RESOLVING SEMANTIC DEGENERACY IN MULTIMODAL DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based methods for dataset distillation (DD) cluster images into prototypes and decode them with diffusion models into a synthetic set, becoming a prominent DD paradigm due to their training-free, architecture-agnostic design. Existing methods transfer this pipeline to multimodal dataset distillation (MDD) on paired image–text data by clustering each modality separately at the target budget and keep the DD decode process unchanged. However, per-modality clustering ignores cross-modal interactions and collapses distinct pairs, budget-sized clustering merges fine-grained semantics into meaningless prototypes, and standard diffusion guidance renders samples from distinct prototypes indistinguishable. We term these failures semantic degeneracy and propose DPRG (Distinction-Preserving Representation and Guidance). On the representation side, we assign each pair jointly and over-cluster before selecting back to the budget, keeping semantically distinct pairs in separate prototypes. On the guidance side, we denoise each target against its hardest image- and text-side alternatives, keeping distinct prototypes apart during synthesis. Extensive experiments on Flickr30K and MS-COCO demonstrate that DPRG outperforms existing baselines across all evaluated distillation budgets and downstream architectures.

1 INTRODUCTION

Modern training sets have grown to a scale at which every architecture comparison, hyperparameter search, or controlled ablation can require another pass through the full dataset. Dataset distillation (DD) addresses this by replacing a large training set with a small synthetic one on which training still yields a comparable model (Wang et al., 2018; Zhao & Bilen, 2021; Cazenavette et al., 2022). Multimodal dataset distillation (MDD) extends this to image–text corpora (Radford et al., 2021; Schuhmann et al., 2022), which are increasingly important for training vision-language models and now reach millions or even billions of pairs. Diffusion-based dataset distillation has recently emerged as a prominent training-free paradigm for image classification (Chan-Santiago et al., 2025; Su et al., 2024). Instead of optimizing synthetic samples through costly training dynamics, these methods cluster real images within each class in a frozen embedding space and decode representative prototypes with a pretrained diffusion model. This cluster-then-generate pipeline avoids repeated model training, reduces distillation cost, and produces synthetic sets that transfer across downstream architectures, making it attractive for scalable dataset compression.

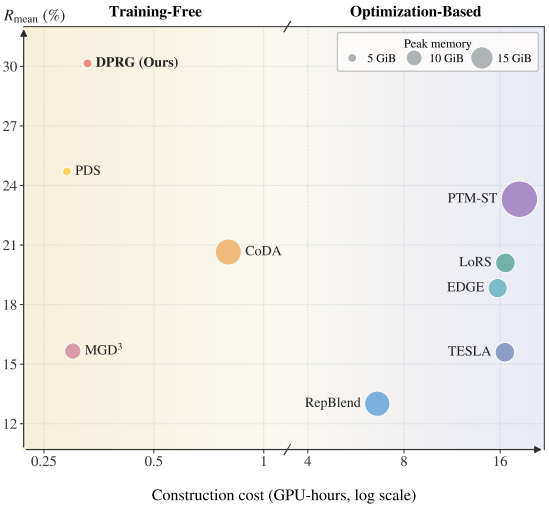


Figure 1: **Accuracy–cost trade-off (Flickr30K, 100 image–text pairs).** DPRG (Ours) achieves the best accuracy with a small construction budget.

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.

However, transferring this paradigm to MDD introduces a difficulty that conventional DD does not face, since multimodal data come with no predefined semantic partition. In image classification, class labels already specify the major semantic units, and clustering only needs to summarize variation within each class. An image–text corpus is instead an unstructured collection of paired observations whose semantic modes must themselves be discovered. Existing diffusion-based MDD methods like PDS (Choi et al., 2026) address this through independently clustering image and caption embeddings into as many groups as the distillation budget, matching two partitions, and merging each matched cluster into a paired prototype that a pretrained diffusion model decodes into one synthetic pair. Setting the number of clusters to the budget, the pipeline aims to cover the data distribution as comprehensively as the budget allows.

Coverage alone is not all that downstream multimodal learning asks for. The dominant way to learn image–text correspondence is contrastive, aligning each matched pair by separating it from the others (Radford et al., 2021; Li et al., 2021), yet a pipeline built for coverage can collapse different semantic modes into nearly identical synthetic pairs, a failure we call *semantic degeneracy*. An effective distilled set must therefore be both semantically representative and semantically distinct, covering the rich cross-modal semantics of the original data under a limited budget while maintaining clear boundaries between different meanings. To achieve this, we propose DPRG (Distinction-Preserving Representation and Guidance), which addresses semantic degeneracy at the three points where it arises.

First, cross-modal relations must be preserved during mode discovery. Clustering each modality separately captures only its marginal structure, and matching the two partitions afterward cannot recover relations that were never represented. DPRG therefore introduces pair coupling, which assigns each pair as a whole so that visual similarity, textual similarity, and cross-modal correspondence jointly determine the discovered semantic modes.

Second, the budget should decide how many semantic modes are kept, not how many exist. Clustering directly into M groups answers both with one number, forcing each cluster to absorb several of the far more numerous modes in an image–text corpus and, without class labels to confine it to a single meaning, leaving its mean between those modes rather than on any of them (Figure 4). DPRG therefore applies over-clustering to expose finer semantic modes, then selects M mutually distant candidates by farthest-first coverage.

Third, a diffusion prior is optimized for likelihood rather than distinctness, leaving too little contrast among the generated pairs for contrastive learning to exploit (Ho & Salimans, 2022). DPRG therefore uses triple CFG: for each target, it constructs image-side and text-side counterfactual conditions from the most confusable alternatives in the distilled set, then contrasts their denoising predictions with the positive branch during early denoising.

Our contributions are as follows.

- We identify *semantic degeneracy*, the collapse of distinct semantic modes into near-identical synthetic pairs that arises when the class-conditional diffusion distillation recipe is transferred to image–text data, and trace it to both prototype discovery and synthesis.
- We propose DPRG, which couples each observed pair during mode discovery, over-clusters beyond the budget and selects M separated modes by farthest-first coverage, and uses hard-negative Triple CFG to contrast positive and counterfactual denoising predictions before synthesis.
- Across Flickr30K, MS-COCO, Flickr8K, and Pascal1K, DPRG improves 113 of 120 detailed recall comparisons and raises R_{mean} in all 20 dataset–budget–backbone settings, and the gain survives a joint change of representation and generator to SigLIP and FLUX.1-dev.

2 Related Work

Dataset distillation and multimodal data distillation. Dataset distillation synthesizes compact training sets that retain the utility of much larger datasets (Wang et al., 2018). Most methods optimize images by matching gradients, features, distributions, or training trajectories (Zhao & Bilen, 2021; 2023; Cazenavette et al., 2022; Cui et al., 2023), often through costly bi-level procedures tied to the distillation architecture. Multimodal distillation must additionally preserve image–text correspondence for cross-modal retrieval (Radford et al., 2021; Li et al., 2021). PDS (Choi et al., 2026)

offers a scalable training-free alternative by constructing CLIP prototype pairs and decoding them with unCLIP. DPRG retains this pipeline while coupling pairs during mode discovery and making generation retrieval-aware.

Mode discovery for diffusion-based distillation. Image-only diffusion distillers discover and guide modes in the generative latent space, as in MGD^3 (Chan-Santiago et al., 2025), or align an off-the-shelf text-to-image prior to density-based core samples, as in CoDA (Zhou et al., 2026). For multimodal data, ProCo (Dang et al., 2026) clusters cross-modal retrieval distributions, whereas PDS (Choi et al., 2026) separately clusters image and text features before matching the modes. DPRG instead over-clusters paired representations, rejects candidates without independent-image support, and allocates the final budget across reliable modes through farthest-first coverage.

Diffusion guidance for conditional synthesis. Classifier-free guidance combines conditional and unconditional predictions during sampling (Ho & Salimans, 2022) and underpins text-to-image and unCLIP decoders (Ramesh et al., 2022; Rombach et al., 2022). DPRG augments this mechanism with retrieval-derived hard negatives, encouraging each synthetic pair to remain distinguishable from its nearest image- and text-side alternatives.

3 PRELIMINARIES

Problem setup. Given a large-scale image-text dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$, where x_n denotes an image and y_n denotes its corresponding caption, multimodal dataset distillation (MDD) aims to construct a compact synthetic dataset $\mathcal{S} = \{(\tilde{x}_m, \tilde{y}_m)\}_{m=1}^{M}$, with $M \ll N$. The distilled dataset should preserve the training utility of the original dataset, such that a vision–language model trained on $\mathcal{S}$ achieves downstream performance comparable to that of a model trained on the full dataset $\mathcal{D}$.

Prototype-based diffusion distillation. Optimization-based distillation constructs $\mathcal{S}$ through iterative, network-dependent objectives based on training losses, gradients, or training trajectories (Wang et al., 2018; Zhao & Bilen, 2021; Cazenavette et al., 2022). A training-free, prototype-based line instead uses pretrained diffusion models for synthesis (Su et al., 2024; Chan-Santiago et al., 2025). At a high level, these methods separate prototype discovery in a frozen feature space from prototype-guided image generation. In the fixed-budget formulation considered here, M prototypes are used to synthesize M samples (Choi et al., 2026). With the pretrained networks held fixed, these two stages avoid student-training loops and student-specific distillation objectives. This removes the associated optimization overhead and decouples synthesis from the downstream student architecture (Su et al., 2024; Choi et al., 2026).

