Cross-Modal Correspondence Matters: Unified Prototype-Guided Data Synthesis for Multimodal Dataset Distillation

Anonymous authors
Paper under double-blind review

Abstract

Multimodal dataset distillation aims to construct compact yet informative image–text datasets that preserve essential knowledge from large-scale datasets for efficient cross-modal learning. Recent prototype-guided approaches typically construct visual and textual prototypes separately, primarily preserving modality-specific structures while insufficiently capturing the joint dependency that defines image–text correspondence. In this paper, we analyze this limitation from an information-theoretic perspective and introduce **UPSyn**, a **U**nified **P**rototype-guided **Syn**thesis framework that designed to adequately preserve cross-modal information from the original dataset. Specifically, we construct two cross-modal common spaces from correlation and geometry perspectives, and jointly cluster the paired data by integrating the structural evidence from both spaces, thereby producing informative prototypes that better maintain image–text dependency. Additionally, we leverage a large language model to aggregate and summarize the entities surrounding each textual prototype into a refined prototype, providing semantically rich guidance for subsequent distilled data generation. Extensive experiments on Flickr30K and MS-COCO demonstrate that our method consistently outperforms coreset-based, pixel-optimization-based and generative-based baselines in bidirectional retrieval, while retaining strong cross-architecture generalization with an efficient training-free paradigm.

1 Introduction

Dataset distillation (DD) (Wang et al., 2018; Nguyen et al., 2021) aims to compress a large-scale training set into a tiny set of informative synthetic samples while preserving its utility for efficient learning of downstream tasks. Existing studies have predominantly focused on unimodal visual data, where both the original and distilled datasets consist solely of images (Zhao et al., 2021; Lee et al., 2024). With the rapid development of Vision–Language Models (VLM) (Radford et al., 2021), this paradigm has recently been extended to multimodal settings, giving rise to Multimodal Dataset Distillation (MDD) that synthesizes compact image–text pairs from large-scale vision–language datasets.

Previous MDD methods primarily rely on iterative pixel-space optimization, including trajectory-, similarity-, and distribution-matching approaches (Wu et al., 2024; Xu et al., 2024; Lee & Chung, 2025). Despite their effectiveness, these methods often incur substantial computational overhead and exhibit strong dependence on the training architecture. More importantly, their matching objectives capture cross-modal relationships only implicitly, without explicitly modeling the joint dependency between paired images and texts, potentially leading to the loss of cross-modal information during distillation.

Recent Prototype-Guided Data Synthesis (PDS) (Choi et al., 2026) provides an effective and efficient paradigm by explicitly discovering representative prototypes in the CLIP space and decoding them into realistic images with a pre-trained unCLIP (Ramesh et al., 2022) decoder. However, PDS follows a sequential "cluster-then-align" strategy, where modality-specific prototypes are independently discovered before establishing their cross-modal correspondence. This decoupled process fails to fully capture the joint dependency underlying image–text correspondence. As shown in Fig 1(left), this manifests as intra-modality aggregation and inter-modality separation (Liang et al., 2022; Zhang et al., 2025), giving rise to two distinct deficiencies. First, intra-modal prototype contraction causes representations within each modality to become overly concentrated, reducing prototype diversity and

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.

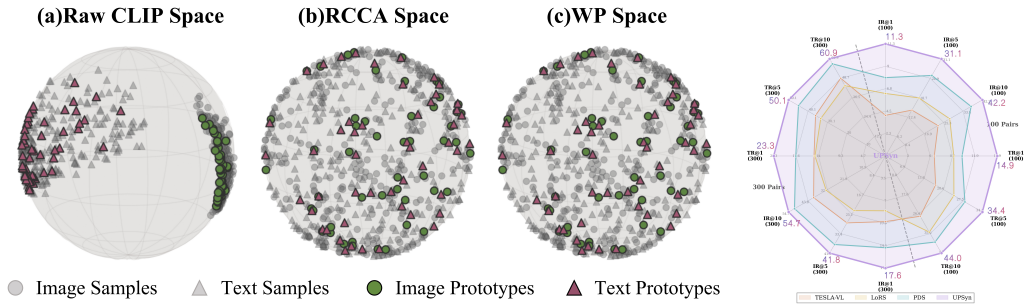

Figure 1: Motivation and overall performance of UPSyn. Left: a subset of Flickr30K image–text samples and prototypes visualized in the raw CLIP, RCCA, and WP spaces. The Raw CLIP Space shows PDS prototypes, whereas the RCCA and WP spaces show UPSyn prototypes. Right: bidirectional retrieval performance under 100- and 300-pair settings.

obscuring fine-grained semantic variations. Second, cross-modal prototype divergence drives image and text prototypes apart, disrupting their shared semantic structure and weakening the correspondence between paired image–text representations. Together, these effects prevent the discovered prototypes from faithfully preserving the rich multimodal semantics of the original paired image–text data.

In this paper, we theoretically characterize modality-specific prototype discovery as relevance-agnostic compression from an information-theoretic perspective, showing via the data processing inequality (Cover & Thomas, 2006) that independent clustering cannot guarantee the preservation of cross-modal information. To address this, we propose **UPSyn**, a unified prototype-guided synthesis framework that explicitly incorporates paired cross-modal structure into prototype discovery. Specifically, UPSyn constructs two common spaces from correlation-oriented and geometry-oriented perspectives. The former employs Regularized Canonical Correlation Analysis (RCCA) (Vinod, 1976; Leurgans et al., 1993) to capture statistical dependency between paired image and text representations, while the latter adopts Whitening–Procrustes (WP) (Kessy et al., 2018; Schönemann, 1966) to reduce modality-specific covariance differences and preserve clustering-relevant geometry. Rather than independently discovering prototypes in each modality, UPSyn jointly integrates the assignment evidence from both spaces through consensus refinement, yielding a shared partition for paired image–text data. The resulting assignments are then mapped back to the original CLIP space to reconstruct image and text prototypes for subsequent synthesis of distilled image–text pairs.

To further enrich the textual semantics associated with each prototype, we introduce **Region-aware Caption Refinement (RCR)**. Given a textual prototype, RCR retrieves its Top-k nearest captions from the original training set and employs a Large Language Model (LLM) to consolidate their shared semantic content into a refined prototype-level description. This process supplements the compact text prototype with richer local semantic information while remaining anchored to its neighborhood in the CLIP text space. The refined caption is subsequently used together with the corresponding image prototype to guide image synthesis with a pre-trained unCLIP decoder. Overall, our contributions are summarized as follows:

- We theoretically and empirically reveal that preserving modality-specific structures alone is insufficient to fully retain the cross-modal dependency underlying image–text correspondence, resulting in prototypes with limited cross-modal semantic representativeness.
- We introduce UPSyn, which integrates correlation-oriented RCCA with geometry-oriented WP to construct a shared multimodal prototype space. This unified space better preserves the cross-modal dependencies of the original paired data during prototype discovery.
- We propose RCR, which leverages an LLM to consolidate entities surrounding each textual prototype into a refined representation. This refinement enriches prototype semantics and provides informative guidance for subsequent synthesis of distilled image–text pairs.
- Extensive experiments on Flickr30K and MS-COCO demonstrate consistent improvements over various distillation baselines with favorable cross-architecture generalization.

2 Preliminaries

2.1 Problem Formulation

Given a paired image–text dataset $\mathcal{D} = \{(I_n, T_n)\}_{n=1}^{N}$, MDD seeks to construct a compact dataset $\mathcal{S} = \{(\widetilde{I}_m, \widetilde{T}_m)\}_{m=1}^{M}$, with $M \ll N$, such that a vision–language model trained on $\mathcal{S}$ achieves downstream performance comparable to that obtained by training on the full dataset $\mathcal{D}$. Let $\mathcal{B}$ be a mini-batch drawn from the underlying paired distribution and $\mathcal{L}$ a vision–language training objective. We formulate this objective as the following bilevel optimization problem:

$$\begin{aligned} \mathcal{S}^* &= \arg\min_{\mathcal{S},\,|\mathcal{S}|=M} \mathbb{E}_{\mathcal{B}}\left[\left|\mathcal{L}(\mathcal{B};\theta^*_{\mathcal{S}}) - \mathcal{L}(\mathcal{B};\theta^*_{\mathcal{D}})\right|\right], \\ \text{s.t.} \qquad \theta^*_{\mathcal{X}} &= \arg\min_{\theta} \mathcal{L}(\mathcal{X};\theta), \qquad \mathcal{X} \in \{\mathcal{D}, \mathcal{S}\}. \end{aligned} \tag{1}$$

Optimization-based methods approximate this objective by optimizing synthetic samples or matching training statistics. Prototype-guided approaches instead summarize frozen vision–language features with representative image–text prototypes that guide synthesis. The quality of these prototypes depends on both within-modality representativeness and preservation of the dependency between paired modalities in the original image–text training data.

2.2 Rethinking Modality-Specific Prototype Discovery

Let X and T denote paired image and text features drawn from P_{XT}. Independent clustering produces deterministic assignments $Z_X = f(X)$ and $Z_T = g(T)$. Since each assignment depends only on its own modality, $Z_X \to X \to T \to Z_T$ forms a Markov chain. The data processing inequality (Cover & Thomas, 2006) gives

$$I(Z_X; Z_T) \le I(Z_X; T) \le I(X; T), \tag{2}$$

where $I(\cdot;\cdot)$ denotes mutual information. Equality requires both compression functions to retain all information in each modality that is relevant to the other. Prototype-guided approaches commonly use K-Means (Lloyd, 1982) to obtain representative cluster centers (Choi et al., 2026). Independent K-Means does not enforce cross-modal information preservation, because it minimizes separate within-modality quantization errors:

$$\mathcal{L}_{\text{sep}} = \min_{f,\{\boldsymbol{\mu}^I_k\}} \mathbb{E}\left[\|X - \boldsymbol{\mu}^I_{f(X)}\|_2^2\right] + \min_{g,\{\boldsymbol{\mu}^T_k\}} \mathbb{E}\left[\|T - \boldsymbol{\mu}^T_{g(T)}\|_2^2\right]. \tag{3}$$

These objectives depend on P_X and P_T, which do not generally determine P_{XT}. Independent prototype discovery is thus *relevance-agnostic compression*: it has no explicit criterion for retaining variation that is informative about the paired modality (Tishby et al., 2000). Matching based only on the compressed representations cannot recover discarded information. This motivates cross-modal modeling before partitioning the paired data into representative multimodal prototype clusters.

