the student model is as follows:

$$S_{t+1} = \begin{cases} \min(S_t + \alpha, S_{\max}), & \Delta H_t > \varepsilon, \\ \max(S_t - \beta, S_{\min}), & \Delta H_t \leq \varepsilon, \end{cases} \tag{5}$$

where S_t denotes the learning step of the student model, and ε is a constant threshold. $\Delta H_t = H_t - H_{t-1}$ denotes the change in conditional information entropy between two steps (from $t-1$ to t). When ΔH is large, it indicates that the model is still absorbing information from non-paired samples, and we expand the trajectory step size. When ΔH converges and becomes small, we decrease the step size to suppress the excessive concentration of information from paired samples. By dynamically controlling the learning trajectory of the student model, our method guides it to avoid over-reliance on paired texts and effectively integrates supplementary information from unpaired texts. This process leads to a more fine-grained cross-modal alignment.

4.3 Theoretical Analysis

Here, we theoretically explain why our P^{2}DE effectively mitigates the semantic gap and leverages information from unpaired samples. **Intra-pair Text Fusion (ITF).** In information theory, entropy (Eq. (2)) measures semantic uncertainty and coverage. In trajectory matching, the image modality X naturally exhibits high entropy because it contains various objects and relations, leading to a broader and more uniform semantic distribution. By contrast, the text modality T is

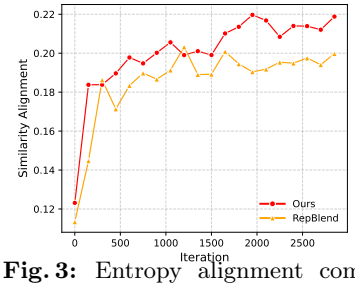

Fig. 3: Entropy alignment comparison under the same distillation setting.

restricted to a single semantic text, which covers only a very narrow region of the image's semantic space and thus yields a lower-entropy, more concentrated distribution. This entropy asymmetry constitutes a cross-modal information bottleneck that directly constrains the mutual information (MI) between modalities. Formally, as defined in Eq. (3), MI quantifies the amount of shared information between two variables, and its attainable upper bound is determined by the modality with lower entropy, i.e., $I(X;T) \leq \min\big(H(X), H(T)\big)$. Due to the semantic gap between modalities, $H(T) \ll H(X)$, so this upper bound is substantially compressed, which fundamentally exacerbates cross-modal alignment. Through ITF, we expand the semantic space of the text modality, increasing $H(T)$, thereby raising $I(X;T)$ and ultimately improving cross-modal alignment.

We further empirically verify the effectiveness of ITF by quantifying the distributional similarity between the image and text modalities during distillation. As shown on the Fig. 3, our method (red) achieves a stronger alignment throughout training, whereas RepBlend [54] (yellow) exhibits limited improvement and converges to a lower alignment level. By achieving better distributional alignment, ITF fosters more consistent cross-modal alignment. The computation details for distributional similarity are provided in Appendix C.

Inter-pair Contribution Controller (ICC). For clarity, we first provide the relevant definitions and assumptions before presenting the formal theorem. Let $y_{ij} \in \{0, 1\}$ indicate whether (i, j) is a paired sample ($y_{ij} = 1$) or a non-paired sample ($y_{ij} = 0$). Define the WBCE loss with separate averaging over paired (p) sample and non-paired (np) sample as:

$$L = \frac{1}{|\mathrm{p}|} \sum_{(i,j)\in \mathrm{p}} \ell_{ij} + \frac{1}{|\mathrm{np}|} \sum_{(i,j)\in \mathrm{np}} \ell_{ij},$$

where $\ell(\cdot, \cdot)$ refers to the binary cross-entropy loss. Define $z_{ij} = s_{ij}/\gamma$, with $\gamma > 0$ representing the temperature. Therefore, we have the following three theorems.

Theorem 1 (WBCE-Induced Gradient Vanishing). *If the training enters a diagonal-dominant regime for each pair (i, j), the gradients will behave as follows: $\frac{\partial L}{\partial s_p} \to 0^-$, $\frac{\partial L}{\partial s_{np}} \to 0^+$. Moreover, the gradient magnitude for each negative pair scales as $\left|\frac{\partial L}{\partial s_{ij}}\right| = \mathcal{O}\left(\frac{\sigma(z_{ij})}{\gamma|\mathrm{np}|}\right)$ which decays to zero and is further diluted by the factor $1/|\mathrm{np}|$ ($num : |\mathrm{np}| \gg |\mathrm{p}|$).*

Remark 1. Existing methods use contrastive loss functions to constraint the synthetic dataset. Theorem 1 indicates, the excessive concentration of similarity reduces the gradient of non-paired samples. As a result, the image fails to learn sufficient descriptive information from non-paired texts, limiting the potential of the distillation process.