(Stage 1) Prototype discovery. A frozen encoder ϕ maps each sample to a feature, the features are partitioned into groups $\{\mathcal{C}_m\}_{m=1}^{M}$, and each group is summarized by its mean,

$$\mathbf{p}_m = \frac{1}{|\mathcal{C}_m|} \sum_{n \in \mathcal{C}_m} \phi(x_n). \tag{1}$$

The number of groups is set to the distillation budget M, so one prototype is discovered per available slot. In the unimodal setting this partition is not the method's to choose: class labels split $\mathcal{D}$ first, and the images-per-class protocol fixes how many slots each class receives, leaving clustering to operate only within a class (Gu et al., 2024; Su et al., 2024; Chan-Santiago et al., 2025).

(Stage 2) Conditional generation. Each prototype supplies a condition c_m to a pretrained diffusion generator G, which decodes one synthetic sample $\tilde{x}_m = G(c_m)$. Sampling uses classifier-free guidance (Ho & Salimans, 2022): writing $\boldsymbol{\epsilon}_\theta$ for the denoiser, z_t for the latent at step t, and $\varnothing$ for the null condition,

$$\hat{\boldsymbol{\epsilon}}(z_t, c_m) = \boldsymbol{\epsilon}_\theta(z_t, \varnothing) + s\big(\boldsymbol{\epsilon}_\theta(z_t, c_m) - \boldsymbol{\epsilon}_\theta(z_t, \varnothing)\big), \tag{2}$$

where the guidance scale $s > 1$ amplifies the target condition against the unconditional prediction. Both stages act on one condition at a time: clustering assigns each sample to one group, and guidance steers each sample toward its own c_m.

Multimodal instantiation. PDS adapts prototype discovery to paired image–text data (Choi et al., 2026). For each pair (x_n, y_n), it extracts ℓ_2-normalized visual and textual CLIP features $\mathbf{x}_n = \phi_{\mathsf{v}}(x_n)$ and $\mathbf{y}_n = \phi_{\mathsf{t}}(y_n)$ (Radford et al., 2021), then clusters the two modalities independently into

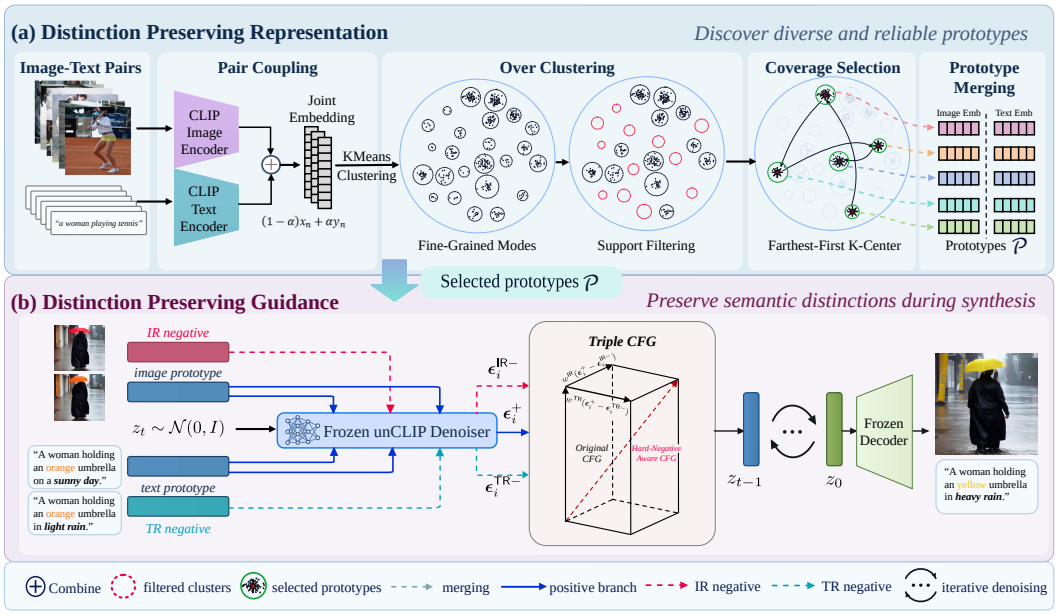


Figure 2: **Overview of** DPRG**. (a)** Distinction-Preserving Representation couples frozen image–text features and over-clusters the joint embeddings to discover fine-grained candidate modes. It filters candidates by support from distinct images and selects M separated modes by farthest-first coverage. Prototype merging forms paired image and text prototypes from each selected mode and retrieves a caption prompt from the full training set. **(b)** Distinction-Preserving Guidance mines the most confusable image-side and text-side alternatives for each prototype. Triple CFG contrasts the positive prediction with these two negative predictions during denoising to generate one sample per prototype. Algorithm 1 in Appendix B gives the complete procedure.

M groups. Because the resulting partitions are not aligned, PDS finds a one-to-one cluster matching that maximizes the number of image–text pairs shared by matched groups and averages the features on each matched intersection to obtain paired visual and textual prototypes.

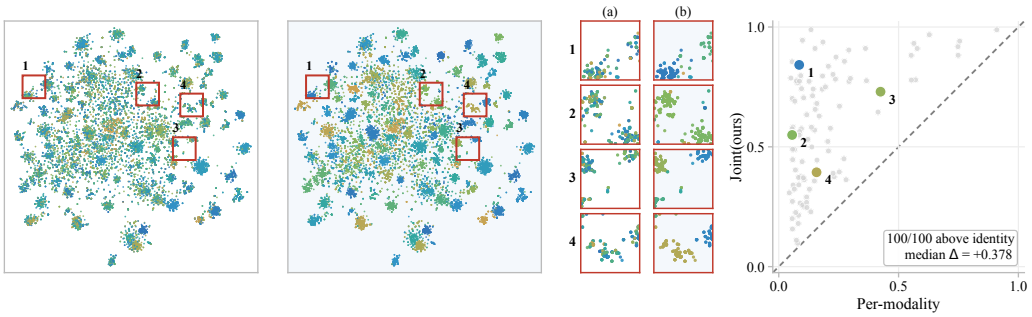
4 Method

4.1 Distinction-Preserving Representation

Pair coupling. Visually similar images can convey different meanings under different captions, and similar captions can refer to different visual evidence. We therefore couple each observed pair before clustering (Figure 2) to preserve distinctions that independent assignments overlook (Figure 3). For each image–text pair (x_n, y_n), let $\mathbf{x}_n$ and $\mathbf{y}_n$ denote its ℓ_2-normalized visual and textual features, and let f_n^{v} and f_n^{t} denote the corresponding original, unnormalized embeddings. We first interpolate the normalized features as $\widetilde{\mathbf{z}}_n = (1-\alpha)\mathbf{x}_n + \alpha\mathbf{y}_n$, where $\alpha \in (0,1)$ controls their relative contributions, and then normalize the pair representation as $\mathbf{z}_n = \widetilde{\mathbf{z}}_n / \|\widetilde{\mathbf{z}}_n\|_2$ before clustering. For two observed pairs n and r, the inner product expands as

$$\widetilde{\mathbf{z}}_n^\top \widetilde{\mathbf{z}}_r = \underbrace{(1-\alpha)^2 \mathbf{x}_n^\top \mathbf{x}_r}_{\text{visual–visual}} + \underbrace{\alpha^2 \mathbf{y}_n^\top \mathbf{y}_r}_{\text{text–text}} + \underbrace{\alpha(1-\alpha)\left(\mathbf{x}_n^\top \mathbf{y}_r + \mathbf{y}_n^\top \mathbf{x}_r\right)}_{\text{bidirectional cross-modal}}. \tag{3}$$

The first two terms measure agreement within each modality, while the cross-modal terms relate each image to the other pair's caption. Dividing by both pair norms gives the similarity between the normalized representations used for clustering, retaining contributions from all three types of evidence. We cluster these joint representations, so the visual and textual evidence from each observed pair shares a single assignment.



(a) Per-modality assignment (b) Joint pair assignment (ours) (c) Local views (d) Purity

Figure 3: **Pair coupling preserves local semantic coherence.** **(a,b)** The same Flickr30K pairs in frozen CLIP image-space t-SNE coordinates, coloured by per-modality or joint assignments to 100 modes. **(c)** Enlarged views of the four marked regions show more coherent assignments under pair coupling. **(d)** Assignment agreement between each pair and its image's 30 nearest neighbours, averaged within each mode. All 100 modes lie above the identity line; highlighted points correspond to the four marked regions.

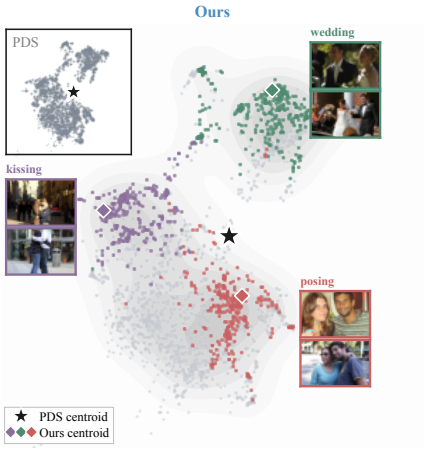


Figure 4: **Budget-level clustering can merge distinct modes.**

Over-clustering. Budget-level clustering can average distinct modes into ambiguous prototypes (Figure 4). To separate semantic discovery from compression, we increase the number of clusters from M to $\bar{M} = \rho M$, where $\rho > 1$, and apply MiniBatch K-Means to the normalized $\mathbf{z}_n$ to obtain $\bar{M}$ candidate clusters. Let $\mathcal{A}_k = \{n : a_n = k\}$ be the member set of candidate k. We remove candidates without sufficient support from distinct training images, producing a reliable set $\mathcal{C}_{\text{rel}}$. The complete filtering rule is given in Section B.1.