A natural way to quantify the cross-modal dependency that should be preserved during prototype discovery is mutual information $I(X;T)$. Direct estimation of $I(X;T)$ in high-dimensional CLIP spaces is difficult. Under a jointly Gaussian model with nonsingular joint covariance, the canonical correlations provide an exact characterization (Hotelling, 1936; Bach & Jordan, 2002). Define

$$\mathbf{M} = \boldsymbol{\Sigma}_{XX}^{-1/2} \boldsymbol{\Sigma}_{XT} \boldsymbol{\Sigma}_{TT}^{-1/2}, \tag{4}$$

where $\boldsymbol{\Sigma}_{XX}$ and $\boldsymbol{\Sigma}_{TT}$ are within-modality covariances and $\boldsymbol{\Sigma}_{XT}$ is the cross-covariance. The singular values $\{\rho_j\}_{j=1}^{r}$ of $\mathbf{M}$ are the canonical correlations, and

$$I(X;T) = -\frac{1}{2}\sum_{j=1}^{r} \log(1 - \rho_j^2). \tag{5}$$

Each term increases with $\rho_j \in [0, 1)$, linking strongly correlated directions to cross-modal dependency. We use this Gaussian relation to motivate RCCA as a correlation-oriented common space. Since clustering also depends on the geometric organization of the representation space, we further introduce WP as a geometry-oriented space, where whitening reduces modality-specific covariance and scaling differences and orthogonal Procrustes aligns the two modalities while preserving the geometry of the whitened space.

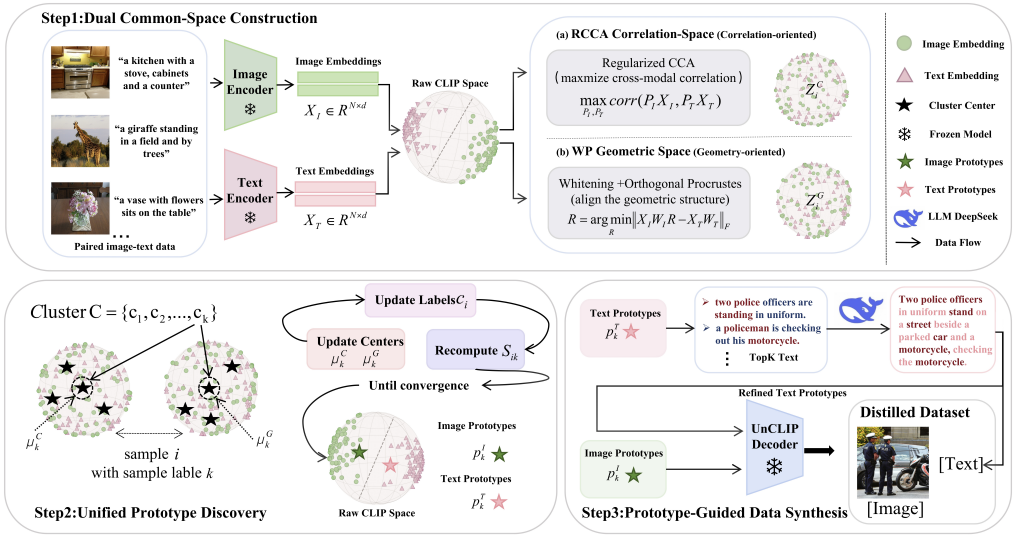

Figure 2: Overview of UPSyn. UPSyn first constructs correlation-oriented RCCA and geometry-oriented WP common spaces from frozen CLIP representations, and then refines a shared image–text partition through cross-space consensus to obtain multimodal prototypes. RCR retrieves the Top-k captions around each text prototype and aggregates them with an LLM, after which the refined text descriptions and image prototypes jointly guide a pre-trained unCLIP decoder to synthesize the final distilled dataset.

3 Method

Overview. UPSyn constructs RCCA and WP common spaces and refines a shared prototype partition through cross-space consensus. The resulting prototypes are mapped back to the original CLIP space, while RCR provides refined textual conditioning for synthesis with a frozen unCLIP decoder. Figure 2 summarizes the framework from common-space construction to prototype-guided data synthesis.

3.1 Unified Cross-Modal Prototype Discovery

We extract paired features $\mathbf{x}_i = f_I(I_i)$ and $\mathbf{t}_i = f_T(T_i)$ with frozen CLIP encoders. We then center them as $\bar{\mathbf{x}}_i = \mathbf{x}_i - \boldsymbol{\mu}_I$ and $\bar{\mathbf{t}}_i = \mathbf{t}_i - \boldsymbol{\mu}_T$, and compute the within-modality covariance matrices $\boldsymbol{\Sigma}_{II} = \frac{1}{N}\sum_{i=1}^{N} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^\top$ and $\boldsymbol{\Sigma}_{TT} = \frac{1}{N}\sum_{i=1}^{N} \bar{\mathbf{t}}_i \bar{\mathbf{t}}_i^\top$, together with the cross-covariance $\boldsymbol{\Sigma}_{IT} = \frac{1}{N}\sum_{i=1}^{N} \bar{\mathbf{x}}_i \bar{\mathbf{t}}_i^\top$.

Correlation-oriented RCCA space. RCCA estimates correlated cross-modal directions using regularized within-modality covariances (Vinod, 1976; Leurgans et al., 1993). Writing $\widetilde{\boldsymbol{\Sigma}}_{II} = \boldsymbol{\Sigma}_{II} + \lambda_I \mathbf{I}$ and $\widetilde{\boldsymbol{\Sigma}}_{TT} = \boldsymbol{\Sigma}_{TT} + \lambda_T \mathbf{I}$, we obtain

$$\begin{aligned} \mathbf{M}_C &= \widetilde{\boldsymbol{\Sigma}}_{II}^{-1/2} \boldsymbol{\Sigma}_{IT} \widetilde{\boldsymbol{\Sigma}}_{TT}^{-1/2} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\top, \quad \boldsymbol{\Lambda} = \mathrm{diag}(\rho_1, \ldots, \rho_r), \\ \mathbf{z}_{i,I}^C &= \boldsymbol{\Lambda}_{d_C}^c \mathbf{U}_{d_C}^\top \widetilde{\boldsymbol{\Sigma}}_{II}^{-1/2} \bar{\mathbf{x}}_i, \quad \mathbf{z}_{i,T}^C = \boldsymbol{\Lambda}_{d_C}^c \mathbf{V}_{d_C}^\top \widetilde{\boldsymbol{\Sigma}}_{TT}^{-1/2} \bar{\mathbf{t}}_i. \end{aligned} \tag{6}$$

We retain the leading d_C canonical components, where d_C denotes the output dimensionality of the RCCA common space. The exponent c controls the weighting of these components according to their canonical correlations, with $c = 1$ in our default configuration. Here, ρ_j denotes the j-th regularized canonical correlation obtained from the corresponding singular value of $\mathbf{M}_C$.

Geometry-oriented Whitening–Procrustes space. Let $\mathbf{W}_I = \widetilde{\boldsymbol{\Sigma}}_{II}^{-1/2}$ and $\mathbf{W}_T = \widetilde{\boldsymbol{\Sigma}}_{TT}^{-1/2}$ denote the whitening matrices for the image and text modalities, respectively. They transform the centered features into $\widehat{\mathbf{X}} = \mathbf{W}_I \bar{\mathbf{X}}$ and $\widehat{\mathbf{T}} = \mathbf{W}_T \bar{\mathbf{T}}$, where samples are arranged as columns. Whitening reduces modality-specific covariance and scaling differences before cross-modal alignment (Kessy et al., 2018). We then seek an orthogonal alignment matrix $\mathbf{R}$ that maps the whitened text representation to

the whitened image space while preserving the geometry of the text representation. Specifically,

$$\begin{aligned} \mathbf{R}^* &= \arg\min_{\mathbf{R}^\top\mathbf{R}=\mathbf{I}} \|\widehat{\mathbf{X}} - \mathbf{R}\widehat{\mathbf{T}}\|_F^2 = \mathbf{U}_G\mathbf{V}_G^\top, \quad \widehat{\mathbf{X}}\widehat{\mathbf{T}}^\top = \mathbf{U}_G\mathbf{\Lambda}_G\mathbf{V}_G^\top, \\ \mathbf{z}_{i,I}^G &= \mathbf{W}_I\bar{\mathbf{x}}_i, \qquad \mathbf{z}_{i,T}^G = \mathbf{R}^*\mathbf{W}_T\bar{\mathbf{t}}_i. \end{aligned} \tag{7}$$

Here, $\mathbf{R}^*$ denotes the optimal orthogonal transformation obtained from the SVD of $\widehat{\mathbf{X}}\widehat{\mathbf{T}}^\top$. It aligns the whitened text features with the image features without introducing additional geometric distortion. Whitening reduces modality-specific covariance and scale differences. The subsequent orthogonal transformation preserves pairwise distances and angles within the whitened text space during alignment.

Pair-level representation. Within each common space, we combine the two modalities of each pair:

$$\mathbf{z}_i^s = \mathrm{Norm}\Big(\alpha\mathbf{z}_{i,I}^s + (1-\alpha)\mathbf{z}_{i,T}^s\Big), \qquad s \in \{C, G\}, \tag{8}$$

where $\alpha \in [0, 1]$ controls the relative contributions of the image and text representations when constructing the pair-level feature; we use $\alpha = 0.5$ by default to balance the two modalities.

Consensus prototype discovery. We initialize shared labels by K-Means on $\mathbf{z}_i^{\text{init}} = [\mathbf{z}_i^C/\sqrt{2}; \mathbf{z}_i^G/\sqrt{2}]$. Each space then maintains its own centroid $\boldsymbol{\mu}_k^s = \mathrm{Norm}(|C_k|^{-1}\sum_{i\in C_k}\mathbf{z}_i^s)$, while both use the same partition $\{C_k\}_{k=1}^K$. From the cosine affinities $s_{ik}^s = \mathrm{sim}(\mathbf{z}_i^s, \boldsymbol{\mu}_k^s)$, we compute

$$S_{ik} = \tfrac{1}{2}(s_{ik}^C + s_{ik}^G) - \gamma|s_{ik}^C - s_{ik}^G|. \tag{9}$$

The first term rewards average support and the second penalizes disagreement, with strength γ. We alternate centroid updates and shared reassignment, $c_i = \arg\max_k S_{ik}$, until convergence. The common spaces determine cluster membership only: the final prototypes are reconstructed as $\mathbf{p}_k^I = \mathrm{Norm}(|C_k|^{-1}\sum_{i\in C_k}\mathbf{x}_i)$ and $\mathbf{p}_k^T = \mathrm{Norm}(|C_k|^{-1}\sum_{i\in C_k}\mathbf{t}_i)$ in the original CLIP space, preserving compatibility with the pre-trained generator for subsequent prototype-guided image synthesis.

3.2 Region-Aware Caption Refinement

The text prototype $\mathbf{p}_k^T$ represents the semantic center of a cluster in the CLIP text space, but it is not itself a natural-language description suitable for rich textual conditioning. Using only its nearest caption may capture only a limited aspect of the underlying semantics. We therefore retrieve the K_t nearest training captions to $\mathbf{p}_k^T$ according to cosine similarity,

$$\mathcal{N}_k^T = \{T_{k,1}, \ldots, T_{k,K_t}\}, \tag{10}$$

and employ an LLM to consolidate their shared and complementary semantic information into a refined prototype-level description $\widetilde{T}_k$. A shortened excerpt of the refinement prompt is shown below:

```
You will obtain the Top-{K} most similar texts and
synthesize them into a single, semantically rich prototype
description.
## Input Texts:
{TEXTS}
## Guidelines:
1.  Identify recurring semantic concepts across the
retrieved texts, including entities, attributes, actions,
and scene context.
2.  Prioritize semantic information that is consistently
supported by multiple texts.  The complete prompt, including
the output constraints and detailed semantic preservation
rules, is provided in Appendix B.5.  By aggregating
information from the local caption neighborhood, T̃_k provides
richer natural-language conditioning while remaining
anchored to the semantics of the discovered prototype.
```

Table 1: Cross-architecture retrieval on Flickr30K (top) and MS-COCO (bottom) using 100 and 300 distilled pairs. Results are mean ± standard deviation. Bold and underlining indicate the best and second-best results within each setting.