Theorem 2 (Similarity Gap–Entropy Concentration). *The conditional probability p_{ij} between image i and text j is defined as:*

$$p_{ij} = \frac{\exp(s_{ij}/\tau)}{\sum_{k=1}^{m} \exp(s_{ik}/\tau)},$$

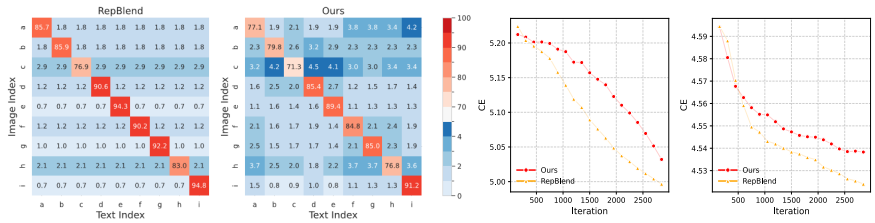

Fig. 4: Left: Heatmaps of similarity matrices. The first heatmap shows the similarity matrix produced by RepBlend, while the second presents the matrix after modulation with conditional information entropy. ICC mitigates information concentration within paired samples and encourages a balanced utilization of non-paired textual information. **Right:** Evolution of conditional information entropy (CE) over training iterations. The two curves correspond to Flickr-30K [34] and MS-COCO [25], respectively. Compared with RepBlend, our method maintains consistently higher entropy, indicating improved information distribution and more effective cross-modal interaction during training.

where s_{ij} denotes the similarity between image i and text j, and $\Delta_i = \hat{s}_{ii} - \max_{j \neq i} \hat{s}_{ij}$ represents the maximum similarity difference between the paired sample and the non-paired samples. It follows that:

$$P_{ii} \geq \frac{1}{1 + (m-1)e^{-\Delta_i/\tau}}.$$

As Δ_i increases, the similarity of the paired sample p_{ii} becomes dominant, which leads to a significant decrease in the conditional information entropy.

Remark 2. Theorem 2 highlights how an increasing similarity gap between paired and non-paired samples leads to an imbalance in the concentration of information, causing a rapid reduction in the conditional information entropy. This emphasizes the challenge of maintaining diverse and informative gradients across paired and non-paired data.

Theorem 3 (Entropy Bound $\Rightarrow$ Non-Paired Gradients). *We define the average entropy ($\bar{h}_j^-$) for the non-paired samples as:*

$$\bar{h}_j^- \triangleq \frac{1}{|\mathcal{N}_j|} \sum_{j \in \mathcal{N}_j} h_{ij}, \quad \mathcal{N}_i = \{j : y_{ij} = 0\},$$

where $\mathcal{N}_i$ represents the set of non-paired samples, h_{ij} is the information entropy between image j and text i. If there exists a threshold $h_\star > 0$ such that $\bar{h}_i^- \geq h_\star$, then there exists a constant $\delta(h_\star) \in (0, \frac{1}{2}]$, for which the following holds:

$$\sum_{j \in \mathcal{N}_i} \left| \frac{\partial L}{\partial s_{ij}} \right| \geq \frac{\delta(h_\star)}{\gamma}.$$

Remark 3. Theorem 3 establishes that, provided a minimum uncertainty is maintained for non-paired samples, the model is constrained to retain a non-zero probability mass at non-diagonal positions. This ensures a small lower bound on the gradients for non-paired samples, preventing the learning signals from diminishing.

As shown in Fig. 4, the conditional entropy of Flickr-30k [34] and MS-COCO [25] decreases as training progresses. Compared to RepBlend [54] (yellow), our method (red) prevents the conditional information entropy from excessively collapsing, thereby increasing the uncertainty of the text modality and providing stronger learning signals for unpaired texts. This difference can be more clearly observed by comparing the similarity matrix heatmaps, where the similarity values of paired samples decrease significantly, while those of unpaired samples increase to varying degrees. Detailed proofs of these theorems can be found in Appendix A.

5 Experiments

Table 1: Results on MS-COCO [25]. Both distillation and validation are performed using NFNet+BERT. The model trained on the full dataset performs: IR@1=14.6, IR@5=38.9, IR@10=53.2; TR@1=20.6, TR@5=46.8, TR@10=61.3.