Coverage selection. The final set must retain broad semantic coverage under a limited budget. We therefore reduce the reliable candidates to M mutually separated modes using farthest-first selection (Gonzalez, 1985). For unit-normalized reliable centers, we use the cosine dissimilarity $d_{\cos}(k,l) = 1 - \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_l$. With all candidate indices below ranging over $\mathcal{C}_{\text{rel}}$, each selection is given by

$$k^* = \begin{cases} \arg\max_{k} \min_{l \neq k} d_{\cos}(k,l), & \mathcal{Q} = \varnothing, \\ \arg\max_{k \notin \mathcal{Q}} \min_{q \in \mathcal{Q}} d_{\cos}(k,q), & \mathcal{Q} \neq \varnothing. \end{cases} \tag{4}$$

After each selection, we add k^* to $\mathcal{Q}$ and repeat until $|\mathcal{Q}| = M$, resolving all ties by the smallest original candidate index. Each step therefore fills the largest current coverage gap.

Prototype merging. Let q_i be the candidate selected for final mode i and let $\mathcal{C}_i = \mathcal{A}_{q_i}$ contain all of its assigned image–text pairs. We build the i-th image and text prototype by averaging the original embeddings assigned to $\mathcal{C}_i$:

$$\mathbf{p}_i^{\mathsf{v}} = \frac{1}{|\mathcal{C}_i|} \sum_{n \in \mathcal{C}_i} f_n^{\mathsf{v}}, \qquad \mathbf{p}_i^{\mathsf{t}} = \frac{1}{|\mathcal{C}_i|} \sum_{n \in \mathcal{C}_i} f_n^{\mathsf{t}}. \tag{5}$$

The original means are retained as unCLIP conditioning vectors. Whenever cosine similarity is required, we use $\bar{\mathbf{p}}_i^{\mathsf{v}} = \mathbf{p}_i^{\mathsf{v}} / \|\mathbf{p}_i^{\mathsf{v}}\|_2$ and $\bar{\mathbf{p}}_i^{\mathsf{t}} = \mathbf{p}_i^{\mathsf{t}} / \|\mathbf{p}_i^{\mathsf{t}}\|_2$. The caption is selected globally as

$$n_i^* = \arg\max_{n \in [N]} (\bar{\mathbf{p}}_i^{\mathsf{t}})^\top \mathbf{y}_n, \qquad \hat{y}_i = y_{n_i^*}. \tag{6}$$

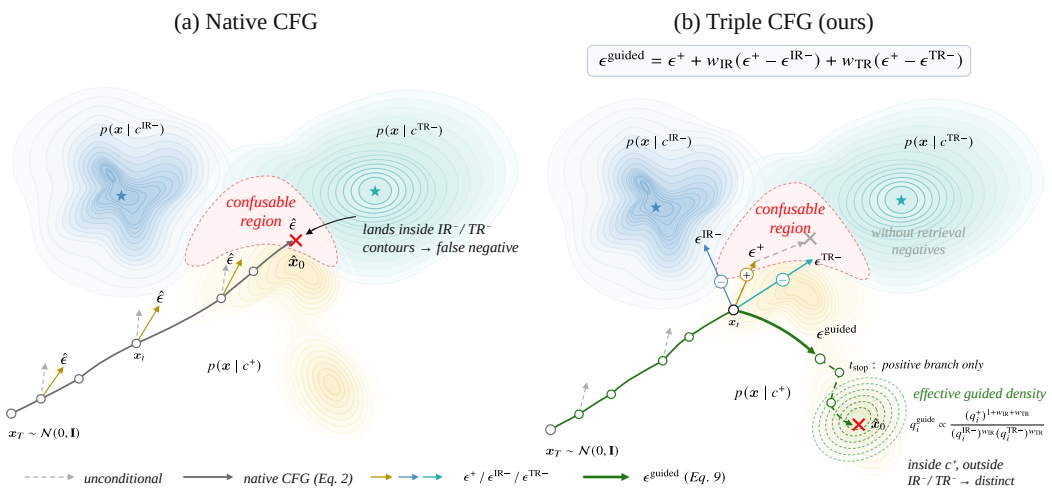


Figure 5: **Triple CFG for distinction-preserving generation. (a)** Native CFG can produce a sample compatible with the positive condition yet confusable with IR or TR alternatives. **(b)** Triple CFG amplifies the positive prediction relative to image-side and text-side negative predictions, steering samples away from confusable alternatives. The dashed green trajectory after t_{stop} uses only the positive branch. Contours and trajectories are schematic. The three-cluster toy model is detailed in Appendix B.2.

4.2 Distinction-Preserving Guidance

Hard-negative mining. In contrastive image–text learning, other pairs act as negatives, so near-identical synthetic pairs can become false negatives for one another. We therefore mine the most confusable alternatives within the distilled set, targeting the distinctions that downstream learning must preserve. For target prototype pair i, the two hard negatives are

$$j_i^{\mathsf{IR}} = \arg\max_{j \neq i} (\bar{\mathbf{p}}_i^{\mathsf{t}})^{\top} \bar{\mathbf{p}}_j^{\mathsf{v}}, \qquad j_i^{\mathsf{TR}} = \arg\max_{j \neq i} (\bar{\mathbf{p}}_j^{\mathsf{t}})^{\top} \bar{\mathbf{p}}_i^{\mathsf{v}}. \tag{7}$$

For prototype i, the positive and two counterfactual negative conditions are

$$c_i^{+} = (\mathbf{p}_i^{\mathsf{v}}, \hat{y}_i), \qquad c_i^{\mathsf{IR}-} = (\mathbf{p}_{j_i^{\mathsf{IR}}}^{\mathsf{v}}, \hat{y}_i), \qquad c_i^{\mathsf{TR}-} = (\mathbf{p}_i^{\mathsf{v}}, \hat{y}_{j_i^{\mathsf{TR}}}). \tag{8}$$

Contrastive synthesis. Standard CFG conditions each sample in isolation, allowing distinct prototypes to yield confusable samples (Figure 5). We therefore contrast the positive prediction with those under the two negative conditions. At each denoising step, the three branches produce $\boldsymbol{\epsilon}_i^{+}$, $\boldsymbol{\epsilon}_i^{\mathsf{IR}-}$, and $\boldsymbol{\epsilon}_i^{\mathsf{TR}-}$, where each prediction already includes the generator's ordinary conditional–unconditional CFG of Equation (2). We combine them as

$$\boldsymbol{\epsilon}_i^{\mathsf{guided}} = \boldsymbol{\epsilon}_i^{+} + w^{\mathsf{IR}}\big(\boldsymbol{\epsilon}_i^{+} - \boldsymbol{\epsilon}_i^{\mathsf{IR}-}\big) + w^{\mathsf{TR}}\big(\boldsymbol{\epsilon}_i^{+} - \boldsymbol{\epsilon}_i^{\mathsf{TR}-}\big), \tag{9}$$

where w^{IR} and w^{TR} control the two retrieval-negative directions. Triple CFG is applied from the beginning of denoising until normalized time t_{stop}; subsequent steps use only the positive branch.

5 Experiments

5.1 Experimental Settings

Datasets and budgets. We conduct experiments on Flickr30K (Plummer et al., 2015), MS-COCO (Lin et al., 2014), Flickr8K (Hodosh et al., 2013), and Pascal1K. Flickr30K and MS-COCO are evaluated with 100 and 300 synthetic image–caption pairs, while Flickr8K and Pascal1K are evaluated with 50, 75, and 100 pairs. We follow the data splits and retrieval annotations of PDS (Choi et al., 2026). Pascal1K uses an 800/100/100 train/validation/test split. Component ablations and hyperparameter analyses are conducted on Flickr30K.

Table 1: Retrieval comparison with TESLA-VL (Cui et al., 2023), LoRS (Xu et al., 2024), and PDS (Choi et al., 2026) on Flickr30K and MS-COCO at 100 and 300 synthetic pairs. Values are reported as mean ± standard deviation. Best and second-best scores are bold and underlined.