Evaluation Model	Pairs	Method	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
Flickr30K								
ResNet	100	TESLA-VL	4.1 ± 0.3	14.7 ± 0.9	22.9 ± 1.2	6.5 ± 0.4	17.8 ± 1.4	27.3 ± 1.4
		LoRS	6.3 ± 0.1	18.6 ± 0.1	28.0 ± 0.2	9.1 ± 0.2	24.3 ± 0.4	34.5 ± 0.8
		PDS	7.9 ± 0.3	25.8 ± 0.4	37.3 ± 0.3	10.2 ± 0.3	28.2 ± 0.9	39.0 ± 0.3
		UPSyn	**11.3 ± 0.1**	**31.1 ± 0.9**	**42.2 ± 0.2**	**14.9 ± 0.2**	**34.4 ± 0.3**	**44.0 ± 0.8**
	300	TESLA-VL	10.3 ± 0.3	28.8 ± 0.6	40.6 ± 0.8	14.9 ± 0.9	36.2 ± 1.1	48.8 ± 1.7
		LoRS	8.6 ± 0.9	23.5 ± 1.5	33.5 ± 1.8	14.7 ± 1.2	33.5 ± 2.8	44.1 ± 3.2
		PDS	14.4 ± 0.4	38.1 ± 0.2	51.4 ± 0.4	18.7 ± 0.5	45.0 ± 0.4	57.8 ± 0.6
		UPSyn	**17.6 ± 0.7**	**41.8 ± 1.0**	**54.7 ± 0.8**	**23.3 ± 0.8**	**50.1 ± 0.6**	**60.9 ± 0.5**
ViT	100	TESLA-VL	2.1 ± 0.3	7.8 ± 0.7	13.1 ± 1.2	2.6 ± 0.6	8.7 ± 0.9	13.7 ± 1.4
		LoRS	2.8 ± 0.1	9.9 ± 0.4	16.1 ± 0.2	5.2 ± 0.3	13.6 ± 0.5	20.5 ± 0.2
		PDS	6.8 ± 0.3	19.2 ± 0.3	28.5 ± 0.4	6.6 ± 0.5	17.5 ± 0.5	26.9 ± 0.5
		UPSyn	**7.3 ± 0.3**	**22.4 ± 0.5**	**32.1 ± 0.3**	**7.4 ± 0.5**	**21.7 ± 0.4**	**31.4 ± 0.8**
	300	TESLA-VL	5.1 ± 0.7	16.2 ± 1.0	24.5 ± 1.1	6.1 ± 0.6	18.0 ± 1.0	27.3 ± 1.4
		LoRS	4.1 ± 0.5	13.1 ± 1.0	20.7 ± 1.3	6.2 ± 0.6	17.1 ± 1.2	25.7 ± 1.8
		PDS	9.1 ± 0.1	27.3 ± 0.4	38.4 ± 0.4	9.6 ± 0.3	26.1 ± 0.5	37.5 ± 1.2
		UPSyn	**11.5 ± 0.3**	**29.2 ± 0.8**	**40.5 ± 0.5**	**11.7 ± 0.3**	**28.6 ± 0.6**	**40.4 ± 0.9**
MS-COCO								
ResNet	100	TESLA-VL	1.4 ± 0.1	5.8 ± 0.2	10.2 ± 0.4	2.1 ± 0.3	7.6 ± 0.3	12.5 ± 0.3
		LoRS	1.8 ± 0.1	6.8 ± 0.2	11.4 ± 0.4	2.3 ± 0.2	7.8 ± 0.5	12.6 ± 0.4
		PDS	2.8 ± 0.1	10.0 ± 0.2	17.3 ± 0.3	4.5 ± 0.2	14.0 ± 0.3	21.4 ± 0.4
		UPSyn	**4.2 ± 0.1**	**12.3 ± 0.2**	**19.8 ± 0.3**	**5.9 ± 0.2**	**15.5 ± 0.2**	**23.1 ± 0.4**
	300	TESLA-VL	3.0 ± 0.1	10.7 ± 0.3	17.6 ± 0.4	5.7 ± 0.2	17.0 ± 0.3	25.6 ± 0.4
		LoRS	2.5 ± 0.2	8.5 ± 0.4	13.8 ± 0.6	3.2 ± 0.2	10.1 ± 0.4	15.9 ± 0.6
		PDS	5.3 ± 0.2	17.2 ± 0.4	27.2 ± 0.6	7.4 ± 0.3	20.7 ± 0.3	30.2 ± 0.4
		UPSyn	**6.5 ± 0.7**	**18.8 ± 0.0**	**28.5 ± 0.3**	**8.6 ± 0.0**	**21.6 ± 0.4**	**31.7 ± 0.9**
ViT	100	TESLA-VL	0.5 ± 0.2	2.1 ± 0.8	3.8 ± 1.4	0.2 ± 0.1	1.0 ± 0.5	1.7 ± 0.9
		LoRS	0.8 ± 0.1	3.0 ± 0.2	5.4 ± 0.4	0.7 ± 0.2	2.4 ± 0.3	4.1 ± 0.5
		PDS	2.3 ± 0.1	8.6 ± 0.1	14.5 ± 0.1	2.2 ± 0.1	7.7 ± 0.2	13.2 ± 0.3
		UPSyn	**3.5 ± 0.1**	**10.2 ± 0.2**	**16.1 ± 0.4**	**3.4 ± 0.1**	**9.5 ± 0.3**	**15.1 ± 0.3**
	300	TESLA-VL	1.5 ± 0.1	5.9 ± 0.1	10.1 ± 0.3	2.2 ± 0.2	7.8 ± 0.2	12.6 ± 0.4
		LoRS	1.0 ± 0.3	3.7 ± 1.1	6.3 ± 1.8	1.3 ± 0.4	4.6 ± 1.4	7.4 ± 2.0
		PDS	4.1 ± 0.1	13.4 ± 0.1	21.2 ± 0.1	3.7 ± 0.1	12.6 ± 0.1	19.8 ± 0.2
		UPSyn	**5.2 ± 0.7**	**14.8 ± 0.1**	**22.5 ± 0.9**	**5.1 ± 0.3**	**14.3 ± 0.3**	**21.8 ± 0.3**

3.3 Prototype-Guided Data Synthesis

Following PDS (Choi et al., 2026), we use a pre-trained unCLIP (Ramesh et al., 2022) decoder to synthesize distilled images from the discovered image prototypes and aggregated text descriptions. Because unCLIP operates in the CLIP-aligned representation space, it can directly consume the prototype information produced by UPSyn as conditioning for the generation of distilled images.

4 Experiments

4.1 Experimental Settings

Datasets and metrics. We evaluate on Flickr30K (Plummer et al., 2015) and MS-COCO (Lin et al., 2014), using 29,000/1,000 and approximately 113,000/5,000 training/test images, respectively. Each image has five human-annotated captions. Following PDS (Choi et al., 2026), we report Recall at $k \in \{1, 5, 10\}$ for text-to-image retrieval (IR) and image-to-text retrieval (TR). Candidates are ranked by cosine similarity; a query is successful when a ground-truth match appears in the top-k results.

Implementation details. Unless otherwise specified, we follow the experimental protocol of PDS (Choi et al., 2026). Frozen CLIP ViT-L/14 (Radford et al., 2021) encodes both modalities as 768-dimensional features. We construct 256-dimensional RCCA and WP spaces with covariance shrinkage 0.1 and RCCA ridge regularization 10^{-3}. The sensitivity to the common-space dimensionality is

Table 2: Subset selection and generative distillation on Flickr30K with 100 pairs. Subset selection retains real pairs; generative methods synthesize compact training data. Results are mean ± standard deviation.

Metric	Subset Selection					Generative Distillation				
	K-center	Herding	CLIP Score	LAION Filtering	Image-based	D^4M	MGD^3	EDGE	PDS	**UPSyn**
IR@1	2.9 ± 0.1	3.6 ± 0.2	2.5 ± 0.1	2.4 ± 0.1	2.2 ± 0.2	1.3 ± 0.1	2.6 ± 0.1	2.9 ± 0.2	7.9 ± 0.3	**11.3 ± 0.1**
IR@5	10.4 ± 0.3	12.6 ± 0.5	8.7 ± 0.3	8.5 ± 0.2	8.6 ± 0.3	5.8 ± 0.2	10.0 ± 0.2	10.6 ± 0.1	25.8 ± 0.4	**31.1 ± 0.9**
IR@10	16.8 ± 0.3	20.1 ± 0.6	14.5 ± 0.3	14.5 ± 0.3	13.6 ± 0.4	9.8 ± 0.3	17.2 ± 0.5	16.9 ± 0.3	37.3 ± 0.3	**42.2 ± 0.2**
TR@1	5.3 ± 0.3	6.7 ± 0.3	4.7 ± 0.4	4.7 ± 0.3	4.0 ± 0.3	1.4 ± 0.2	3.4 ± 0.2	3.7 ± 0.4	10.2 ± 0.3	**14.9 ± 0.2**
TR@5	15.6 ± 0.5	19.4 ± 0.7	12.6 ± 0.7	12.4 ± 0.6	10.5 ± 0.4	5.6 ± 0.5	11.6 ± 0.3	11.2 ± 0.5	28.2 ± 0.9	**34.4 ± 0.3**
TR@10	24.1 ± 0.4	28.2 ± 0.6	18.9 ± 0.7	19.3 ± 0.4	16.2 ± 0.6	9.5 ± 0.3	17.5 ± 0.4	16.7 ± 0.5	39.0 ± 0.3	**44.0 ± 0.8**

further investigated in Appendix E.3. We use $\alpha = 0.5$, $\gamma = 0.1$, at most 30 consensus iterations, and an assignment-change tolerance of 10^{-3}. The sensitivity to α and γ is analyzed in Appendix E.1, while the convergence behavior of the consensus refinement is further examined in Appendix E.4. The number of clusters equals the number of distilled pairs. For RCR, we use DeepSeek-V4-Flash (DeepSeek-AI, 2026) as the LLM-based caption aggregator to consolidate the retrieved captions into a refined prototype-level description. The pre-trained unCLIP decoder uses classifier-free guidance (Ho & Salimans, 2022) with scale 5.0, and a fixed denoising budget of 20 steps is applied to each prototype. Generated images are then resized to 224×224 for evaluation. Following Xu et al. (2024); Choi et al. (2026), we evaluate ResNet-50 (He et al., 2016) and ViT-Ti/16 (Dosovitskiy et al., 2021) backbones. The text encoder remains frozen while the image encoder and a linear projection are trained on the distilled pairs. Experiments use a single NVIDIA T4 GPU. Additional settings and baseline descriptions are provided in Appendices C and D for reference and reproducibility.