Pairs	Ratio	Metric	Coreset Selection				Dataset Distillation					
			Rand –	Herd [47] ICML'09	K-Cent [48] IJGIS'11	Forget [42] ICLR'19	MTT-VL [49] TMLR'24	TESLA-VL [50] ICML'24	LoRS [50] ICML'24	CovMatch [20] NeurIPS'25	RepBlend [54] NeurIPS'25	Ours –
100	0.8‰	IR@1	0.3	0.5	0.4	0.3	1.3±0.1	0.3±0.2	1.8±0.1	2.8±0.1	4.1±0.3	**5.9**±0.4
		IR@5	1.3	1.4	1.4	1.5	5.4±0.3	1.0±0.4	7.1±0.2	10.5±0.2	13.9±0.8	**18.7**±1.0
		IR@10	2.7	3.5	2.5	2.5	9.5±0.5	1.8±0.5	12.2±0.2	17.7±0.3	22.3±0.5	**28.8**±1.3
		TR@1	0.8	0.8	1.4	0.7	2.5±0.3	2.0±0.2	3.3±0.2	3.8±0.1	5.2±0.5	**6.7**±0.3
		TR@5	3.0	2.1	3.7	2.6	10.0±0.5	7.7±0.5	12.2±0.3	13.1±0.3	17.9±0.9	**21.6**±0.8
		TR@10	5.0	4.9	5.5	4.8	15.7±0.4	13.5±0.3	19.6±0.3	21.1±0.2	28.0±0.3	**32.2**±0.7
200	1.7‰	IR@1	0.6	0.9	0.7	0.6	1.7±0.1	0.1±0.1	2.4±0.1	3.8±0.1	6.1±0.8	**6.8**±0.4
		IR@5	2.3	2.4	2.1	2.8	6.5±0.4	0.2±0.1	9.3±0.2	13.4±0.1	19.3±0.7	**21.4**±1.0
		IR@10	4.4	4.1	5.8	4.9	12.3±0.8	0.5±0.1	15.5±0.2	21.8±±0.2	29.8±0.5	**32.4**±1.2
		TR@1	1.0	1.0	1.2	1.1	3.3±0.2	0.7±0.2	4.3±0.1	5.3±0.2	6.9±0.6	**8.2**±0.4
		TR@5	4.0	3.6	3.8	3.5	11.9±0.6	3.1±0.5	14.2±0.3	17.3±0.2	21.8±0.9	**24.6**±0.7
		TR@10	7.2	7.7	7.5	7.0	19.4±1.2	5.3±0.8	22.6±0.2	27.0±0.2	32.3±0.7	**36.4**±0.8
500	4.4‰	IR@1	1.1	1.7	1.1	0.8	2.5±0.5	0.8±0.2	2.8±0.2	5.4±0.1	6.2±0.1	**7.2**±0.4
		IR@5	5.0	5.3	6.3	5.8	8.9±0.7	3.6±0.6	9.9±0.5	18.0±0.1	19.9±0.3	**22.2**±1.0
		IR@10	8.7	9.9	10.5	8.2	15.8±1.5	6.7±0.9	16.5±0.7	28.2±0.1	30.6±0.1	**33.4**±1.4
		TR@1	1.9	1.9	2.5	2.1	5.0±0.4	1.7±0.4	5.3±0.5	8.1±0.3	7.0±0.2	**8.7**±0.4
		TR@5	7.5	7.8	8.7	8.2	17.2±1.3	5.9±0.8	18.3±1.5	23.5±0.3	22.0±0.3	**25.5**±0.7
		TR@10	12.5	13.7	14.3	13.0	26.0±1.9	10.2±1.0	27.9±1.4	34.6±0.6	32.9±0.6	**37.5**±0.6

In this section, we perform extensive experiments on multiple benchmark datasets to demonstrate the advantage of the proposed P^{2}DE framework. We begin by describing the experimental setup, including the datasets, baseline methods, and implementation details. The main results are reported in Tab. 1 and Tab. 2, followed by detailed ablation studies that investigate the contribution of each component.

5.1 Experimental Setup

Baseline methods. To comprehensively evaluate our proposed method, we conduct extensive comparisons with a range of representative approaches in MDD and core-set selection. The comparison includes recent leading approaches such as MTT-VL [49], TESLA-VL [50], LoRS [50], and RepBlend [54], as well

Table 2: Results on Flickr-30k [34]. Both distillation and validation are performed using NFNet+BERT. The model trained on the full dataset performs: IR@1=23.16, IR@5=53.98, IR@10=66.62; TR@1=33.8, TR@5=65.7, TR@10=76.9.