Method	Pairs	Image Retrieval			Text Retrieval			R_{mean}
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
Flickr30K / NF-ResNet50								
TESLA-VL (Cui et al., 2023)	100	4.1 ± 0.3	14.7 ± 0.9	22.9 ± 1.2	6.5 ± 0.4	17.8 ± 1.4	27.3 ± 1.4	15.6 ± 0.4
LoRS (Xu et al., 2024)	100	6.3 ± 0.1	18.6 ± 0.1	28.0 ± 0.2	9.1 ± 0.2	24.3 ± 0.4	34.5 ± 0.8	20.1 ± 0.2
PDS (Choi et al., 2026)	100	7.9 ± 0.3	25.8 ± 0.4	37.3 ± 0.3	10.2 ± 0.3	28.2 ± 0.9	39.0 ± 0.3	24.7 ± 0.2
DPRG (Ours)	100	**10.7** ± 0.1	**31.2** ± 0.5	**43.1** ± 0.2	**14.7** ± 0.7	**34.8** ± 0.9	**46.5** ± 1.3	**30.2** ± 0.2
TESLA-VL (Cui et al., 2023)	300	10.3 ± 0.3	28.8 ± 0.6	40.6 ± 0.8	14.9 ± 0.9	36.2 ± 1.1	48.8 ± 1.7	29.9 ± 0.4
LoRS (Xu et al., 2024)	300	8.6 ± 0.9	23.5 ± 1.5	33.5 ± 1.8	14.7 ± 1.2	33.5 ± 2.8	44.1 ± 3.2	26.3 ± 0.8
PDS (Choi et al., 2026)	300	14.4 ± 0.4	38.1 ± 0.2	51.4 ± 0.4	18.7 ± 0.5	45.0 ± 0.4	57.8 ± 0.6	37.6 ± 0.2
DPRG (Ours)	300	**16.6** ± 0.2	**41.5** ± 0.1	**53.5** ± 0.1	**20.8** ± 0.3	**47.4** ± 0.3	**59.6** ± 0.6	**39.9** ± 0.1
Flickr30K / ViT								
TESLA-VL (Cui et al., 2023)	100	2.1 ± 0.3	7.8 ± 0.7	13.1 ± 1.2	2.6 ± 0.6	8.7 ± 0.9	13.7 ± 1.4	8.0 ± 0.4
LoRS (Xu et al., 2024)	100	2.8 ± 0.1	9.9 ± 0.4	16.1 ± 0.2	5.2 ± 0.3	13.6 ± 0.5	20.5 ± 0.2	11.4 ± 0.1
PDS (Choi et al., 2026)	100	6.8 ± 0.3	19.2 ± 0.3	28.5 ± 0.4	6.6 ± 0.5	17.5 ± 0.5	26.9 ± 0.5	17.6 ± 0.2
DPRG (Ours)	100	**7.6** ± 0.1	**21.4** ± 0.1	**31.6** ± 0.1	**7.6** ± 0.5	**20.0** ± 0.8	**28.7** ± 0.2	**19.5** ± 0.2
TESLA-VL (Cui et al., 2023)	300	5.1 ± 0.7	16.2 ± 1.0	24.5 ± 1.1	6.1 ± 0.6	18.0 ± 1.0	27.3 ± 1.4	16.2 ± 0.4
LoRS (Xu et al., 2024)	300	4.1 ± 0.5	13.1 ± 1.0	20.7 ± 1.3	6.2 ± 0.6	17.1 ± 1.2	25.7 ± 1.8	14.5 ± 0.5
PDS (Choi et al., 2026)	300	9.1 ± 0.1	27.3 ± 0.4	**38.4** ± 0.4	9.6 ± 0.3	26.1 ± 0.5	37.5 ± 1.2	24.7 ± 0.2
DPRG (Ours)	300	**10.6** ± 0.2	**27.8** ± 0.3	38.3 ± 0.2	**10.5** ± 0.3	**27.9** ± 0.6	**37.8** ± 0.9	**25.5** ± 0.3
MS-COCO / NF-ResNet50								
TESLA-VL (Cui et al., 2023)	100	1.4 ± 0.1	5.8 ± 0.2	10.2 ± 0.4	2.1 ± 0.3	7.6 ± 0.3	12.5 ± 0.3	6.6 ± 0.1
LoRS (Xu et al., 2024)	100	1.8 ± 0.1	6.8 ± 0.2	11.4 ± 0.4	2.3 ± 0.2	7.8 ± 0.5	12.6 ± 0.4	7.1 ± 0.1
PDS (Choi et al., 2026)	100	2.8 ± 0.1	10.0 ± 0.2	17.3 ± 0.3	4.5 ± 0.2	14.0 ± 0.3	21.4 ± 0.4	11.7 ± 0.1
DPRG (Ours)	100	**3.2** ± 0.1	**11.3** ± 0.1	**18.7** ± 0.1	**5.0** ± 0.3	**14.4** ± 0.1	**21.9** ± 0.3	**12.4** ± 0.0
TESLA-VL (Cui et al., 2023)	300	3.0 ± 0.1	10.7 ± 0.3	17.6 ± 0.4	5.7 ± 0.2	17.0 ± 0.3	25.6 ± 0.4	13.3 ± 0.1
LoRS (Xu et al., 2024)	300	2.5 ± 0.2	8.5 ± 0.4	13.8 ± 0.6	3.2 ± 0.2	10.1 ± 0.4	15.9 ± 0.6	9.0 ± 0.2
PDS (Choi et al., 2026)	300	5.3 ± 0.2	17.2 ± 0.4	27.2 ± 0.6	7.4 ± 0.3	20.7 ± 0.3	30.2 ± 0.4	18.0 ± 0.2
DPRG (Ours)	300	**6.1** ± 0.1	**19.1** ± 0.0	**29.1** ± 0.0	**8.6** ± 0.1	**23.3** ± 0.2	**34.2** ± 0.5	**20.1** ± 0.1
MS-COCO / ViT								
TESLA-VL (Cui et al., 2023)	100	0.5 ± 0.2	2.1 ± 0.8	3.8 ± 1.4	0.2 ± 0.1	1.0 ± 0.5	1.7 ± 0.9	1.6 ± 0.3
LoRS (Xu et al., 2024)	100	0.8 ± 0.1	3.0 ± 0.2	5.4 ± 0.4	0.7 ± 0.2	2.4 ± 0.3	4.1 ± 0.5	2.7 ± 0.1
PDS (Choi et al., 2026)	100	2.3 ± 0.1	8.6 ± 0.1	14.5 ± 0.1	2.2 ± 0.1	7.7 ± 0.2	13.2 ± 0.3	8.1 ± 0.1
DPRG (Ours)	100	**2.7** ± 0.0	**9.9** ± 0.1	**16.0** ± 0.1	**2.9** ± 0.1	**9.8** ± 0.0	**15.6** ± 0.2	**9.5** ± 0.1
TESLA-VL (Cui et al., 2023)	300	1.5 ± 0.1	5.9 ± 0.1	10.1 ± 0.3	2.2 ± 0.2	7.8 ± 0.2	12.6 ± 0.4	6.7 ± 0.1
LoRS (Xu et al., 2024)	300	1.0 ± 0.3	3.7 ± 1.1	6.3 ± 1.8	1.3 ± 0.4	4.6 ± 1.4	7.4 ± 2.0	4.1 ± 0.5
PDS (Choi et al., 2026)	300	4.1 ± 0.1	13.4 ± 0.1	21.2 ± 0.1	3.7 ± 0.1	12.6 ± 0.1	19.8 ± 0.2	12.5 ± 0.1
DPRG (Ours)	300	**4.7** ± 0.0	**14.7** ± 0.1	**22.5** ± 0.2	**4.5** ± 0.1	**14.3** ± 0.5	**22.0** ± 0.5	**13.8** ± 0.2

Prototype construction and synthesis. Unless otherwise specified, we use the CLIP ViT-L/14 image and text encoders (Radford et al., 2021) and synthesize 768×768 images with the Stable Diffusion 2.1 unCLIP decoder (Ramesh et al., 2022). Generation uses classifier-free guidance (Ho & Salimans, 2022) with scale 5.0, 100 denoising steps, and noise level 10. DPRG applies MiniBatch K-Means (Sculley, 2010) to joint image–caption representations, filters candidates by distinct-image support, and selects the final budget by farthest-first coverage (Gonzalez, 1985). Unless varied explicitly, $\alpha = 0.5$, $\rho = 6$, $w^{\mathsf{IR}} = 1$, $w^{\mathsf{TR}} = 1.25$, and $t_{\mathsf{stop}} = 0.5$. Appendix A.2 specifies the SigLIP (Zhai et al., 2023)–FLUX.1-dev configuration.

Baselines and fairness. We compare with TESLA-VL (Cui et al., 2023), LoRS (Xu et al., 2024), and PDS (Choi et al., 2026).

Downstream retrieval evaluation. Following LoRS and PDS, we test whether each distilled set remains useful with a frozen NF-ResNet50 (Brock et al., 2021) or ViT (Dosovitskiy, 2020) image encoder. Synthetic and test images are resized to 224×224. The CLIP text encoder is frozen, and only a projection head that maps caption embeddings into the selected visual feature space is trained. We report the fixed epoch-100 checkpoint using image retrieval (IR) and text retrieval (TR) recall at 1, 5, and 10, together with their arithmetic mean R_{mean}. Each directly evaluated distilled

setting averages three generation seeds and three projection-head seeds per generation, for nine runs in total. Appendix A.3 gives the full optimization, preprocessing, metric, and statistical protocol.

5.2 Comparison Experiments

Larger benchmarks: Flickr30K and MS-COCO. Table 1 reports all six recall measures. DPRG improves 47 of the 48 detailed metrics over PDS across both retrieval directions and evaluation backbones. The gain is largest under a tight budget: on Flickr30K at $M = 100$, R_{mean} rises from 24.7 to 30.2 with NF-ResNet50 and from 17.6 to 19.5 with ViT. At this budget, DPRG also exceeds both TESLA-VL and LoRS using 300 pairs on each backbone, showing that a training-free pipeline can attain their retrieval utility with one third of the synthetic budget. Its advantage over PDS persists at $M = 300$ and on MS-COCO.

Small- and medium-scale benchmarks: Pascal1K and Flickr8K. Appendix C reports the complete results. DPRG yields a higher R_{mean} for every budget and backbone on both datasets, despite the greater variance of individual recall measures on the smaller test sets. Across all four benchmarks, DPRG improves 113 of 120 detailed recall comparisons, ties one, and decreases six, while improving R_{mean} in all 20 dataset–budget–backbone settings.

Transfer to SigLIP and FLUX.1-dev. We replace the prototype encoder with SigLIP and the generator with FLUX.1-dev equipped with an IP-Adapter, while matching cached features, checkpoints, resolution, denoising schedule, and downstream evaluation between PDS and DPRG. Appendix C.2 reports the complete results in Tables 7 and 8. DPRG improves R_{mean} in all eight dataset–budget–backbone settings and 47 of the 48 detailed recall measures. Thus, the advantage survives a simultaneous change of representation and generator families. Because this comparison changes prototype construction and generation guidance together, it establishes system-level transfer rather than attributing the gain to an individual component.

5.3 Analysis Experiments

We evaluate whether the components act coherently, whether the prototypes are more distinguishable, and whether the conclusions are robust to the clustering algorithm and hyperparameters.

Component ablation. We enumerate all 2^3 combinations of pair-level representation (Pair), over-clustering (Over), and Triple CFG on Flickr30K at both budgets. Figure 6 shows each component added to PDS and removed from DPRG, and Appendix D.1 reports every configuration. Pair and Triple CFG improve both contrasts, and Triple CFG remains beneficial under every state of the other two. Over behaves differently, costing from 1.8 to 3.6 points of R_{mean} at $M = 300$ without Pair and gaining from 1.8 to 3.5 points with it across both backbones and both Triple CFG states. Clustering one modality at a time yields finer candidates that stay cross-modally inconsistent, so pair assignment must make them coherent before support filtering and coverage selection can use them. The complete pipeline is strongest at every budget and backbone, which makes the three components complementary rather than additive.