4.2 Main Results

Cross-architecture generalization. Table 1 shows that UPSyn consistently improves over PDS across both datasets, distillation budgets, and evaluation backbones. The gains are particularly pronounced on Flickr30K. Under the 100-pair ResNet setting, UPSyn improves all six retrieval metrics by an average of 4.9 points over PDS, with gains of 3.4, 5.3, and 4.9 points on IR@1/5/10 and 4.7, 6.2, and 5.0 points on TR@1/5/10, respectively. The largest improvement appears on TR@5, increasing from 28.2 to 34.4. With 300 pairs, the improvement remains consistent, ranging from 3.1 to 5.1 points across the six metrics. Similar cross-architecture gains are observed with ViT: under the 100-pair setting, IR@10 and TR@10 improve by 3.6 and 4.5 points, respectively, while the 300-pair setting yields gains of 1.9–2.9 points across all metrics under this distilled-pair budget.

On the larger MS-COCO dataset, UPSyn also consistently outperforms PDS in every evaluated setting, although the absolute gains are more moderate. With 100 pairs, ResNet improvements range from 1.4 to 2.5 points, including a 2.5-point gain on IR@10, while ViT gains range from 1.2 to 1.9 points. The same trend persists with 300 pairs, where UPSyn improves all six metrics under both ResNet and ViT evaluation. These results indicate that the distilled image–text pairs produced by UPSyn remain effective across different dataset scales, compression budgets, and unseen vision architectures for both image and text retrieval.

Subset selection and generative distillation. Table 2 compares 100-pair distilled datasets on Flickr30K. Subset selection methods, including K-center (Zanjirani Farahani & Hekmatfar, 2009), Herding (Welling, 2009), and DataComp-style filtering (Gadre et al., 2023), directly retain samples from the original dataset. For D^4M (Su et al., 2024) and MGD^3 (Chan Santiago et al., 2025), we retain the VAE image encoder (Kingma & Welling, 2014; Rombach et al., 2022), introduce CLIP text features, and match the resulting image and text prototypes before applying their original synthesis procedures. Their relatively weak retrieval performance is consistent with the lack of explicit alignment between the VAE image space and the CLIP text space. EDGE (Zhao et al., 2025) and PDS are generative methods developed specifically for multimodal dataset distillation. EDGE achieves 16.9 IR@10 and 16.7 TR@10, while PDS improves these results to 37.3 and 39.0, respectively. UPSyn further increases them to 42.2 and 44.0, corresponding to gains of 4.9 and 5.0 points over PDS. UPSyn also outperforms all subset selection baselines across the six retrieval metrics in this setting.

Table 3: Ablation study on Flickr30K with 100 distilled pairs. RCCA and WP denote the two common spaces, *Cons.* denotes consensus refinement, and RCR denotes our Region-Aware Caption Refinement. Joint K-Means concatenates the paired CLIP image and text features and performs joint clustering in the concatenated space. Bold and underlining indicate the best and second-best results per backbone.

Eval.	Variant	RCCA	WP	Cons.	RCR	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	PDS					7.9 ± 0.3	25.8 ± 0.4	37.3 ± 0.3	10.2 ± 0.3	28.2 ± 0.9	39.0 ± 0.3
	Joint K-Means					8.8 ± 0.4	26.2 ± 0.4	38.0 ± 0.3	11.5 ± 0.1	29.2 ± 0.2	41.4 ± 0.6
	Only RCCA	✓				9.4 ± 0.4	28.5 ± 0.8	40.7 ± 0.9	12.7 ± 0.3	31.0 ± 0.5	42.2 ± 0.7
	Only WP		✓			9.8 ± 0.7	28.9 ± 0.3	41.0 ± 0.1	11.6 ± 0.2	29.3 ± 0.1	40.4 ± 0.4
	Dual Space	✓	✓			10.3 ± 0.7	28.9 ± 0.1	41.4 ± 0.8	13.2 ± 0.6	32.1 ± 0.1	43.2 ± 0.4
	UPSyn w/o RCR	✓	✓	✓		10.5 ± 0.2	29.9 ± 0.6	41.8 ± 0.2	14.1 ± 0.6	32.3 ± 0.3	43.5 ± 0.1
	UPSyn	✓	✓	✓	✓	**11.3 ± 0.1**	**31.1 ± 0.9**	**42.2 ± 0.2**	**14.9 ± 0.2**	**34.4 ± 0.3**	**44.0 ± 0.8**
ViT	PDS					6.8 ± 0.3	19.2 ± 0.3	28.5 ± 0.4	6.6 ± 0.5	17.5 ± 0.5	26.9 ± 0.5
	Joint K-Means					6.3 ± 0.2	19.4 ± 0.9	28.8 ± 0.2	5.8 ± 0.6	16.5 ± 1.0	25.3 ± 0.6
	Only RCCA	✓				6.8 ± 0.5	20.1 ± 0.3	29.5 ± 0.6	5.9 ± 0.3	17.1 ± 0.9	26.0 ± 0.7
	Only WP		✓			6.7 ± 0.9	19.2 ± 0.7	28.2 ± 0.8	5.5 ± 0.6	16.6 ± 0.9	24.3 ± 0.4
	Dual Space	✓	✓			6.2 ± 0.2	19.3 ± 1.0	29.6 ± 0.2	6.1 ± 0.5	17.0 ± 1.0	26.0 ± 0.2
	UPSyn w/o RCR	✓	✓	✓		6.4 ± 0.5	21.0 ± 0.6	30.0 ± 0.8	6.4 ± 0.6	20.2 ± 0.3	30.0 ± 0.7
	UPSyn	✓	✓	✓	✓	**7.3 ± 0.3**	**22.4 ± 0.5**	**32.1 ± 0.3**	**7.4 ± 0.5**	**21.7 ± 0.4**	**31.4 ± 0.8**

4.3 Ablation Studies

Prototype-discovery components. Table 3 isolates the contribution of each component in UPSyn. Compared with PDS, Joint K-Means directly clusters concatenated CLIP image and text representations and improves all six metrics on ResNet, including TR@10 from 39.0 to 41.4, providing initial evidence that cross-modal correspondence matters for prototype discovery. Its less consistent gains on ViT further suggest that naive feature concatenation alone is insufficient for robust cross-architecture generalization. Only RCCA provides more consistent gains across backbones, while Only WP mainly benefits ResNet. Combining the two spaces and applying consensus refinement further strengthens performance; for example, ViT TR@5/10 increase from 17.0/26.0 to 20.2/30.0. Finally, RCR yields additional gains on both backbones, with improvements of up to 2.1 points, and the complete UPSyn achieves the best overall performance across all reported metrics.

Table 4: Region-Aware Caption Refinement on Flickr30K with 100 distilled pairs. Different Top-k retrieved captions for text refinement. Bold and underlining indicate the best and second-best results per backbone.

Eval. Model	Top-k	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	1	10.5 ± 0.2	29.9 ± 0.6	41.8 ± 0.2	14.1 ± 0.6	32.3 ± 0.3	43.5 ± 0.1
	3	10.6 ± 0.1	29.0 ± 0.6	40.9 ± 0.8	12.7 ± 0.6	32.2 ± 0.9	42.8 ± 0.1
	5	10.7 ± 0.1	29.0 ± 0.7	40.8 ± 0.8	12.8 ± 0.8	32.8 ± 0.3	43.1 ± 0.1
	10	**11.3 ± 0.1**	**31.1 ± 0.9**	**42.2 ± 0.2**	**14.9 ± 0.2**	**34.4 ± 0.3**	**44.0 ± 0.8**
	20	10.5 ± 0.3	28.9 ± 0.3	41.0 ± 0.3	12.8 ± 0.3	32.3 ± 0.6	42.9 ± 0.6
ViT	1	6.4 ± 0.5	21.0 ± 0.6	30.0 ± 0.8	6.4 ± 0.6	20.2 ± 0.3	30.0 ± 0.7
	3	6.5 ± 0.6	20.7 ± 0.1	30.3 ± 0.3	6.5 ± 0.2	19.2 ± 0.5	28.3 ± 0.4
	5	**7.3 ± 0.3**	**22.4 ± 0.5**	**32.1 ± 0.3**	**7.4 ± 0.5**	**21.7 ± 0.4**	**31.4 ± 0.8**
	10	6.3 ± 0.6	20.6 ± 0.8	30.4 ± 0.9	5.7 ± 0.8	18.7 ± 0.7	28.9 ± 1.0
	20	6.4 ± 0.1	20.5 ± 0.2	29.7 ± 0.3	6.2 ± 0.2	20.6 ± 0.4	29.8 ± 0.7

Top-k caption aggregation. Table 4 studies the effect of aggregating multiple prototype-relevant captions. On ResNet, increasing k from 1 to 10 improves IR@1 from 10.5 to 11.3, IR@5 from 29.9 to 31.1, and TR@5 from 32.3 to 34.4. On ViT, $k = 5$ performs best overall, improving IR@1 from 6.4 to 7.3 and TR@1 from 6.4 to 7.4. These results indicate that aggregating multiple nearby captions can provide richer prototype-level semantics than relying on a single description. However, further increasing the neighborhood to $k = 20$ does not yield additional gains, suggesting that overly large neighborhoods may introduce redundant or less relevant semantic content. Overall, a moderate number of retrieved captions provides the most effective textual conditioning, with the optimal k varying across evaluation backbones used to assess the resulting distilled image–text pairs.

Efficiency analysis. Table 5 compares the computational cost and retrieval performance of representative generative distillation methods. EDGE incurs a substantially larger preprocessing cost due to generative-model optimization, whereas PDS and UPSyn operate directly on frozen representations. UPSyn without RCR requires only 417.1 seconds of preprocessing and 7.1 seconds per synthesized image, while consistently outperforming PDS across all six retrieval metrics. Introducing RCR adds

Table 5: Efficiency and retrieval performance on Flickr30K with 100 distilled image–text pairs. Preprocessing time measures the cost before image generation, while synthesis time denotes the average generation time per image. Lower is better for time, and higher is better for retrieval.

Method	Preproc. (s) ↓	Synth. (s/img) ↓	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
EDGE	194,190.0	11.5	2.9 ± 0.2	10.6 ± 0.1	16.9 ± 0.3	3.7 ± 0.4	11.2 ± 0.5	16.7 ± 0.5
PDS	620.8	34.2	7.9 ± 0.3	25.8 ± 0.4	37.3 ± 0.3	10.2 ± 0.3	28.2 ± 0.9	39.0 ± 0.3
UPSyn w/o RCR	**417.1**	**7.1**	10.5 ± 0.2	29.9 ± 0.6	41.8 ± 0.2	14.1 ± 0.6	32.3 ± 0.3	43.5 ± 0.1
UPSyn w/ RCR	567.1	7.7	**11.3 ± 0.1**	**31.1 ± 0.9**	**42.2 ± 0.2**	**14.9 ± 0.2**	**34.4 ± 0.3**	**44.0 ± 0.8**

only moderate preprocessing and generation overhead, increasing the costs to approximately 567.1 seconds and 7.7 seconds per image, respectively, while further improving all six retrieval metrics. The complete UPSyn achieves 11.3/31.1/42.2 on IR@1/5/10 and 14.9/34.4/44.0 on TR@1/5/10. These results show that the performance gains of UPSyn do not come at the expense of substantially higher distillation cost, providing a favorable performance–efficiency trade-off for multimodal learning.