Pairs	Ratio	Metric	Coreset Selection				Dataset Distillation					
			Rand –	Herd [47] ICML'09	K-Cent [48] IJGIS'11	Forget [42] ICLR'19	MTT-VL [49] TMLR'24	TESLA-VL [50] ICML'24	LoRS [50] ICML'24	CovMatch [20] NeurIPS'25	RepBlend [54] NeurIPS'25	Ours –
100	0.3%	IR@1	1.0	0.7	0.7	0.7	4.7±0.2	0.5±0.2	8.3±0.2	10.1±0.2	11.5±0.4	**13.5**±0.3
		IR@5	4.0	2.8	3.1	2.4	15.7±0.5	2.3±0.2	24.1±0.2	28.6±0.4	32.0±0.7	**36.4**±1.7
		IR@10	6.5	5.3	6.1	5.6	24.6±1.0	4.7±0.4	35.1±0.3	40.9±0.6	44.5±0.6	**49.5**±1.7
		TR@1	1.3	1.1	0.6	1.2	9.9±0.3	5.5±0.5	11.8±0.2	14.8±0.9	16.2±0.8	**18.7**±0.4
		TR@5	5.9	4.7	5.0	4.2	28.3±0.5	19.5±0.9	35.8±0.6	38.0±0.4	41.7±0.9	**46.5**±1.6
		TR@10	10.1	7.9	7.6	9.7	39.1±0.7	28.9±1.0	49.2±0.5	50.6±0.6	55.5±0.4	**61.8**±1.0
200	0.7%	IR@1	1.1	1.5	1.5	1.2	4.6±0.9	0.2±0.1	8.6±0.3	12.3±0.4	12.7±0.8	**14.8**±0.5
		IR@5	4.8	5.5	5.4	3.1	16.0±1.6	1.3±0.2	25.3±0.2	33.6±0.3	34.7±0.6	**39.4**±0.8
		IR@10	9.2	9.3	9.9	8.4	25.5±2.6	2.5±0.2	36.6±0.3	45.8±0.2	47.6±0.5	**52.2**±1.0
		TR@1	2.1	2.3	2.2	1.5	10.2±0.8	2.8±0.5	14.5±0.5	17.4±0.5	18.6±0.7	**23.0**±0.8
		TR@5	8.7	8.4	8.2	8.4	28.7±1.0	10.4±1.5	38.7±0.5	41.7±0.5	46.0±0.8	**52.3**±1.0
		TR@10	13.2	14.4	13.5	10.2	41.9±1.9	17.4±1.6	53.4±0.5	55.8±0.5	60.0±0.6	**66.9**±1.1
500	1.7%	IR@1	2.4	3.0	3.5	1.8	6.6±0.3	1.1±0.2	10.0±0.2	14.7±0.3	17.0±0.6	**17.0**±0.2
		IR@5	10.5	10.0	10.4	9.0	20.2±1.2	7.3±0.4	28.9±0.7	38.4±0.4	42.5±0.5	**43.1**±0.4
		IR@10	17.4	17.0	17.3	15.9	30.0±2.1	12.6±0.5	41.6±0.6	51.4±0.3	55.9±0.6	**56.6**±0.5
		TR@1	5.2	5.1	4.9	3.6	13.3±0.6	5.1±0.2	15.5±0.7	19.9±0.6	22.5±0.4	**22.7**±0.5
		TR@5	18.3	16.4	16.4	12.3	32.8±1.8	15.3±0.5	39.8±0.4	46.7±0.9	**53.2**±0.3	52.7±0.7
		TR@10	25.7	24.3	23.3	19.3	46.8±0.8	23.8±0.3	53.7±0.3	59.5±0.7	66.7±0.3	**67.3**±0.6

as a series of classic core-set selection techniques, including Random Selection, Herd [47], K-Cent [48], and Forget [42].

Datasets and Metrics. To ensure a fair evaluation, we conduct experiments on widely adopted multimodal benchmarks, including the vision–language datasets Flickr-30K [34] and MS-COCO [25], as well as the audio–text dataset AudioCaps [19]. Flickr-30K contains 31K images, MS-COCO includes over 330K images across 80 categories, and AudioCaps comprises approximately 49K audio clips paired with human-written captions. Performance is measured using Top-K recall (R@K) for cross-modal retrieval, where IR@K and TR@K denote text-to-image and image-to-text retrieval. Following prior work, we adopt NFNet [2], RegNet [36], and ResNet-50 [14] as image encoders and BERT [6] as the text encoder. All experiments are conducted under consistent settings, repeated three times on a single NVIDIA V100 GPU.

Implementation Details. Following prior work on vision–language dataset distillation, we implement a CLIP-style architecture using the aforementioned image and text encoders. The image encoder is initialized with ImageNet-pretrained weights, and the text encoder with the official pretrained weights of their respective language models. We collect 20 expert trajectories, each consisting of 10 training epochs. The hyperparameter settings follow those in RepBlend, as summarized in Tab. 8 and Tab. 9 of Appendix B.