Analysis on hyperparameters. Figure 10 in Appendix D.1 varies pair mixing, over-clustering, and the two guidance weights on Flickr30K validation at $M = 100$. Balanced mixing, intermediate over-clustering, and nonzero weight on both retrieval-negative branches work best. The appendix also reports stop-guidance timing.

Qualitative comparison. Figure 7 compares DPRG without Triple CFG (a) and the full method (b) on Flickr30K at $M = 100$. The former repeats similar dog-on-grass scenes, whereas the latter spans more varied actions and contexts.



Figure 7: Triple CFG: (a) off; (b) on.

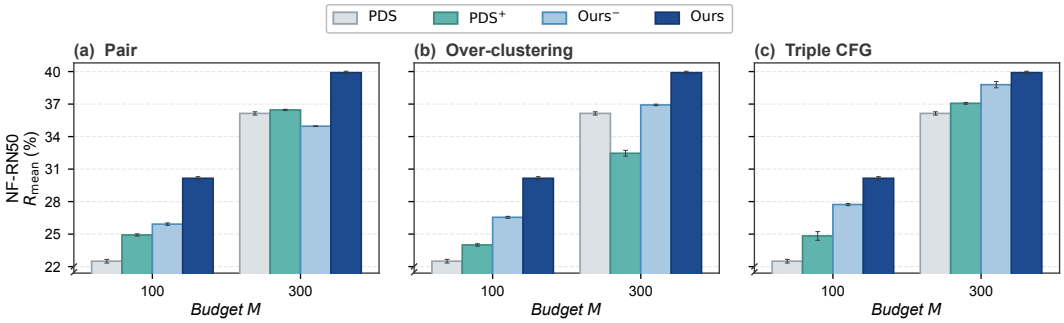


Figure 6: Component ablation on Flickr30K with NF-ResNet50. Panels identify the target component. PDS^{+} adds the target to PDS, whereas Ours^{-} removes it from DPRG. Bars and error bars show the mean and one sample standard deviation under the nine-run protocol. The corresponding ViT results are reported in Appendix D.1.

Retrieval-margin analysis. We use retrieval margins to assess semantic distinctness within generated sets before downstream training. DPRG's average margin is $1.6\times$–$2.5\times$ that of PDS across all tested dataset–budget settings (Table 2), indicating better separation from cross-modal confusions.

We compute the cosine-similarity gap between each generated image–caption pair and its hardest non-matching alternative in both retrieval directions. Let S denote their within-set similarity matrix. For pair i, the margins are

$$m_{\mathsf{I}\to\mathsf{T}}(i) = S_{ii} - \max_{j\neq i} S_{ij}, \qquad m_{\mathsf{T}\to\mathsf{I}}(i) = S_{ii} - \max_{j\neq i} S_{ji}. \tag{10}$$

We report their arithmetic average, $\bar{m}_{\mathsf{avg}} = \frac{1}{2}(\bar{m}_{\mathsf{I}\to\mathsf{T}} + \bar{m}_{\mathsf{T}\to\mathsf{I}})$.

Table 2: Hard-negative margins within the generated sets. Higher is better.

Method	Pairs	I→T↑	T→I↑	Average↑
Flickr30K				
PDS (Choi et al., 2026)	100	0.0541	0.0513	0.0527
DPRG (Ours)	100	**0.0850**	**0.0788**	**0.0819**
PDS (Choi et al., 2026)	300	0.0362	0.0339	0.0350
DPRG (Ours)	300	**0.0621**	**0.0577**	**0.0599**
MS-COCO				
PDS (Choi et al., 2026)	100	0.0457	0.0430	0.0443
DPRG (Ours)	100	**0.0794**	**0.0794**	**0.0794**
PDS (Choi et al., 2026)	300	0.0232	0.0204	0.0218
DPRG (Ours)	300	**0.0543**	**0.0534**	**0.0538**

Coverage visualization. Figure 8 overlays the two prototype sets on the fixed t-SNE map used for the pair-assignment visualization. Keeping M of the ρM candidates gives up a complete partition without giving up reach, since farthest-first selection spreads the retained prototypes across the space. Appendix D.4 describes how the plot is constructed and Section D.2 gives the downstream comparison.

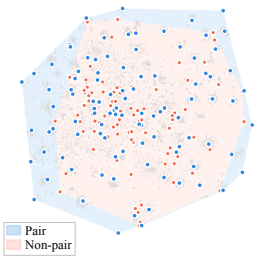


Figure 8: Prototype locations in a fixed Flickr30K t-SNE projection.

Robustness to clustering algorithms. Appendix D.2 replaces MiniBatch K-Means with GMM and Ward clustering while holding the remaining pipeline fixed. DPRG outperforms its matched PDS variant for every solver and both backbones; MiniBatch K-Means gives the strongest DPRG result at the lowest normalized cost. The gain therefore does not arise from substituting a more expensive clustering algorithm.

6 Conclusion and Limitations

We introduced DPRG, combining pair-level semantic units, over-clustering with coverage selection, and hard-negative Triple CFG. Across four datasets, DPRG improves 113 of 120 detailed recall comparisons and raises R_{mean} in all 20 dataset–budget–backbone settings. Transfer and ablation results support the portability and complementary roles of its components.

Limitations. DPRG applies only to encoders with a joint image–text space such as CLIP and SigLIP, since both pair interpolation and hard-negative mining operate there.

AI USE STATEMENT

We used GPT-5.6 to assist with language polishing in the relevant sections. The tool was not used to generate research claims, experimental results, mathematical arguments, or references. All AI-assisted text was reviewed and revised by the authors, and all factual claims, citations, and equations were manually verified. The authors take full responsibility for the final content of this work.

REFERENCES

Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International conference on machine learning*, pp. 1059–1071. PMLR, 2021.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10708–10717. IEEE, 2022.

Jeffrey A Chan-Santiago, Praveen Tirupattur, Gaurav Kumar Nayak, Gaowen Liu, and Mubarak Shah. Mgd ^{3}: Mode-guided dataset distillation using diffusion models. *arXiv preprint arXiv:2505.18963*, 2025.

Junhyeok Choi, Sangwoo Mo, and Minwoo Chae. Multimodal dataset distillation made simple by prototype-guided data synthesis. *arXiv preprint arXiv:2602.19756*, 2026.

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

Zhuohang Dang, Minnan Luo, Chengyou Jia, Hangwei Qian, Xinyu Zhang, Xiaojun Chang, and Ivor Tsang. Correspondence coverage matters for multi-modal dataset distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20693–20701, 2026.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22, 1977.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical computer science*, 38:293–306, 1985.

Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15793–15803. IEEE, 2024.

Shengbin Guo, Hang Zhao, Senqiao Yang, Chenyang Jiang, Yuhang Cheng, Xiangru Peng, Rui Shao, and Zhuotao Tian. Multimodal dataset distillation via phased teacher models. *arXiv preprint arXiv:2603.25388*, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

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. Springer, 2014.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pp. 1177–1178, 2010.

Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. D 4 m: Dataset distillation via disentangled diffusion model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5809–5818. IEEE, 2024.

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

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

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

Yue Xu, Zhilin Lin, Yusong Qiu, Cewu Lu, and Yong-Lu Li. Low-rank similarity mining for multimodal dataset distillation. *arXiv preprint arXiv:2406.03793*, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Ziruo Zhang, JIAWEI DU, Zuozhu Liu, Joey Tianyi Zhou, et al. Beyond modality collapse: Representation blending for multimodal dataset distillation. *Advances in Neural Information Processing Systems*, 38:114809–114829, 2026.

Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pp. 12674–12685. PMLR, 2021.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6503–6512. IEEE, 2023.

Zhenghao Zhao, Haoxuan Wang, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. Efficient multimodal dataset distillation via generative models. *Advances in Neural Information Processing Systems*, 38:67912–67933, 2026.

Letian Zhou, Songhua Liu, and Xinchao Wang. Coda: From text-to-image diffusion models to training-free dataset distillation. In *International Conference on Learning Representations*, volume 2026, pp. 81947–81980, 2026.

A EXPERIMENTAL SETTINGS

This appendix specifies the data, synthesis, and downstream evaluation settings used in the main comparison and transfer experiments.

The TESLA-VL, LoRS, and PDS results in Table 1 are taken from PDS (Choi et al., 2026). The Flickr8K PDS result at $M = 100$ with NF-ResNet50 in Table 5, and the PDS results in Table 10, use the same source. We evaluate PDS settings absent from that source with Table 3. For source rows, R_{mean} is the arithmetic mean of the six recalls. Retrieval tables use the same mean $\pm$ standard-deviation notation.

A.1 MAIN EXPERIMENT CONFIGURATION

Tables 3 and 4 list the shared settings and method-specific choices for each experiment.

Table 3: Main CLIP–unCLIP protocol. Shared settings span both methods; method-specific choices are listed separately.