4.4 Qualitative Analysis

Qualitative analysis. Figure 3 illustrates how RCR enriches the semantic content around each text prototype. The samples highlighted by blue boxes correspond to the nearest retrieved examples to the prototypes and can be viewed as representative local anchors. As shown in the figure, these nearest samples usually capture only a limited and relatively narrow aspect of the underlying prototype semantics. Starting from each prototype, RCR retrieves its Top-k neighboring captions and aggregates their complementary information into a refined prototype-level description. Consequently, the synthesized samples contain richer and more diverse semantic content. For example, the first group integrates cues related to police officers, street scenes, cars, and motorcycles, while the second combines recurring concepts such as broccoli, vegetables, meat, and plated dishes. These examples show that RCR extends the semantic coverage of a single prototype-neighbor description by exploiting the local semantic neighborhood around the prototype.



Figure 3: Qualitative visualization of RCR on Flickr30K. Each row shows the Top-5 real image–caption pairs retrieved by a text prototype (left) and the image synthesized using the RCR-refined caption together with its LLM-aggregated description (right).

5 Conclusion

UPSyn addresses relevance-agnostic prototype compression by explicitly modeling paired cross-modal structure before prototype discovery. It constructs correlation-oriented RCCA and geometry-oriented WP common spaces and integrates their assignment evidence through consensus refinement to obtain a shared multimodal prototype partition in the original CLIP space. RCR further enriches textual conditioning by consolidating prototype-relevant captions with an LLM. Experiments on Flickr30K and MS-COCO demonstrate consistent improvements over prototype-guided and generative distillation baselines across data budgets and evaluation backbones, with favorable cross-architecture generalization and no iterative synthetic-data optimization. Future work will explore richer cross-modal dependency modeling and scalable prototype representations.

AI Use Statement

Generative AI tools assisted with language polishing and manuscript organization. The authors produced and verified the technical claims, experimental results, and analyses and take responsibility for the final content of this submission.

References

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002. URL https://jmlr.org/papers/v3/bach02a.html.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10718–10727, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Cazenavette_Dataset_Distillation_by_Matching_Training_Trajectories_CVPR_2022_paper.html.

Jeffrey A. Chan Santiago, Praveen Tirupattur, Gaurav Kumar Nayak, Gaowen Liu, and Mubarak Shah. MGD^3: Mode-guided dataset distillation using diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52861–52876, 2025. URL https://proceedings.mlr.press/v267/santiago25a.html.

Junhyeok Choi, Sangwoo Mo, and Minwoo Chae. Multimodal dataset distillation made simple by prototype-guided data synthesis. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4a896a8bb065774169d9ad65f19208b7-Abstract-Conference.html.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006. ISBN 9780471241959. URL https://www.wiley.com/en-dk/Elements+of+Information+Theory%2C+2nd+Edition-p-9780471241959.

Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to ImageNet-1K with constant memory. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6565–6590, 2023. URL https://proceedings.mlr.press/v202/cui23e.html.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. URL https://arxiv.org/abs/2606.19348.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2304.14108.

Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15793–15803, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Gu_Efficient_Dataset_Distillation_via_Minimax_Diffusion_CVPR_2024_paper.html.

Shengbin Guo, Hang Zhao, Senqiao Yang, Chenyang Jiang, Yuhang Cheng, Xiangru Peng, Rui Shao, and Zhuotao Tian. Multimodal dataset distillation via phased teacher models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3a6935d11910d6f9142b0a1e36fc6753-Abstract-Conference.html.

Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fc740286fb4840cb56f99400817967da-Abstract-Conference.html.

David R. Hardoon, Sándor Szedmák, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004. doi: 10.1162/0899766042321814. URL https://direct.mit.edu/neco/article/16/12/2639/6880/Canonical-Correlation-Analysis-An-Overview-with.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://arxiv.org/abs/1512.03385.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3–4):321–377, 1936. doi: 10.1093/biomet/28.3-4.321. URL https://academic.oup.com/biomet/article-abstract/28/3-4/321/220073.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985. doi: 10.1007/BF01908075. URL https://doi.org/10.1007/BF01908075.

Jongoh Jeong, Hoyong Kwon, Minseok Kim, and Kuk-Jin Yoon. Multimodal distribution matching for vision-language dataset distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23072–23082, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jeong_Multimodal_Distribution_Matching_for_Vision-Language_Dataset_Distillation_CVPR_2026_paper.html.

Jongoh Jeong, Sun-Kyung Lee, and Kuk-Jin Yoon. Rank-aware hyperbolic alignment for vision-language dataset distillation. In *European Conference on Computer Vision*, 2026b. URL https://andyj1.github.io/raha/.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018. doi: 10.1080/00031305.2016.1277159. URL https://arxiv.org/abs/1512.00809.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Dong Bok Lee, Seanie Lee, Joonho Ko, Kenji Kawaguchi, Juho Lee, and Sung Ju Hwang. Self-supervised dataset distillation for transfer learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fe42607d1ededbda8b629333b4a639f7-Abstract-Conference.html.

Yongmin Lee and Hye Won Chung. CovMatch: Cross-covariance guided multimodal dataset distillation with trainable text encoder. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2510.18583.

S. E. Leurgans, R. A. Moyeed, and B. W. Silverman. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(3):725–740, 1993. doi: 10.1111/j.2517-6161.1993.tb01936.x. URL https://academic.oup.com/jrsssb/article/55/3/725/7028233.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/702f4db7543a7432431df588d57bc7c9-Abstract-Conference.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. URL https://arxiv.org/abs/1405.0312.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982. doi: 10.1109/TIT.1982.1056489. URL https://doi.org/10.1109/TIT.1982.1056489.

Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=l-PrrQrK0QR.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015. URL https://arxiv.org/abs/1505.04870.

Ding Qi, Jian Li, Shuguang Dou, Zifan Song, Junyao Gao, Yabiao Wang, Chengjie Wang, and Cai Zhao. Asynchronous matching with dynamic sampling for multimodal dataset distillation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/01830c92c6558179fa6d7fb1edff692c-Abstract-Conference.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. URL https://arxiv.org/abs/2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://arxiv.org/abs/2112.10752.

Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451. URL https://link.springer.com/article/10.1007/BF02289451.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002. URL https://jmlr.org/papers/v3/strehl02a.html.

Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. D^4M: Dataset distillation via disentangled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5809–5818, 2024. URL https://arxiv.org/abs/2407.15138.

Duo Su, Huyu Wu, Huanran Chen, Yiming Shi, Yuzhu Wang, Xi Ye, and Jun Zhu. Diffusion models as dataset distillation priors. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2c7b9949db45ccc7fa5bf857a37330da-Abstract-Conference.html.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint arXiv:physics/0004057*, 2000. URL https://arxiv.org/abs/physics/0004057.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010. URL https://jmlr.org/papers/v11/vinh10a.html.

H. D. Vinod. Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4(2): 147–166, 1976. doi: 10.1016/0304-4076(76)90010-5. URL https://www.sciencedirect.com/science/article/pii/0304407676900105.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. URL https://arxiv.org/abs/1811.10959.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1121–1128, 2009. doi: 10.1145/1553374.1553517. URL https://icml.cc/Conferences/2009/papers/447.pdf.

Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2308.07545.

Yue Xu, Zhilin Lin, Yusong Qiu, Cewu Lu, and Yong-Lu Li. Low-rank similarity mining for multimodal dataset distillation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55144–55161, 2024. URL https://proceedings.mlr.press/v235/xu24q.html.

Reza Zanjirani Farahani and Masoud Hekmatfar (eds.). *Facility Location: Concepts, Models, Algorithms and Case Studies*. Contributions to Management Science. Physica-Verlag, 2009. doi: 10.1007/978-3-7908-2151-2. URL https://link.springer.com/book/10.1007/978-3-7908-2151-2.

Xin Zhang, Ziruo Zhang, Jiawei Du, Zuozhu Liu, and Joey Tianyi Zhou. Beyond modality collapse: Representation blending for multimodal dataset distillation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a6a90bcc2aa470c3871b2d39a67d26e8-Abstract-Conference.html.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6514–6523, 2023. URL https://openaccess.thecvf.com/content/WACV2023/html/Zhao_Dataset_Condensation_With_Distribution_Matching_WACV_2023_paper.html.

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

Zhenghao Zhao, Haoxuan Wang, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. Efficient multimodal dataset distillation via generative models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2509.15472.

A Related Work

A.1 Dataset Distillation

Dataset distillation (DD) aims to compress a large-scale training set into a small set of informative synthetic samples such that models trained on the distilled data can retain competitive performance (Wang et al., 2018; Nguyen et al., 2021). Early approaches mainly optimized synthetic samples to reproduce the learning behavior induced by real data. Gradient Matching (Zhao et al., 2021) optimizes synthetic data by matching the gradients produced by real and distilled samples, while Matching Training Trajectories (MTT) (Cazenavette et al., 2022) further constrains synthetic data to reproduce long-term expert training dynamics (Guo et al., 2024). Distribution Matching (DM) (Zhao & Bilen, 2023) instead aligns feature distributions between real and synthetic data, reducing the cost of complex bi-level optimization. More recent methods incorporate generative models to improve scalability (Su et al., 2026; Chan Santiago et al., 2025). For example, Minimax Diffusion (Gu et al., 2024) and D^4M (Su et al., 2024) exploit diffusion models to generate representative and diverse samples for large-scale or high-resolution settings. Despite these advances, most conventional DD methods are designed for unimodal visual data and do not explicitly model cross-modal correspondence or modality-specific representation discrepancies in paired vision–language data.

A.2 Vision–Language Dataset Distillation

Vision–language dataset distillation extends conventional dataset distillation to paired image–text data, requiring the preservation of both intra-modal information and cross-modal correspondence. Early work such as VLDD (Wu et al., 2024) adapts trajectory matching to jointly distill images and text representations, while LoRS (Xu et al., 2024) further preserves low-rank image–text similarity structures. Subsequent methods explore different forms of multimodal structure preservation: RepBlend (Zhang et al., 2025) addresses modality collapse, CovMatch (Lee & Chung, 2025) matches cross-covariance statistics, PTM-ST (Guo et al., 2026) and AMD (Qi et al., 2026) improve trajectory-based optimization, and MDM (Jeong et al., 2026a) and RAHA (Jeong et al., 2026b) model multimodal geometry and latent structure. Generative approaches such as EDGE (Zhao et al., 2025) and PDS (Choi et al., 2026) further reduce the reliance on direct synthetic-data optimization. In contrast, our work follows the efficient prototype-guided synthesis paradigm and explicitly incorporates both cross-modal correlation and geometric consistency into multimodal prototype discovery for paired image–text data.