5.2 Main Results

On MS-COCO [25], as shown in Tab. 1, our method outperforms all baselines across various distillation budgets (100, 200, and 500 pairs). In the low-budget setting (100 pairs), we achieve 18.7% in IR@10 and 32.2% in TR@10, which are 6 and 4 percentage points higher than RepBlend, respectively. With 200 pairs, the performance further improves to 32.4% (IR@10) and 36.4% (TR@10), with an average improvement of 3% $\sim$ 4%. Even with 500 pairs, our method shows significant improvements, with a 4.6% increase in TR@10. A similar trend is observed

on Flickr-30k [34], as shown in Tab. 2. Our method achieves 61.8% and 66.9% in TR@10 with 100 and 200 pairs, respectively, significantly outperform LoRS and RepBlend. Our method maintains the best performance across all metrics even with 500 pairs. This demonstrates the effectiveness of P^{2}DE for cross-modal retrieval tasks. To further demonstrate the cross-modal generalization ability of our approach, we conduct additional experiments on the audio–text dataset AudioCaps [19], where our method consistently outperforms RepBlend across different synthetic data scales; detailed results are provided in Appendix F.

5.3 Ablation Study

Intra-pair Text Fusion & Inter-pair Contribution Controller. We analyze the individual contributions of the Intra-pair Text Fusion (ITF) module and the Inter-pair Contribution Controller (ICC) module in our method. By selectively adjusting either ITF or ICC, we conduct an ablation study on the Flickr-30K dataset using NFNet+BERT. As shown in Tab. 3, all modules contribute positively to the retrieval metrics (IR@1/5/10 and TR@1/5/10).

Table 3: Comparison of accuracy (%) with and without the key components in the proposed method.

Ablations		100						200						Avg.
w/ ITF	w/ ICC	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	
		11.5	32.0	44.5	16.2	41.7	55.5	12.7	34.7	47.6	18.6	46.0	60.0	35.1
✓		12.6	32.4	45.5	16.4	43.1	57.3	13.1	36.0	49.3	21.7	47.6	62.8	36.5 ↑ 1.4
	✓	**13.8**	37.1	50.2	18.1	48.3	61.6	14.1	37.8	50.0	21.8	50.4	65.7	39.1 ↑ 4.0
✓	✓	**13.8**	**38.4**	**51.4**	**18.5**	**48.4**	**62.9**	**14.3**	**38.5**	**51.3**	**23.0**	**51.1**	**66.0**	**39.8**↑ 4.7

Parameter Analysis. For the Inter-pair Contribution Controller (ICC), we evaluate a variety of parameter settings on the Flickr-30k dataset [34]. The changes in average performance under different parameter configurations are provided in Appendix E. As observed, our method consistently outperforms RepBlend [54] across all settings, achieving up to a 5% performance improvement, which demonstrates the robustness of the proposed ICC module.

5.4 More Discussion

Cross-architecture generalization. We evaluate the cross-architecture generalization of P^{2}DE by distilling synthetic data with NFNet+BERT and testing on different projection heads (ResNet and RegNet for image, BERT for text). As shown in Tab. 4 on Flickr-30K (500 pairs), our method consistently outperforms others across architectures, demonstrating strong robustness, superior transferability, and stable generalization performance.

Large budget evaluation. As shown in Tab. 5, our method consistently outperforms prior approaches on both Flickr-30K and MS-COCO under the 500 and 1000 synthetic-pair settings. The performance improvements are observed

Table 4: Cross architecture generalization. The data is synthesized with NFNET+BERT and evaluated on various architectures.

Pairs	Evaluate Model	Method	Flickr-30k IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
499	ResNet+BERT	LoRS	2.0±0.1	7.6±1.6	12.4±0.4	5.6±1.3	19.1±0.5	28.3±1.4
		RepBlend	1.7±0.3	6.6±1.0	11.5±1.5	5.2±0.3	16.8±0.8	26.2±0.5
		Ours	**2.4**±0.3	**8.9**±0.7	**14.9**±1.4	**6.8**±0.2	**21.6**±0.8	**31.6**±0.7
499	RegNet+BERT	LoRS	1.4±0.1	6.2±0.7	10.3±1.7	3.5±0.9	15.2±1.5	22.7±1.2
		RepBlend	1.3±0.1	6.0±0.1	9.9±0.2	4.4±0.1	13.1±0.2	21.4±0.6
		Ours	**1.6**±0.3	**6.4**±0.6	**11.1**±1.4	**5.3**±0.6	**15.7**±0.4	**24.6**±0.5

Table 5: Performance comparison at larger synthetic data budgets on Flickr-30K and MS-COCO.