Configuration	PDS	DPRG
Shared data, representation, synthesis, and evaluation		
Datasets and budgets	Flickr30K/MS-COCO: $M \in \{100, 300\}$; Flickr8K/Pascal1K: $M \in \{50, 75, 100\}$	
Prototype encoder	CLIP ViT-L/14 (768-dimensional image and text features)	
Cluster solver	MiniBatch K-Means; batch size 3998, $n_{\mathsf{init}} = 20$, 100 iterations	
Image generator	Stable Diffusion 2.1 unCLIP-small; 768×768, 100 denoising steps, native guidance 5.0, noise level 10	
Evaluation preprocessing	Bicubic resize to 224×224 with CLIP normalization	
Downstream evaluation	Frozen NF-ResNet50 or ViT image encoder and frozen CLIP text encoder; train only the text projection head; report epoch 100 with IR/TR $R@1$, $R@5$, and $R@10$; three generation seeds and three projection-head seeds per generation	
Method-specific prototype construction and guidance		
Input pair filtering	Lowest-similarity 20% removed	All pairs
Clustering unit	Independent image/text embeddings	Pair-level joint embeddings
Pair mixing and over-clustering	–	$\alpha = 0.5$, $\rho = 6$
Candidate reduction	Hungarian matching	Support filtering + farthest-first
Retrieval guidance	Positive branch only	$w^{\mathsf{IR}} = 1$, $w^{\mathsf{TR}} = 1.25$; $t_{\mathsf{stop}} = 0.5$

A.2 SIGLIP–FLUX TRANSFER CONFIGURATION

Table 4: SigLIP–FLUX transfer protocol with shared and method-specific settings.

Configuration	PDS	DPRG
Shared data, representation, synthesis, and evaluation		
Datasets and budgets	Flickr30K and MS-COCO; $M \in \{100, 300\}$	
Prototype encoder	SigLIP SO400M/14@384 (1152-dimensional image and text features)	
Cluster solver	MiniBatch K-Means; batch size 10000, $n_{\mathsf{init}} = 20$, 100 iterations	
Generator and conditioning	FLUX.1-dev in bfloat16; 1024×1024, 24 denoising steps, native guidance 5.5; InstantX FLUX.1-dev IP-Adapter with 128 image tokens and scale 0.6; maximum prompt length 512; InstantX scheduler	
Downstream evaluation	Bicubic resize to 224×224 with CLIP normalization; same frozen NF-ResNet50/ViT evaluator and epoch-100 projection-head protocol; three generation seeds and three projection-head seeds per generation	
Method-specific prototype construction and guidance		
Input pair filtering	Lowest-similarity 20% removed	All pairs
Clustering unit	Independent image/text embeddings	Pair-level joint embeddings
Pair mixing and over-clustering	–	$\alpha = 0.5$, $\rho = 6$
Candidate reduction	Hungarian matching	Support filtering + farthest-first
Retrieval guidance	Positive branch only	$w^{\mathsf{IR}} = 1$, $w^{\mathsf{TR}} = 1.25$; $t_{\mathsf{stop}} = 0.5$

A.3 DOWNSTREAM RETRIEVAL EVALUATION

Flickr30K, Flickr8K, and MS-COCO use the annotations and train/test partitions provided by PDS (Choi et al., 2026); Pascal1K uses an 800/100/100 train/validation/test split. PDS and DPRG use the same split, annotations, synthetic budget, and downstream evaluator for each dataset–budget setting.

The evaluator uses bicubic resizing to 224×224 and CLIP normalization for synthetic training images and real test images. Following LoRS (Xu et al., 2024) and PDS (Choi et al., 2026), we pair a frozen NF-ResNet50 (Brock et al., 2021) or ViT (Dosovitskiy, 2020) image backbone with the frozen CLIP text encoder and optimize only the projection head from caption embeddings to the selected image feature space.

The projection head consists of a linear projection, GELU, a second linear layer, dropout, a residual connection, and LayerNorm. We minimize symmetric image-to-text and text-to-image cross-entropy at temperature 0.07 with SGD (learning rate 0.2, momentum 0.9, weight decay 5×10^{-4}, batch size 128). IR ranks images for each caption, whereas TR ranks captions for each image. We report $R@1$, $R@5$, $R@10$, and their arithmetic mean at the fixed epoch-100 checkpoint, without checkpoint selection.

A.4 Statistical Scope

All distilled-set results produced by our evaluation pipeline use three independent generation seeds. Every generated set is then evaluated with three projection-head initialization seeds, giving nine generation–evaluation combinations. We compute the reported mean and sample standard deviation over these nine runs. Source-reported rows retain their published per-metric dispersions. Full-dataset evaluation has no generation stage and therefore averages three projection-head seeds.

B Algorithm Details

Complete procedure. Algorithm 1 summarizes the complete DPRG procedure, from paired prototype construction to contrastive synthesis.

Algorithm 1 Distinction-Preserving Representation and Guidance (DPRG)

Given: paired data $\mathcal{D}$, budget M, frozen encoders $(\phi_{\mathsf{v}}, \phi_{\mathsf{t}})$, decoder G; α, ρ ($\bar{M} = \rho M$), w^{IR}, w^{TR}, t_{stop}.
Return: distilled set $\mathcal{S} = \{(\tilde{x}_i, \hat{y}_i)\}_{i=1}^{M}$.

Stage 1 Distinction-Preserving Representation

Pair coupling

$f_n^{\mathsf{v}}, f_n^{\mathsf{t}} \leftarrow \phi_{\mathsf{v}}(x_n), \phi_{\mathsf{t}}(y_n)$; normalize to $\mathbf{x}_n, \mathbf{y}_n$.
$\mathbf{z}_n \leftarrow \mathrm{Norm}\big((1-\alpha)\mathbf{x}_n + \alpha \mathbf{y}_n\big)$.

Over-clustering

$\{\mathcal{A}_k, \boldsymbol{\mu}_k\}_{k=1}^{\bar{M}} \leftarrow$ MiniBatchKMeans$(\{\mathbf{z}_n\}, \bar{M})$.
$s_k \leftarrow \#\{\text{distinct images in } \mathcal{A}_k\}$; $\tau \leftarrow \lceil G/\bar{M} \rceil$.
$\mathcal{C}_{\text{rel}} \leftarrow \{k : s_k \geq \tau\}$.

Coverage selection

$\{q_i\}_{i=1}^{M} \leftarrow$ FarthestFirst$(\{\boldsymbol{\mu}_k\}_{k \in \mathcal{C}_{\text{rel}}}, M)$.

Prototype merging

for $i = 1, \ldots, M$ **do**
  $\mathbf{p}_i^{\mathsf{v}}, \mathbf{p}_i^{\mathsf{t}} \leftarrow \mathrm{Mean}_{n \in \mathcal{A}_{q_i}}\big(f_n^{\mathsf{v}}, f_n^{\mathsf{t}}\big)$.
  $\hat{y}_i \leftarrow \arg\max_{y_n \in \mathcal{D}} (\bar{\mathbf{p}}_i^{\mathsf{t}})^\top \mathbf{y}_n$.
end for

Stage 2 Distinction-Preserving Guidance

Hard-negative mining

for $i = 1, \ldots, M$ **do**
  $j_i^{\mathsf{IR}} \leftarrow \arg\max_{j \neq i} (\bar{\mathbf{p}}_i^{\mathsf{t}})^\top \bar{\mathbf{p}}_j^{\mathsf{v}}$.
  $j_i^{\mathsf{TR}} \leftarrow \arg\max_{j \neq i} (\bar{\mathbf{p}}_j^{\mathsf{t}})^\top \bar{\mathbf{p}}_i^{\mathsf{v}}$.
  $c_i^{+} \leftarrow (\mathbf{p}_i^{\mathsf{v}}, \hat{y}_i)$; $c_i^{\mathsf{IR}-} \leftarrow (\mathbf{p}_{j_i^{\mathsf{IR}}}^{\mathsf{v}}, \hat{y}_i)$.
  $c_i^{\mathsf{TR}-} \leftarrow (\mathbf{p}_i^{\mathsf{v}}, \hat{y}_{j_i^{\mathsf{TR}}})$; $u_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Contrastive synthesis

  for $t = T, \ldots, 1$ **do**
    if $t \geq t_{\text{stop}} T$ **then**
      $\boldsymbol{\epsilon} \leftarrow \boldsymbol{\epsilon}_i^{+} + w^{\mathsf{IR}}(\boldsymbol{\epsilon}_i^{+} - \boldsymbol{\epsilon}_i^{\mathsf{IR}-}) + w^{\mathsf{TR}}(\boldsymbol{\epsilon}_i^{+} - \boldsymbol{\epsilon}_i^{\mathsf{TR}-})$
    else $\boldsymbol{\epsilon} \leftarrow \boldsymbol{\epsilon}_i^{+}$
    $u_{t-1} \leftarrow$ SchedulerStep$(u_t, \boldsymbol{\epsilon})$.
  end for; $\tilde{x}_i \leftarrow G(u_0)$.
end for

B.1 Candidate Discovery and Support Filtering

Candidate discovery Stage 1 applies MiniBatch K-Means (Sculley, 2010) to normalized pair representations $\mathbf{z}_n$ with $\bar{M} = \rho M$ clusters. The assignments define the candidate modes used by support filtering and farthest-first selection. For candidate k with member set $\mathcal{A}_k = \{n : a_n = k\}$, its support counts distinct source images,

$$s_k = \left|\{g \in [G] : \exists n \in \mathcal{A}_k \text{ such that } \gamma(n) = g\}\right|. \tag{11}$$

Support filtering Repeated captions from one image contribute at most one unit. At candidate resolution $\bar{M}$, we retain

$$\mathcal{C}_{\mathrm{rel}} = \{k : s_k \geq \tau\}, \qquad \tau = \left\lceil \frac{G}{\bar{M}} \right\rceil. \tag{12}$$

We use support to reject candidates. Every candidate that passes the threshold enters farthest-first selection with equal eligibility; a configuration with fewer than M retained candidates is infeasible.

B.2 Three-Cluster Toy Model

We examine the mechanism illustrated schematically in Figure 5 using an analytic conditional diffusion model with exact Gaussian scores and computed sampling trajectories.