B Additional Methodological Details

B.1 Derivation of the RCCA Common Space

Let $\mathbf{X} = [\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_N]$ and $\mathbf{T} = [\bar{\mathbf{t}}_1, \ldots, \bar{\mathbf{t}}_N]$ in $\mathbb{R}^{d\times N}$ collect the centered CLIP representations. Their empirical covariance matrices are

$$\mathbf{\Sigma}_{II} = \frac{1}{N}\mathbf{X}\mathbf{X}^\top, \qquad \mathbf{\Sigma}_{TT} = \frac{1}{N}\mathbf{T}\mathbf{T}^\top, \qquad \mathbf{\Sigma}_{IT} = \frac{1}{N}\mathbf{X}\mathbf{T}^\top. \tag{11}$$

Using the additive regularization notation of the main text, define $\widetilde{\mathbf{\Sigma}}_{II} = \mathbf{\Sigma}_{II}+\lambda_I\mathbf{I}$ and $\widetilde{\mathbf{\Sigma}}_{TT} = \mathbf{\Sigma}_{TT}+\lambda_T\mathbf{I}$. RCCA (Vinod, 1976; Leurgans et al., 1993) solves

$$\begin{aligned} \max_{\mathbf{a},\mathbf{b}} \quad & \mathbf{a}^\top\mathbf{\Sigma}_{IT}\mathbf{b}, \\ \text{s.t.} \quad & \mathbf{a}^\top\widetilde{\mathbf{\Sigma}}_{II}\mathbf{a} = 1, \qquad \mathbf{b}^\top\widetilde{\mathbf{\Sigma}}_{TT}\mathbf{b} = 1. \end{aligned} \tag{12}$$

The singular value decomposition $\widetilde{\mathbf{\Sigma}}_{II}^{-1/2}\mathbf{\Sigma}_{IT}\widetilde{\mathbf{\Sigma}}_{TT}^{-1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ gives the leading d_C projection directions,

$$\mathbf{P}_I = \widetilde{\mathbf{\Sigma}}_{II}^{-1/2}\mathbf{U}_{d_C}, \qquad \mathbf{P}_T = \widetilde{\mathbf{\Sigma}}_{TT}^{-1/2}\mathbf{V}_{d_C}. \tag{13}$$

The unweighted coordinates are $\mathbf{q}_{i,I} = \mathbf{P}_I^\top\bar{\mathbf{x}}_i$ and $\mathbf{q}_{i,T} = \mathbf{P}_T^\top\bar{\mathbf{t}}_i$. Applying the canonical-correlation weighting in Eq. (6) yields $\mathbf{z}_{i,I}^C = \mathbf{\Lambda}_{d_C}^c\mathbf{q}_{i,I}$ and $\mathbf{z}_{i,T}^C = \mathbf{\Lambda}_{d_C}^c\mathbf{q}_{i,T}$. This distinguishes estimation of the canonical directions from their subsequent weighting for clustering.

B.2 Derivation of the Whitening–Procrustes Common Space

WP combines regularized whitening with orthogonal alignment. In the full-dimensional formulation, the whitening operators are $\mathbf{W}_I = \widetilde{\mathbf{\Sigma}}_{II}^{-1/2}$ and $\mathbf{W}_T = \widetilde{\mathbf{\Sigma}}_{TT}^{-1/2}$. They produce $\mathbf{X}_w = \mathbf{W}_I\mathbf{X}$ and $\mathbf{T}_w = \mathbf{W}_T\mathbf{T}$. The same Procrustes derivation applies when the whitening operators map both modalities to a lower-dimensional space, as in our experiments with a common output dimensionality for both modalities.

Whitening reduces covariance and scale differences but does not determine the relative orientation of the modalities. Orthogonal Procrustes (Schönemann, 1966) estimates this orientation by

$$\begin{aligned}\mathbf{R}^* &= \arg\min_{\mathbf{R}^\top\mathbf{R}=\mathbf{I}} \|\mathbf{X}_w - \mathbf{R}\mathbf{T}_w\|_F^2 \\ &= \arg\max_{\mathbf{R}^\top\mathbf{R}=\mathbf{I}} \operatorname{tr}(\mathbf{R}^\top\mathbf{X}_w\mathbf{T}_w^\top).\end{aligned} \tag{14}$$

The equivalence follows by expanding the squared norm: the two squared-norm terms are constant under an orthogonal transformation. With $\mathbf{X}_w\mathbf{T}_w^\top = \mathbf{U}_G\mathbf{\Lambda}_G\mathbf{V}_G^\top$, the solution is $\mathbf{R}^* = \mathbf{U}_G\mathbf{V}_G^\top$. Hence

$$\mathbf{z}_{i,I}^G = \mathbf{W}_I\bar{\mathbf{x}}_i, \qquad \mathbf{z}_{i,T}^G = \mathbf{R}^*\mathbf{W}_T\bar{\mathbf{t}}_i. \tag{15}$$

For any whitened vectors $\mathbf{u}, \mathbf{v}$, orthogonality gives

$$\|\mathbf{R}^*\mathbf{u} - \mathbf{R}^*\mathbf{v}\|_2 = \|\mathbf{u} - \mathbf{v}\|_2, \qquad (\mathbf{R}^*\mathbf{u})^\top(\mathbf{R}^*\mathbf{v}) = \mathbf{u}^\top\mathbf{v}. \tag{16}$$

The alignment step therefore preserves distances and angles in the whitened text space. This property does not imply that whitening preserves the geometry of the original, unwhitened CLIP features.

B.3 Analysis of Consensus Refinement

For pair i and cluster k, the consensus score is

$$S_{ik} = \tfrac{1}{2}(s_{ik}^C + s_{ik}^G) - \gamma|s_{ik}^C - s_{ik}^G|. \tag{17}$$

Let $s_{ik}^{\max} = \max(s_{ik}^C, s_{ik}^G)$ and $s_{ik}^{\min} = \min(s_{ik}^C, s_{ik}^G)$. Then

$$S_{ik} = (\tfrac{1}{2} - \gamma)s_{ik}^{\max} + (\tfrac{1}{2} + \gamma)s_{ik}^{\min}. \tag{18}$$

For $0 \le \gamma \le 1/2$, the coefficients are nonnegative and sum to one. Increasing γ shifts weight toward the weaker supporting view. At $\gamma = 0$, the score is the arithmetic mean; at $\gamma = 1/2$, it reduces to $\min(s_{ik}^C, s_{ik}^G)$. Our default $\gamma = 0.1$ balances average support with agreement. This interpretation concerns the assignment rule; empirical convergence is examined in Appendix E.4.

B.4 Complete Algorithm

Algorithm 1 summarizes common-space construction, consensus refinement, prototype reconstruction, RCR-based caption refinement, and prototype-guided synthesis. The iteration limit $L_{\max}$ is distinguished from the distilled-set size M in the problem formulation. After consensus refinement converges, the final cluster assignments are used to reconstruct image and text prototypes in the original CLIP space, followed by RCR and fixed-budget unCLIP synthesis.

B.5 Prompt Template for RCR

For completeness, we provide the full prompt used in Region-Aware Caption Refinement (RCR). Given the Top-K_t captions retrieved around a text prototype, the LLM is instructed to consolidate their shared and complementary semantic information into a single prototype-level description. The prompt template is as follows:

```
You will obtain the Top-{K} most similar texts and
synthesize them into a single, semantically rich prototype
description.
## Input Texts:
{TEXTS}
## Guidelines:
```

Algorithm 1 UPSyn: Unified Prototype-Guided Synthesis

Require: Paired data $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^{N}$; frozen CLIP encoders f_I, f_T; prototype count K; common-space settings; fusion weight α; disagreement weight γ; iteration limit $L_{\max}$; tolerance ϵ; caption count K_t; LLM-based caption aggregator $\mathcal{A}$; pre-trained unCLIP decoder G.
Ensure: Distilled dataset $\mathcal{S}$.

Extract $\mathbf{x}_i = f_I(I_i)$ and $\mathbf{t}_i = f_T(T_i)$ for all i.
Construct RCCA and WP pair-level representations $\{\mathbf{z}_i^C, \mathbf{z}_i^G\}_{i=1}^{N}$ using Eqs. (6)–(8).
Initialize shared assignments $\{c_i\}_{i=1}^{N}$ by K-Means on $[\mathbf{z}_i^C/\sqrt{2}; \mathbf{z}_i^G/\sqrt{2}]$.
for $m = 1, \ldots, L_{\max}$ **do**
    Form clusters $C_k = \{i : c_i = k\}$ and update dual-space centroids $\boldsymbol{\mu}_k^C$ and $\boldsymbol{\mu}_k^G$.
    Compute affinities $s_{ik}^s = \mathrm{sim}(\mathbf{z}_i^s, \boldsymbol{\mu}_k^s)$ for all i, k and $s \in \{C, G\}$.
    Compute consensus scores $S_{ik} = \frac{1}{2}(s_{ik}^C + s_{ik}^G) - \gamma|s_{ik}^C - s_{ik}^G|$ for all i, k.
    Update $c_i^{\text{new}} \leftarrow \arg\max_k S_{ik}$ for all i.
    Compute assignment-change ratio $\delta \leftarrow N^{-1} \sum_i \mathbb{I}[c_i^{\text{new}} \neq c_i]$.
    Set $c_i \leftarrow c_i^{\text{new}}$ for all i.
    if $\delta < \epsilon$ **then**
        break
    end if
end for
Form the final clusters $\{C_k\}_{k=1}^{K}$ from the converged assignments.
Reconstruct image and text prototypes $\{\mathbf{p}_k^I, \mathbf{p}_k^T\}_{k=1}^{K}$ in the original CLIP space.
for $k = 1, \ldots, K$ **do**
    Retrieve the Top-K_t captions nearest to $\mathbf{p}_k^T$ according to CLIP text cosine similarity.
    Refine the retrieved captions with RCR: $\widetilde{T}_k \leftarrow \mathcal{A}(T_{k,1}, \ldots, T_{k,K_t})$.
    Synthesize $\widetilde{I}_k \leftarrow G(\mathbf{p}_k^I, \widetilde{T}_k; 20 \text{ steps})$.
end for
return $\mathcal{S} = \{(\widetilde{I}_k, \widetilde{T}_k)\}_{k=1}^{K}$.

```
1. Identify recurring semantic concepts across the
retrieved texts, including entities, attributes, actions,
and scene context.
2. Prioritize semantic information that is consistently
supported by multiple texts.
3. Merge compatible complementary details from different
samples when they remain consistent with the shared semantic
theme.
4. Express the common semantics in a generalized and
coherent form rather than copying the wording or sentence
structure of individual captions.
5. Do not overemphasize details that appear in only one
sample unless they provide useful information consistent
with the remaining texts.
6. Preserve fine-grained semantic identities whenever they
are supported by the retrieved evidence. For example, use
``dingo'' rather than the broader term ``dog'' when the
specific species can be reliably identified.
7. Do not introduce objects, attributes, relationships, or
events that are not supported by the retrieved texts.
8. The generated description should represent the semantic
center of the retrieved neighborhood rather than any
individual caption.
9. Keep the output within 220 tokens and use fluent,
precise, and natural language.
Return only the final prototype description without
explanations or additional formatting.
```

C Additional Experimental Details

C.1 Implementation Settings

Table 6 summarizes the default configuration. Unless otherwise specified, all prototypes are synthesized with a fixed 20-step denoising budget. The final UPSyn configuration employs RCR with Top-k retrieved captions for textual refinement, while the corresponding ablation studies vary the caption neighborhood size and other method-specific parameters as described in the main text.