Pairs	Method	Flickr30k IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	Avg	COCO IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	Avg
500	Rand	2.4	10.5	17.4	5.2	18.3	25.7	13.3	1.1	5.0	8.7	1.9	7.5	12.5	6.1
	LoRS [50]	10.0	28.9	41.6	15.5	39.8	53.7	31.6	2.1	8.0	13.7	2.8	9.9	16.2	8.8
	RepBlend [54]	17.0	42.5	**55.9**	22.5	**53.2**	66.7	43.0	6.2	19.9	30.6	7.0	22.0	32.9	19.8
	CovMatch [20]	14.7	38.4	51.4	19.9	46.7	59.5	38.4	5.4	18.0	28.2	8.1	23.5	34.6	19.6
	EDGE [57]	6.7	21.0	30.5	13.3	35.6	47.5	25.8	1.8	6.5	11.2	2.9	9.5	15.7	7.9
	Ours	**17.0**	**43.1**	55.6	**22.7**	52.7	**67.3**	**43.2**	**7.2**	**22.2**	**33.4**	**8.7**	**25.5**	**37.5**	**22.4**
1000	Rand	3.9	13.0	20.7	5.3	17.1	27.7	14.6	1.6	6.5	11.5	2.4	9.9	16.7	8.1
	LoRS [50]	11.2	31.3	42.9	14.4	37.5	50.5	31.3	2.5	9.4	15.4	3.6	11.8	19.1	10.3
	RepBlend [54]	15.3	41.4	53.5	21.1	47.7	60.5	39.9	6.5	19.9	28.9	8.8	24.6	34.5	20.5
	CovMatch [20]	15.1	39.0	51.7	20.1	46.9	58.9	38.5	5.4	18.2	28.7	7.9	24.0	35.1	19.8
	EDGE [57]	9.9	28.2	40.5	14.5	38.3	51.7	30.5	2.8	9.8	16.2	3.9	13.0	21.0	11.1
	Ours	**16.4**	**44.6**	**56.1**	**24.3**	**52.9**	**66.3**	**43.4**	**7.2**	**24.4**	**34.3**	**9.3**	**27.5**	**37.9**	**23.4**

across multiple retrieval metrics, indicating that our approach effectively preserves cross-modal semantic alignment even under highly constrained budgets.

Supervision analysis. We evaluate the quality and effectiveness of our supervision signal from three perspectives. First, we conduct a hallucination check on the fused captions, obtaining a 1.8% sentence-level hallucination rate and a 0.6% object-level hallucination rate, indicating that ITF introduces only negligible noise (see Appendix G for details). Second, we compare direct generation (DG) with our multi-caption supervised strategy (ITF). As shown in Tab. 6, ITF consistently outperforms DG on both image-to-text and text-to-image retrieval, demonstrating that multi-caption, image-grounded fusion provides more faithful and effective supervision for multimodal distillation. Third, to verify that the gain does not simply come from longer textual inputs, we further compare ITF with naive caption concatenation in Appendix I. The results show that ITF improves performance through image-grounded semantic fusion rather than merely combining multiple captions into a longer textual input.

Efficiency and overhead. Time. P^{2}DE incurs only a minor training time increase (1.80%–5.73%) over RepBlend across budgets. **Memory.** RepBlend peaks at 10.17GB during online distillation. Our ITF runs *offline* (9.39GB) without affecting online memory, while enabling ICC raises the online peak to 16.23GB due to additional cross-modal consistency computation. **Parameters.** P^{2}DE uses frozen pre-trained models and introduces no additional trainable parameters.

Table 6: Ablation results on Flickr-30K (100 pairs).

Method	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
Base	11.5	32.0	44.5	16.2	41.7	55.5
DG	10.6	30.6	43.3	14.9	42.1	56.8
Ours (ITF)	**12.6**	**32.4**	**45.5**	**16.4**	**43.1**	**57.3**

Table 7: Distillation time overhead of P^{2}DE on MS-COCO.

pairs	RepBlend (s)	P^{2}DE (s)	Overhead
100	25488	27740	1.80%
200	35580	36492	2.60%
500	39300	41552	5.73%

Overall, P^{2}DE achieves notable performance gains with modest computational and memory overhead.

Statistical significance. We compute the win/tie/loss counts between our method and the compared SOTA method RepBlend [54] using paired t-tests at the 0.05 significance level. The results show that our method outperforms the baselines in all cases, with statistically significant improvements in 77.8% of cases (14 out of 18).