Conditional distributions. The condition $c \in \mathbb{R}^2$ has image-side and text-side coordinates. We define

$$p_0(x \mid c) = \mathcal{N}(x; Ac, I_2), \qquad A = \begin{pmatrix} 1.6 & -1.6 \\ -1.2 & -1.2 \end{pmatrix}, \tag{13}$$

and select $c^+ = (0,0)^\top$, $c^{\mathsf{IR}-} = (-1,0)^\top$, and $c^{\mathsf{TR}-} = (0,-1)^\top$ as the positive and two negative conditions. The unconditional reference is $p_0(x) = \mathcal{N}(x; 0, 4I_2)$, consistent with the condition prior $c \sim \mathcal{N}(0, 3A^{-1}A^{-\top})$.

Guided dynamics. We use a variance-preserving diffusion with $\beta(t) = 0.1 + 29.9t$ for $t \in [0,1]$ and attenuation $a(t) = \exp[-(0.1t + 14.95t^2)/2]$. Its exact conditional and unconditional scores are

$$s_c(t,x) = -(x - a(t)Ac), \qquad s_u(t,x) = -\frac{x}{1 + 3a(t)^2}. \tag{14}$$

Native CFG uses $g_c = s_u + 5(s_c - s_u)$. Triple CFG combines these scores according to Equation (9), with $w^{\mathsf{IR}} = w^{\mathsf{TR}} = 0.5$. We compare native CFG, IR-only, TR-only, and joint guidance by integrating the probability-flow ODE

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\frac{\beta(t)}{2}\big(x + g(t,x)\big) \tag{15}$$

backward from $t = 1$ to $t = 0$ using 1,000 fourth-order Runge–Kutta steps. Guidance remains active throughout this toy experiment, and the process starts from $\mathcal{N}(0, I_2)$; $a(1) = 5.39 \times 10^{-4}$ to three significant figures.

Paired sampling and measurements. All variants use the same 1,000 standard-normal initial draws for each of ten NumPy random seeds, $0, \ldots, 9$, yielding 10,000 endpoints per variant. The example trajectory uses the first draw from seed 5. For condition c, define its 95% highest-density region as $\mathcal{R}_c = \{x : \|x - Ac\|_2^2 \leq -2\log(0.05)\}$. We count an endpoint as confusable when it lies in $\mathcal{R}_{\mathsf{IR}-} \cap \mathcal{R}_{\mathsf{TR}-}$, and as distinct positive when it lies in $\mathcal{R}_+ \setminus (\mathcal{R}_{\mathsf{IR}-} \cup \mathcal{R}_{\mathsf{TR}-})$.

Results and checks. Every sampled native-CFG endpoint falls inside both negative regions, while joint guidance places every endpoint inside the positive region and outside both negatives. Each single-negative variant leaves every endpoint inside the other negative region. A local sensitivity check varies the CFG scale over $\{4, 5, 6\}$ and each negative weight over $\{0.4, 0.5, 0.6\}$. It uses the same 1,000 seed-0 draws and an affine flow map computed with 500 Runge–Kutta steps. The distinct-positive fraction is at least 98.7% across all 27 settings. Independent integrating-factor verification reproduces the saved endpoints within 2×10^{-10} maximum absolute error.

Interpretation. Because the conditional score is affine in c, input-condition extrapolation and explicit three-branch score combination coincide in this model; their paired endpoints agree to numerical precision. Joint negative guidance separates the positive mode under exact scores and integrated probability-flow dynamics.

C Additional Datasets and Model Transfer

We report Flickr8K and Pascal1K results, followed by a transfer experiment with a different representation model and image generator.

C.1 FLICKR8K AND PASCAL1K

Tables 5 and 6 complete the four-dataset comparison in Section 5.2. DPRG improves R_{mean} at every budget and for both backbones on Flickr8K and Pascal1K. Pascal1K uses a 100-image test split, and its recall measures show the corresponding split-level variation.

Table 5: Flickr8K retrieval with PDS (Choi et al., 2026) at $M \in \{50, 75, 100\}$. Values are reported as mean ± standard deviation. Best and second-best values are bold and underlined.

Method	Pairs	Image Retrieval			Text Retrieval			R_{mean}
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
NF-ResNet50								
PDS	50	4.4 ± 0.1	16.6 ± 0.0	26.5 ± 0.2	5.8 ± 0.4	17.0 ± 0.2	25.3 ± 1.0	15.9 ± 0.3
DPRG (Ours)	50	**5.4** ± 0.2	**18.2** ± 0.1	**27.0** ± 0.1	**8.2** ± 0.4	**20.3** ± 0.6	**29.3** ± 0.6	**18.1** ± 0.3
PDS	75	6.0 ± 0.2	19.7 ± 0.3	30.2 ± 0.2	7.5 ± 0.6	20.6 ± 0.7	27.7 ± 0.5	18.6 ± 0.3
DPRG (Ours)	75	**8.3** ± 0.0	**25.6** ± 0.2	**36.6** ± 0.3	**11.2** ± 0.5	**27.4** ± 0.3	**38.7** ± 0.5	**24.6** ± 0.2
PDS	100	8.5 ± 0.2	27.0 ± 0.3	39.8 ± 0.3	10.3 ± 0.2	27.1 ± 0.7	38.3 ± 0.4	25.2 ± 0.2
DPRG (Ours)	100	**11.4** ± 0.1	**30.0** ± 0.4	**41.9** ± 0.2	**14.4** ± 0.4	**33.2** ± 0.4	**43.6** ± 1.2	**29.1** ± 0.1
ViT								
PDS	50	4.0 ± 0.1	14.1 ± 0.2	21.9 ± 0.4	3.6 ± 0.4	11.7 ± 0.6	19.0 ± 0.6	12.4 ± 0.3
DPRG (Ours)	50	**4.1** ± 0.1	**14.7** ± 0.2	**23.0** ± 0.1	**5.6** ± 0.5	**16.6** ± 0.5	**24.6** ± 0.1	**14.8** ± 0.2
PDS	75	4.2 ± 0.1	15.1 ± 0.1	24.2 ± 0.2	3.6 ± 0.2	13.5 ± 0.7	20.2 ± 0.8	13.5 ± 0.2
DPRG (Ours)	75	**5.9** ± 0.3	**19.7** ± 0.1	**29.5** ± 0.1	**7.9** ± 0.6	**21.4** ± 0.6	**30.2** ± 0.5	**19.1** ± 0.1
PDS	100	6.5 ± 0.1	19.8 ± 0.1	29.9 ± 0.2	6.0 ± 0.2	18.1 ± 0.7	27.0 ± 0.7	17.9 ± 0.3
DPRG (Ours)	100	**7.2** ± 0.2	**22.8** ± 0.2	**32.9** ± 0.2	**9.3** ± 0.3	**24.4** ± 0.4	**33.2** ± 0.8	**21.6** ± 0.3

Table 6: Pascal1K retrieval with PDS (Choi et al., 2026) at $M \in \{50, 75, 100\}$. Values are reported as mean ± standard deviation. Best and second-best values are bold and underlined.

Method	Pairs	Image Retrieval			Text Retrieval			R_{mean}
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
NF-ResNet50								
PDS	50	27.1 ± 0.9	69.1 ± 0.6	83.4 ± 0.5	32.7 ± 4.0	62.7 ± 3.5	79.0 ± 2.0	59.0 ± 1.0
DPRG (Ours)	50	**28.1** ± 0.5	**76.2** ± 0.7	**90.0** ± 0.5	**37.0** ± 2.0	**74.0** ± 1.0	**87.0** ± 3.5	**65.4** ± 1.0
PDS	75	33.1 ± 0.3	78.4 ± 0.4	88.3 ± 0.6	39.3 ± 1.2	**70.3** ± 1.5	**84.0** ± 2.0	65.6 ± 0.7
DPRG (Ours)	75	**34.7** ± 0.8	**79.3** ± 0.8	**89.5** ± 0.7	**43.7** ± 1.5	**70.3** ± 2.9	83.0 ± 1.7	**66.7** ± 0.3
PDS	100	31.3 ± 0.6	73.7 ± 0.5	83.1 ± 0.8	**39.7** ± 0.6	70.0 ± 2.0	83.7 ± 2.5	63.6 ± 0.3
DPRG (Ours)	100	**37.6** ± 0.2	**77.3** ± 1.0	**91.3** ± 0.5	39.0 ± 4.0	**71.3** ± 2.3	**87.7** ± 0.6	**67.4** ± 0.5
ViT								
PDS	50	**28.4** ± 0.9	63.9 ± 0.8	78.1 ± 0.9	**29.3** ± 1.2	54.0 ± 1.0	67.0 ± 0.0	53.5 ± 0.4
DPRG (Ours)	50	25.1 ± 0.4	**65.4** ± 1.8	**80.4** ± 0.5	28.3 ± 1.5	**59.7** ± 1.5	**71.3** ± 2.1	**55.0** ± 0.8
PDS	75	26.0 ± 0.2	66.5 ± 0.4	81.7 ± 0.5	30.7 ± 1.5	61.3 ± 0.6	**79.3** ± 1.5	57.6 ± 0.5
DPRG (Ours)	75	**34.0** ± 0.5	**72.1** ± 0.6	**85.3** ± 0.5	**34.0** ± 2.7	**61.7** ± 0.6	78.3 ± 1.5	**60.9** ± 0.1
PDS	100	28.7 ± 1.0	64.3 ± 0.3	79.1 ± 0.6	29.7 ± 1.2	58.0 ± 1.0	74.3 ± 0.6	55.7 ± 0.3
DPRG (Ours)	100	**32.3** ± 0.6	**72.6** ± 0.7	**85.5** ± 0.6	**34.0** ± 1.7	**65.3** ± 1.5	**82.3** ± 1.5	**62.0** ± 0.5

C.2 SIGLIP–FLUX TRANSFER

Tables 7 and 8 report the transfer results using the settings in Table 4.

Table 7: SigLIP–FLUX transfer on Flickr30K; best and second-best values are bold and underlined.