Table 6: Default implementation settings for UPSyn.

Component	Setting
Feature encoder	CLIP ViT-L/14 (Radford et al., 2021)
CLIP feature dimension	768
Common-space dimension	256
RCCA covariance shrinkage	0.1
RCCA ridge regularization	10^{-3}
WP whitening shrinkage	0.1
Modality fusion coefficient α	0.5
Consensus coefficient γ	0.1
Maximum consensus iterations	30
Convergence tolerance	10^{-3}
Image generator	Pre-trained unCLIP decoder
Classifier-free guidance scale	5.0
Evaluation image resolution	224×224
Distilled dataset sizes	100 and 300 pairs
Evaluation backbones	ResNet-50 and ViT-Ti/16
Hardware	Single NVIDIA T4 GPU

C.2 Cross-Architecture Evaluation Protocol

Following Xu et al. (2024); Choi et al. (2026), we evaluate whether data distilled using CLIP features can train alternative vision backbones. ResNet-50 and ViT-Ti/16 are trained on the distilled image–text pairs with a linear projection, while the text encoder remains frozen. We report IR@1/5/10 and TR@1/5/10. When an overall mean is used in the supplementary analyses, it is the arithmetic average of these six metrics; the cross-backbone mean further averages the two evaluation backbones.

D Baseline Details

We compare optimization-based MDD, generative distillation, and subset selection under the evaluation protocol described in the main paper for training and evaluating the distilled image–text pairs.

D.1 Dataset Subset Selection

Herding. Herding (Welling, 2009) iteratively selects samples so that the subset feature mean approximates the full-data mean in the chosen representation space.

K-center. The K-center baseline (Zanjirani Farahani & Hekmatfar, 2009) expands coverage by selecting the sample with the largest minimum distance to the current subset in the representation space.

CLIP score filtering. Image–text pairs are ranked by the cosine similarity of their CLIP representations. Selecting the highest-scoring pairs favors alignment but does not explicitly enforce coverage of the data distribution across the available training pairs.

LAION filtering. Following the filtering protocol of Gadre et al. (2023), this baseline removes low-quality or unsuitable text and then applies CLIP-based image–text similarity filtering.

Image-based filtering. Visual embeddings identify representative regions of the image distribution, followed by CLIP-based filtering to construct the final subset (Gadre et al., 2023).

D.2 Optimization-Based Multimodal Distillation

TESLA-VL. TESLA-VL (Xu et al., 2024; Cui et al., 2023) adapts memory-efficient trajectory matching to vision–language distillation. It reduces the memory and computational costs of conventional trajectory matching and uses a weighted binary cross-entropy objective in place of standard contrastive training for the distilled image–text pairs.

LoRS. LoRS (Xu et al., 2024) extends TESLA-VL with a learnable low-rank similarity matrix. It optimizes the synthetic multimodal data and their similarity structure to preserve cross-modal relationships within the compact distilled image–text dataset.

D.3 Learning-Free Generative Baselines Adapted to MDD

D^{4}M. D^{4}M (Su et al., 2024) extracts VAE features and clusters them into prototypes. Forward diffusion corrupts these prototypes to initialize reverse diffusion for image synthesis. For multimodal evaluation, we retain the VAE image encoder, add a CLIP text encoder, and match the image and text clusters before applying the original synthesis procedure to generate the distilled images.

MGD^3. MGD^3 (Chan Santiago et al., 2025) also discovers prototypes in VAE feature space but uses them as explicit guidance during reverse diffusion rather than only for initialization. We adapt it by incorporating CLIP text features and matching the resulting image and text prototypes before prototype-guided synthesis of the distilled images.

D.4 Generative Multimodal Dataset Distillation

EDGE. EDGE (Zhao et al., 2025) trains a diffusion model with bidirectional contrastive and diversity objectives to improve image–text correspondence and variation among generated samples. A multimodal large language model produces additional captions for the synthesized images. Unlike UPSyn, EDGE optimizes the generative model during distillation on the target image–text dataset.

PDS. PDS (Choi et al., 2026) independently clusters CLIP image and text features, matches the clusters, and uses their prototypes to guide a frozen unCLIP decoder. UPSyn retains this learning-free synthesis paradigm but determines prototype membership jointly through cross-modal common spaces and consensus over paired image–text representations.

E Additional Ablation Studies

E.1 Sensitivity to α and γ

Table 7 varies the modality fusion weight α with $\gamma = 0.1$ and the disagreement penalty γ with $\alpha = 0.5$. Balanced fusion provides a favorable overall configuration, although individual metrics prefer different values: $\alpha = 0.3$ gives the highest IR@1, whereas $\alpha = 0.5$ gives the highest IR@5, IR@10, and TR@1. Removing the disagreement penalty ($\gamma = 0$) reduces all six metrics relative to $\gamma = 0.1$. Larger penalties do not uniformly improve results; $\gamma = 0.5$ favors IR@1 and IR@10 but gives lower text-retrieval scores. We therefore use $\alpha = 0.5$ and $\gamma = 0.1$ as a balanced default.

E.2 Canonical-Correlation Weighting

Table 8 varies $c \in \{0, 0.5, 1\}$ at a fixed RCCA dimension of 256, with RCR enabled for all settings. On ResNet, $c = 1$ improves text retrieval relative to $c = 0$, including TR@10 from 40.7 to 44.0, but slightly reduces image retrieval. On ViT, it improves all six metrics, including IR@10 from 27.6 to 29.8 and TR@10 from 25.1 to 27.6. These results support retaining canonical-correlation weighting, while showing that its effect differs across retrieval directions. We use $c = 1$ in the default configuration for the correlation-oriented common space.

Table 7: Sensitivity to α and γ on Flickr30K with 100 pairs and ResNet evaluation. We fix $\gamma = 0.1$ when varying α, and $\alpha = 0.5$ when varying γ. Bold and underlining indicate the best and second-best results in each group.

Parameter	Value	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
α ($\gamma = 0.1$)	0.3	**11.0 ± 0.1**	29.6 ± 0.1	41.4 ± 0.3	12.6 ± 0.2	31.9 ± 0.7	43.0 ± 0.2
	0.4	9.9 ± 0.3	29.2 ± 0.6	40.7 ± 0.4	12.9 ± 0.8	32.3 ± 0.3	**43.6 ± 0.3**
	0.5	10.5 ± 0.2	**29.9 ± 0.6**	**41.8 ± 0.2**	**14.1 ± 0.6**	**32.3 ± 0.3**	43.5 ± 0.1
	0.6	10.3 ± 0.1	28.7 ± 0.4	40.6 ± 0.3	12.0 ± 0.3	30.9 ± 0.3	42.1 ± 0.5
	0.7	10.6 ± 0.3	28.6 ± 0.2	40.8 ± 0.2	13.3 ± 0.4	30.6 ± 0.4	41.0 ± 0.3
γ ($\alpha = 0.5$)	0	8.8 ± 0.2	27.6 ± 0.4	39.2 ± 0.4	10.9 ± 0.3	31.4 ± 0.4	42.4 ± 0.8
	0.05	10.5 ± 0.3	28.8 ± 0.3	40.4 ± 0.4	11.7 ± 0.3	29.2 ± 0.3	41.5 ± 0.3
	0.1	10.5 ± 0.2	**29.9 ± 0.6**	41.8 ± 0.2	**14.1 ± 0.6**	**32.3 ± 0.3**	**43.5 ± 0.1**
	0.2	9.3 ± 0.2	28.2 ± 0.2	40.0 ± 0.2	12.2 ± 0.3	30.2 ± 0.4	41.2 ± 0.6
	0.5	**10.8 ± 0.1**	29.6 ± 0.3	**42.0 ± 0.2**	12.8 ± 0.2	31.6 ± 0.3	43.2 ± 1.1

Table 8: Canonical-correlation weighting in RCCA on Flickr30K at output dimension 256. Results are mean ± standard deviation. Bold and underlining indicate the best and second-best results per backbone, respectively.

Eval. Model	c	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	0	**9.8 ± 0.1**	28.5 ± 0.2	**40.3 ± 0.1**	11.6 ± 0.4	30.4 ± 0.3	40.7 ± 0.8
	0.5	9.7 ± 0.3	**28.7 ± 0.3**	40.3 ± 0.3	11.9 ± 0.2	32.0 ± 0.5	43.5 ± 0.7
	1	9.1 ± 0.4	28.1 ± 0.8	39.8 ± 0.9	**12.8 ± 0.3**	**32.4 ± 0.5**	**44.0 ± 0.7**
ViT	0	6.3 ± 0.2	18.9 ± 0.2	27.6 ± 0.2	5.4 ± 0.3	16.0 ± 0.2	25.1 ± 0.9
	0.5	5.6 ± 0.2	19.0 ± 0.2	28.2 ± 0.3	5.4 ± 0.5	15.9 ± 0.3	24.9 ± 0.6
	1	**6.4 ± 0.5**	**20.3 ± 0.3**	**29.8 ± 0.6**	**7.1 ± 0.3**	**18.7 ± 0.9**	**27.6 ± 0.7**

E.3 Common-Space Dimensionality

Table 9 compares $d \in \{128, 256, 512\}$,with RCR enabled for all settings. RCCA at 256 dimensions provides strong text retrieval on both backbones, whereas 512 dimensions can favor image retrieval. WP at 128 dimensions performs best on most ResNet metrics; at 256 dimensions it gives stronger ViT results, including IR@1 of 7.5 and TR@10 of 29.6. Thus, increasing dimension is not uniformly beneficial. The 256-dimensional configuration provides a practical cross-architecture compromise and is used for both spaces in UPSyn. Smaller spaces may discard useful structure, whereas larger spaces may retain additional modality-specific variation beyond the shared cross-modal structure.

Table 9: Common-space dimensionality on Flickr30K. We vary the RCCA and WP output dimensions while holding the other settings fixed. Results are mean ± standard deviation. Bold and underlining indicate the best and second-best results within each common space and evaluation backbone.