6 Conclusion

We have reformulated the optimization problem of multimodal dataset distillation through the lens of information theory. We propose a novel, theoretically grounded method that aims to maximize the intrinsic interactivity and effective information capacity between image and text modalities. By leveraging conditional information entropy, our method dynamically adjusts the information ratio between paired and unpaired samples, enhancing the information capacity of the synthetic dataset. We conduct extensive experiments on multiple datasets to validate our method, and it significantly outperforms other approaches across multiple cases.

7 Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants U24A20322, 62576094 and 62422118. This work is also supported by Hong Kong UGC under grants UGC/FDS11/E03/24, UGC/FDS11/E03/25, and Hong Kong Research Grants Council under Grant 11219324. This research work is also supported by the Big Data Computing Center of Southeast University.

References

1. Andrew, G., Arora, R., Bilmes, J.A., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning, ICML. pp. 1247–1255 (2013)
2. Brock, A., De, S., Smith, S.L., Simonyan, K.: High-performance large-scale image recognition without normalization. In: International Conference on Machine Learning, ICML. pp. 1059–1071 (2021)

3. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, NeurIPS (2020)
4. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.: Dataset distillation by matching training trajectories. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 10708–10717 (2022)
5. Cui, J., Wang, R., Si, S., Hsieh, C.: Scaling up dataset distillation to imagenet-1k with constant memory. In: International Conference on Machine Learning, ICML. pp. 6565–6590 (2023)
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. pp. 4171–4186 (2019)
7. Ding, G., Chen, R., Yao, A.: Condensing action segmentation datasets via generative network inversion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 17733–17742 (2025)
8. Du, J., Jiang, Y., Tan, V.Y.F., Zhou, J.T., Li, H.: Minimizing the accumulated trajectory error to improve dataset distillation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 3749–3758 (2023)
9. Du, J., Shi, Q., Zhou, J.T.: Sequential subset matching for dataset distillation. In: Advances in Neural Information Processing Systems, NeurIPS (2023)
10. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems, NeurIPS (2013)
11. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: International Conference on Acoustics, Speech and Signal Processing, ICASSP. pp. 776–780 (2017)
12. Gu, J., Vahidian, S., Kungurtsev, V., Wang, H., Jiang, W., You, Y., Chen, Y.: Efficient dataset distillation via minimax diffusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 15793–15803 (2024)
13. Guo, Z., Wang, K., Cazenavette, G., Li, H., Zhang, K., You, Y.: Towards lossless dataset distillation via difficulty-aligned trajectory matching. In: International Conference on Learning Representations, ICLR (2024)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 770–778 (2016)
15. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
16. Ishmael Belghazi, M., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., Devon Hjelm, R.: Mine: mutual information neural estimation. arXiv preprint arXiv:1801.04062 (2018)
17. Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning, ICML. pp. 4904–4916 (2021)

18. Kang, W., Huang, H., Shang, Y., Shah, M., Yan, Y.: Robin3d: Improving 3d large language model via robust instruction tuning. In: IEEE/CVF International Conference on Computer Vision, ICCV. pp. 3905–3915 (2025)
19. Kim, C.D., Kim, B., Lee, H., Kim, G.: Audiocaps: Generating captions for audios in the wild. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT. pp. 119–132 (2019)
20. Lee, Y., Chung, H.W.: Covmatch: Cross-covariance guided multimodal dataset distillation with trainable text encoder. In: Advances in Neural Information Processing Systems, NeurIPS (2025)
21. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, ICML. pp. 12888–12900 (2022)
22. Li, J., Selvaraju, R.R., Gotmare, A., Joty, S.R., Xiong, C., Hoi, S.C.: Align before fuse: Vision and language representation learning with momentum distillation. In: Advances in Neural Information Processing Systems, NeurIPS (2021)
23. Li, N., Liu, A.A., Zhang, J., Cui, J.: Latent video dataset distillation. arXiv preprint arXiv:2504.17132 (2025)
24. Li, Y., Li, W.: Data distillation for text classification. arXiv preprint arXiv:2104.08448 (2021)
25. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: European Conference on Computer Vision, ECCV. pp. 740–755 (2014)
26. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 26286–26296 (2024)
27. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems, NeurIPS (2023)
28. Lu, H., Isonuma, M., Mori, J., Sakata, I.: Unidetox: Universal detoxification of large language models via dataset distillation. In: International Conference on Learning Representations, ICLR (2025)
29. Maekawa, A., Kosugi, S., Funakoshi, K., Okumura, M.: Dilm: Distilling dataset into language model for text-level dataset distillation. In: Findings of the Association for Computational Linguistics: NAACL. pp. 3138–3153 (2024)
30. Nguyen, D., Li, Z., Bateni, M., Mirrokni, V., Razaviyayn, M., Mirzasoleiman, B.: Synthetic text generation for training large language models via gradient matching. arXiv preprint arXiv:2502.17607 (2025)
31. Nguyen, T., Chen, Z., Lee, J.: Dataset meta-learning from kernel ridge-regression. In: International Conference on Learning Representations, ICLR (2021)
32. Nguyen, T., Novak, R., Xiao, L., Lee, J.: Dataset distillation with infinitely wide convolutional networks. In: Advances in Neural Information Processing Systems, NeurIPS (2021)
33. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
34. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. International Journal of Computer Vision, IJCV pp. 74–93 (2017)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable

visual models from natural language supervision. In: International Conference on Machine Learning, ICML. pp. 8748–8763 (2021)
36. Radosavovic, I., Kosaraju, R.P., Girshick, R.B., He, K., Dollár, P.: Designing network design spaces. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 10425–10433 (2020)
37. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 4035–4045 (2018)
38. Schmid, F., Koutini, K., Widmer, G.: Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation. In: International Conference on Acoustics, Speech and Signal Processing ICASSP. pp. 1–5 (2023)
39. Shang, Y., Cai, M., Xu, B., Lee, Y.J., Yan, Y.: Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In: IEEE/CVF International Conference on Computer Vision, ICCV. pp. 22857–22867 (2025)
40. Su, D., Hou, J., Gao, W., Tian, Y., Tang, B.: D^4m: Dataset distillation via disentangled diffusion model. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 5809–5818 (2024)
41. Sun, P., Shi, B., Yu, D., Lin, T.: On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 9390–9399 (2024)
42. Toneva, M., Sordoni, A., des Combes, R.T., Trischler, A., Bengio, Y., Gordon, G.J.: An empirical study of example forgetting during deep neural network learning. In: International Conference on Learning Representations, ICLR (2019)
43. Wang, K., Zhao, B., Peng, X., Zhu, Z., Yang, S., Wang, S., Huang, G., Bilen, H., Wang, X., You, Y.: CAFE: learning to condense dataset by aligning features. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 12186–12195 (2022)
44. Wang, S., Yang, Y., Liu, Z., Sun, C., Hu, X., He, C., Zhang, L.: Dataset distillation with neural characteristic function: A minmax perspective. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 25570–25580 (2025)
45. Wang, T., Zhu, J.Y., Torralba, A., Efros, A.A.: Dataset distillation. arXiv preprint arXiv:1811.10959 (2018)
46. Wang, Z., Xu, Y., Lu, C., Li, Y.: Dancing with still images: Video distillation via static-dynamic disentanglement. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 6296–6304 (2024)
47. Welling, M.: Herding dynamical weights to learn. In: International Conference on Machine Learning, ICML. pp. 1121–1128 (2009)
48. Wolf, G.W.: Facility location: concepts, models, algorithms and case studies. series: Contributions to management science. International Journal Of Geographical Information Science, IJGIS **25**, 331–333 (2011)
49. Wu, X., Zhang, B., Deng, Z., Russakovsky, O.: Vision-language dataset distillation. Transactions on Machine Learning Research, TMLR (2024)
50. Xu, Y., Lin, Z., Qiu, Y., Lu, C., Li, Y.: Low-rank similarity mining for multimodal dataset distillation. In: International Conference on Machine Learning, ICML (2024)
51. Yin, Z., Xing, E.P., Shen, Z.: Squeeze, recover and relabel: Dataset condensation at imagenet scale from A new perspective. In: Advances in Neural Information Processing Systems, NeurIPS (2023)

52. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. Transactions on Machine Learning Research, TMLR (2022)
53. Zhang, H., Li, S., Lin, F., Wang, W., Qian, Z., Ge, S.: DANCE: dual-view distribution alignment for dataset condensation. In: International Joint Conference on Artificial Intelligence, IJCAI. pp. 1679–1687 (2024)
54. Zhang, X., Zhang, Z., Du, J., Liu, Z., Zhou, J.T.: Beyond modality collapse: Representations blending for multimodal dataset distillation. In: Advances in Neural Information Processing Systems, NeurIPS (2025)
55. Zhao, B., Bilen, H.: Dataset condensation with distribution matching. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV. pp. 6503–6512 (2023)
56. Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. In: International Conference on Learning Representations, ICLR (2021)
57. Zhao, Z., Wang, H., Wu, J., Shang, Y., Liu, G., Yan, Y.: Efficient multimodal dataset distillation via generative models. In: Advances in Neural Information Processing Systems, NeurIPS (2025)
58. Zhou, Y., Nezhadarya, E., Ba, J.: Dataset distillation using neural feature regression. In: Advances in Neural Information Processing Systems, NeurIPS (2022)