Eval. Model	Space	Dim.	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	RCCA	128	9.5 ± 0.1	27.7 ± 0.3	**40.0 ± 0.5**	12.0 ± 0.4	30.8 ± 1.0	41.8 ± 0.6
		256	9.1 ± 0.4	28.1 ± 0.8	39.8 ± 0.9	**12.8 ± 0.3**	**32.4 ± 0.5**	**44.0 ± 0.7**
		512	**10.4 ± 0.1**	**28.3 ± 0.3**	39.8 ± 0.2	12.0 ± 0.2	31.1 ± 0.8	42.3 ± 0.6
	WP	128	**10.1 ± 0.3**	28.9 ± 0.3	41.4 ± 0.4	**13.5 ± 0.5**	**33.0 ± 0.5**	**43.3 ± 0.9**
		256	9.9 ± 0.7	**29.8 ± 0.3**	**42.1 ± 0.1**	12.6 ± 0.2	31.1 ± 0.1	42.5 ± 0.4
		512	8.8 ± 0.3	27.7 ± 0.3	39.8 ± 0.2	11.2 ± 0.2	29.6 ± 0.6	40.1 ± 0.5
ViT	RCCA	128	6.2 ± 0.3	19.2 ± 0.2	28.7 ± 0.3	6.2 ± 0.4	17.5 ± 0.8	26.4 ± 0.5
		256	6.4 ± 0.5	20.3 ± 0.3	29.8 ± 0.6	**7.1 ± 0.3**	**18.7 ± 0.9**	27.6 ± 0.7
		512	**6.4 ± 0.1**	**20.8 ± 0.3**	**30.7 ± 0.2**	5.7 ± 0.2	18.5 ± 0.6	**27.6 ± 0.4**
	WP	128	6.3 ± 0.1	18.7 ± 0.2	27.9 ± 0.2	4.7 ± 0.4	16.4 ± 0.7	25.4 ± 0.8
		256	**7.5 ± 0.9**	**20.1 ± 0.7**	**29.1 ± 0.8**	**6.3 ± 0.6**	**18.2 ± 0.9**	**29.6 ± 0.4**
		512	5.6 ± 0.1	18.4 ± 0.3	28.1 ± 0.3	5.8 ± 0.2	**18.2 ± 0.6**	26.5 ± 0.3

E.4 Convergence of Consensus Refinement

Figure 4 examines Flickr30K with $K = 100$. Assignment changes decline from several percent to approximately 0.1% after 28 iterations, despite local fluctuations. The mean consensus score increases and then saturates. In this experiment, refinement stabilizes within the preset 30-iteration limit; the plot is empirical evidence of convergence, not a general convergence guarantee.

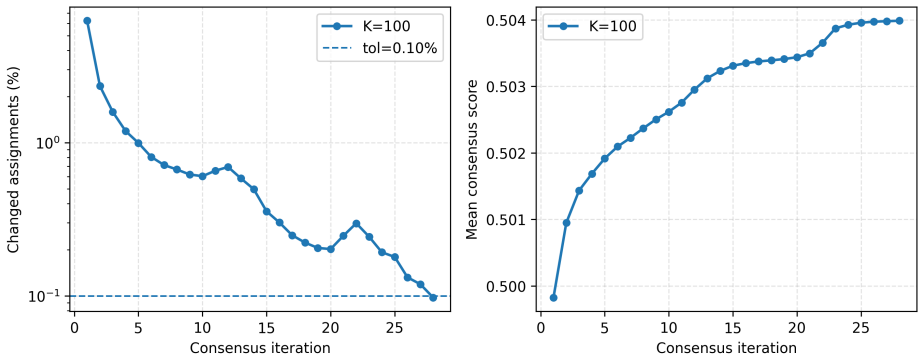

Figure 4: Consensus refinement on Flickr30K with $K = 100$. Left: percentage of changed assignments, with tolerance 0.1% shown as a dashed line. Right: mean consensus score. The observed refinement stabilizes within 30 iterations.

F Further Analysis

F.1 Comparison of the Common Spaces

Table 10 compares CCA, KCCA, and RCCA as candidate correlation-oriented common spaces (Hardoon et al., 2004). To isolate the effect of the common-space construction, all variants are evaluated without RCR. RCCA achieves the highest paired image–text similarity of 0.628, compared with 0.483 for CCA and 0.497 for KCCA. This stronger cross-modal alignment is also reflected in downstream retrieval performance. Under ResNet evaluation, RCCA achieves the best results across all six metrics, reaching 9.4/28.5/40.7 on IR@1/5/10 and 12.7/31.0/42.2 on TR@1/5/10. A similar trend is observed on ViT, where RCCA again performs best on all six retrieval metrics. In contrast, KCCA does not consistently improve over linear CCA despite introducing nonlinear features. These results indicate that regularization provides a more effective and stable correlation-oriented representation for prototype discovery, motivating our use of RCCA in UPSyn.

Table 10: Comparison of candidate correlation-oriented common spaces on Flickr30K with 100 distilled pairs. Paired similarity denotes the mean cosine similarity between matched image and text features.

Eval.	Space	Paired Sim. ↑	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ResNet	CCA	0.483	8.8 ± 0.1	27.6 ± 0.4	40.0 ± 0.2	12.1 ± 0.3	30.8 ± 0.6	40.4 ± 0.4
	KCCA	0.497	9.1 ± 0.2	27.1 ± 0.3	39.3 ± 0.2	12.6 ± 0.6	29.8 ± 1.0	40.8 ± 0.7
	RCCA	**0.628**	**9.4 ± 0.4**	**28.5 ± 0.8**	**40.7 ± 0.9**	**12.7 ± 0.3**	**31.0 ± 0.5**	**42.2 ± 0.7**
ViT	CCA	0.483	6.5 ± 0.2	19.5 ± 0.1	28.8 ± 0.4	5.6 ± 0.3	16.9 ± 0.6	25.9 ± 0.5
	KCCA	0.497	6.1 ± 0.2	19.8 ± 0.2	29.3 ± 0.2	5.1 ± 0.2	15.7 ± 0.6	25.5 ± 0.8
	RCCA	**0.628**	**6.8 ± 0.5**	**20.1 ± 0.3**	**29.5 ± 0.6**	**5.9 ± 0.3**	**17.1 ± 0.9**	**26.0 ± 0.7**

Figure 5 shows complementary retrieval behavior, with the reported results taken from the 256-dimensional setting in Table 9. WP gives higher mean image retrieval (27.25 versus 25.68), while RCCA gives higher mean text retrieval (29.75 versus 28.73). Their partitions have adjusted Rand index (ARI) (Hubert & Arabie, 1985) 0.445 and normalized mutual information (NMI) (Vinh et al., 2010) 0.695, indicating substantial but incomplete agreement. The two views therefore provide related, non-identical assignment evidence, motivating consensus refinement rather than relying on either view alone (Strehl & Ghosh, 2002).

G Qualitative Examples

Figures 6 and 7 show synthetic images and their retrieved captions on Flickr30K and MS-COCO, respectively, illustrating the distilled image–text pairs on both datasets.

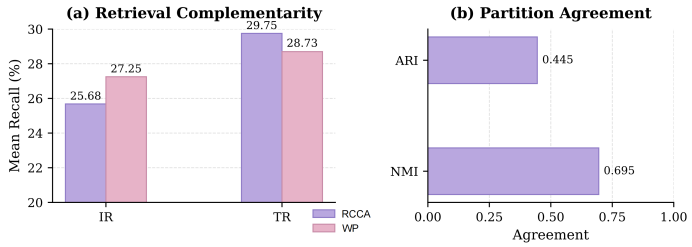

Figure 5: Complementarity of RCCA and WP. Left: mean recall for the two retrieval directions. Right: agreement between their cluster partitions, measured by ARI and NMI.



Figure 6: Synthetic images and retrieved captions on Flickr30K.

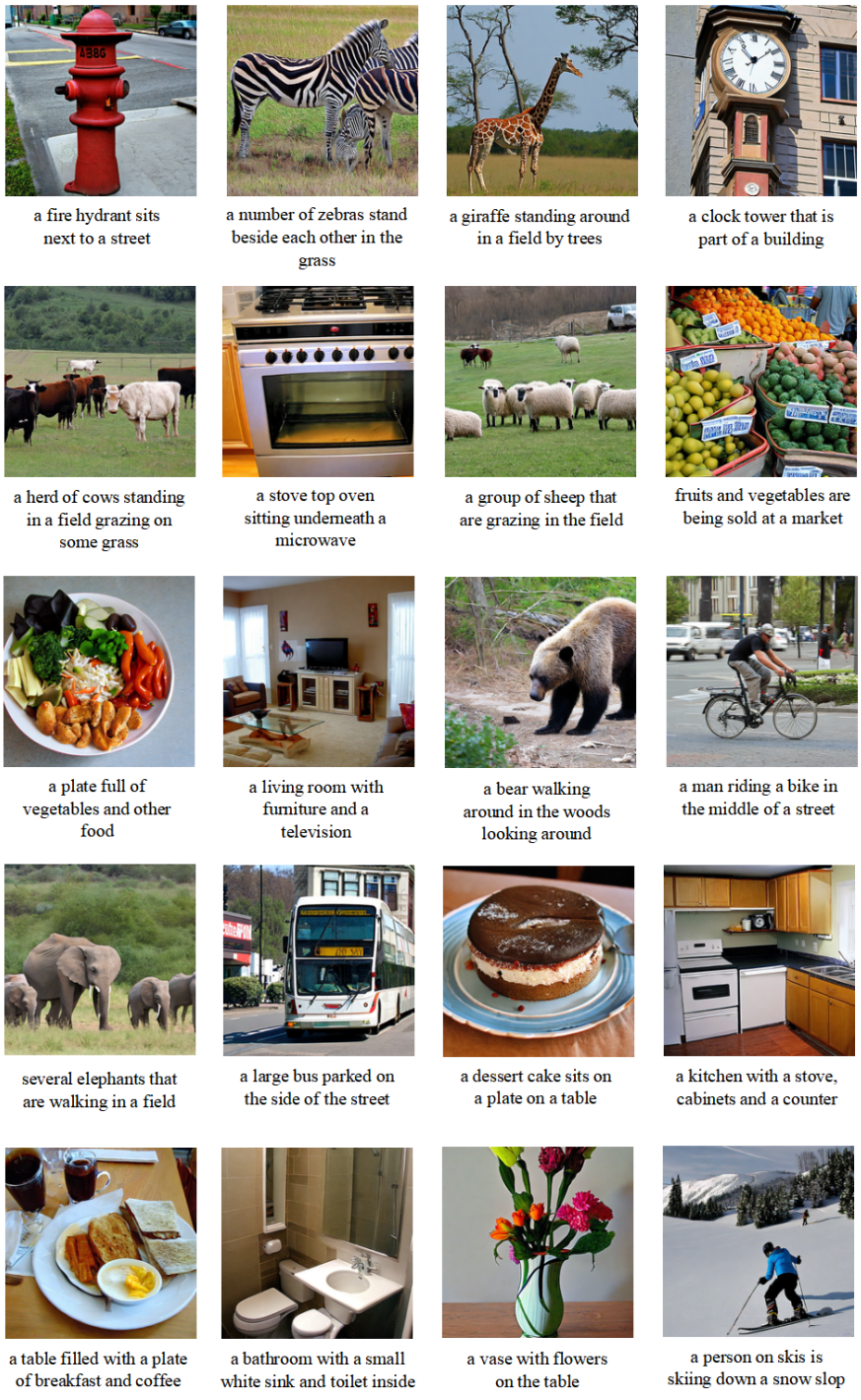

Figure 7: Synthetic images and retrieved captions on MS-COCO.