Method	Pairs	Image Retrieval			Text Retrieval			R_{mean}
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
NF-ResNet50								
PDS	100	7.5 ± 0.2	23.5 ± 0.3	34.3 ± 0.6	9.1 ± 0.7	24.8 ± 2.5	36.3 ± 2.2	22.6 ± 1.0
DPRG (Ours)	100	**8.5** ± 0.3	**25.6** ± 0.6	**37.1** ± 0.7	**11.3** ± 0.5	**27.6** ± 1.1	**38.5** ± 1.2	**24.8** ± 0.3
PDS	300	14.6 ± 0.3	37.1 ± 0.7	49.2 ± 1.2	17.9 ± 1.1	40.5 ± 1.8	52.9 ± 2.0	35.4 ± 1.1
DPRG (Ours)	300	**15.8** ± 0.1	**39.6** ± 0.3	**52.2** ± 0.1	**21.5** ± 0.6	**45.7** ± 1.3	**56.8** ± 1.5	**38.6** ± 0.5
ViT								
PDS	100	**5.7** ± 0.3	17.7 ± 1.0	26.6 ± 1.1	5.1 ± 0.7	14.5 ± 1.5	22.6 ± 0.8	15.4 ± 0.8
DPRG (Ours)	100	5.3 ± 0.4	**18.1** ± 0.6	**27.5** ± 0.6	**6.2** ± 1.1	**17.6** ± 1.3	**25.5** ± 1.4	**16.7** ± 0.9
PDS	300	9.9 ± 0.4	26.7 ± 0.4	37.3 ± 0.3	9.8 ± 1.1	26.2 ± 1.3	35.3 ± 0.6	24.2 ± 0.7
DPRG (Ours)	300	**10.0** ± 0.1	**28.5** ± 0.9	**39.5** ± 0.8	**10.6** ± 0.4	**28.7** ± 0.9	**39.0** ± 1.3	**26.1** ± 0.6

Table 8: SigLIP–FLUX transfer on MS-COCO; best and second-best values are bold and underlined.

Method	Pairs	Image Retrieval			Text Retrieval			R_{mean}
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
NF-ResNet50								
PDS	100	2.4 ± 0.1	9.0 ± 0.0	15.7 ± 0.2	3.3 ± 0.2	10.7 ± 0.1	17.0 ± 0.4	9.7 ± 0.1
DPRG (Ours)	100	**3.0** ± 0.1	**11.0** ± 0.2	**18.0** ± 0.2	**4.0** ± 0.2	**13.1** ± 0.1	**20.3** ± 0.3	**11.6** ± 0.2
PDS	300	5.1 ± 0.3	16.1 ± 0.5	25.3 ± 0.6	6.4 ± 0.3	19.5 ± 0.7	29.1 ± 0.8	16.9 ± 0.4
DPRG (Ours)	300	**5.9** ± 0.1	**18.4** ± 0.2	**28.1** ± 0.3	**8.1** ± 0.2	**22.4** ± 0.7	**32.5** ± 0.2	**19.2** ± 0.2
ViT								
PDS	100	2.0 ± 0.2	7.7 ± 0.4	12.9 ± 0.5	1.7 ± 0.2	6.4 ± 0.5	10.8 ± 0.8	6.9 ± 0.4
DPRG (Ours)	100	**2.4** ± 0.0	**8.7** ± 0.1	**14.4** ± 0.2	**2.1** ± 0.2	**7.5** ± 0.1	**12.5** ± 0.4	**7.9** ± 0.1
PDS	300	3.7 ± 0.0	12.7 ± 0.1	19.8 ± 0.3	3.4 ± 0.1	11.2 ± 0.8	18.0 ± 0.7	11.5 ± 0.3
DPRG (Ours)	300	**4.3** ± 0.0	**13.8** ± 0.4	**21.3** ± 0.6	**4.4** ± 0.3	**13.7** ± 0.7	**20.9** ± 0.6	**13.1** ± 0.4

D Ablation and Diagnostic Studies

We study component and hyperparameter ablations, clustering and selection controls, input filtering, and visualization details.

D.1 Components and Hyperparameters

We first vary the core components, clustering solver, and raw-pair selection rule. At $M = 100$ with NF-ResNet50, the all-disabled component row uses unfiltered pairs and gives $R_{\text{mean}} = 22.5 \pm 0.2$. The published PDS row uses its filtering step and gives 24.7. The component table therefore measures Pair, Over, and Triple CFG relative to the unfiltered control pipeline.

Component ablation Figure 9 provides the ViT counterpart to Figure 6. Table 9 reports the exact values for all 2^3 combinations of Pair, Over, and Triple CFG.

Table 9: Full component ablation on Flickr30K. A checkmark denotes an enabled component.

Pair	**Over**	**Triple CFG**	$M = 100$, **NF**	$M = 100$, **ViT**	$M = 300$, **NF**	$M = 300$, **ViT**
×	×	×	22.5 ± 0.2	16.4 ± 0.2	36.1 ± 0.2	21.7 ± 0.2
×	×	✓	24.8 ± 0.4	18.0 ± 0.1	37.1 ± 0.1	24.2 ± 0.0
×	✓	×	24.0 ± 0.1	16.2 ± 0.1	32.5 ± 0.3	19.9 ± 0.3
×	✓	✓	25.9 ± 0.1	18.4 ± 0.3	35.0 ± 0.0	21.1 ± 0.2
✓	×	×	24.9 ± 0.1	16.8 ± 0.1	36.5 ± 0.1	22.0 ± 0.2
✓	×	✓	26.6 ± 0.1	18.2 ± 0.2	36.9 ± 0.1	22.0 ± 0.0
✓	✓	×	27.7 ± 0.1	18.0 ± 0.1	38.8 ± 0.3	23.8 ± 0.0
✓	✓	✓	$\mathbf{30.2 \pm 0.2}$	$\mathbf{19.5 \pm 0.2}$	$\mathbf{39.9 \pm 0.1}$	$\mathbf{25.5 \pm 0.3}$

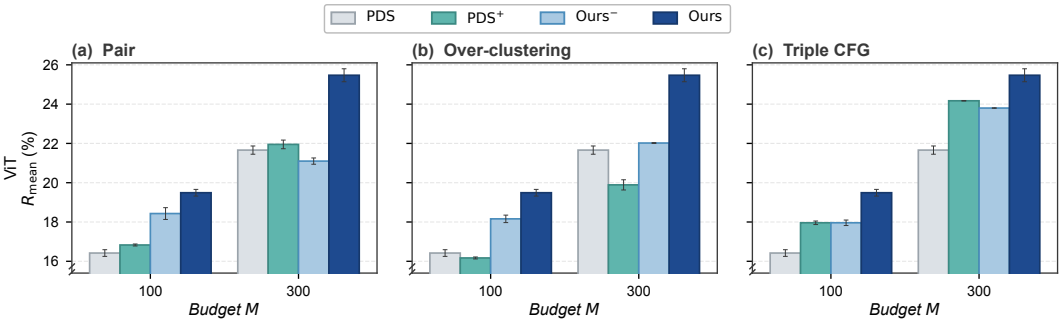
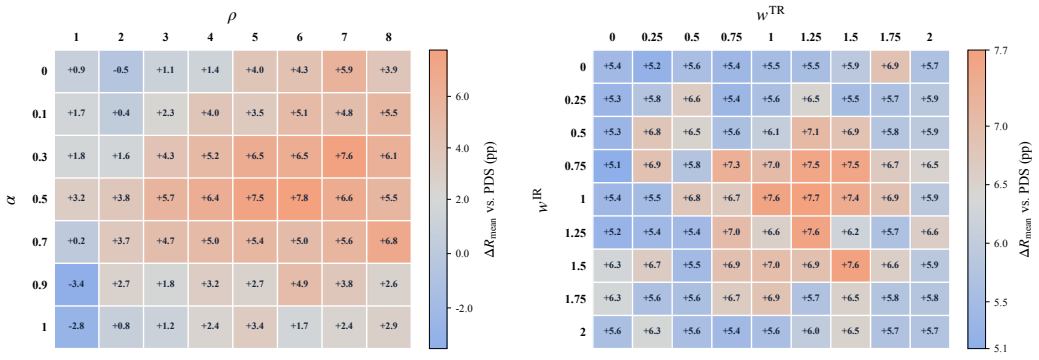


Figure 9: Component ablation on Flickr30K with ViT. PDS$^+$ adds the target to PDS, whereas Ours$^-$ removes it from DPRG. Bars and error bars show the mean and one sample standard deviation.

Analysis on hyperparameters. Both hyperparameter analyses in Figure 10 use the Flickr30K validation set and the PDS control with $R^{\text{PDS}}_{\text{mean}} = 22.5$. The default cell is 30.3 on validation; Table 9 reports 30.2 on test.



(a) Pair mixing α and over-clustering ratio ρ.

(b) IR and TR guidance weights.

Figure 10: Hyperparameter heatmaps on the Flickr30K validation set at $M = 100$ with NF-ResNet50. Each cell reports $\Delta R_{\text{mean}} = R_{\text{mean}} - R^{\text{PDS}}_{\text{mean}}$ relative to the PDS control ($R^{\text{PDS}}_{\text{mean}} = 22.5$), in percentage points to one decimal place.

(i) Pair mixing and over-clustering. Figure 10(a) varies α and ρ on the Flickr30K validation set at $M = 100$. The highest values occur near balanced pair mixing and intermediate over-clustering; weak mixing and extreme candidate counts are less effective.

(ii) IR/TR guidance weights. Figure 10(b) varies w^{IR} and w^{TR} with $\alpha = 0.5$, $\rho = 6$, and $t_{\mathsf{stop}} = 0.5$. The strongest values assign nonzero weight to both branches; removing either branch or heavily favoring one direction lowers retrieval.

(iii) Stop-guidance timing. Figure 11 varies when Triple CFG is disabled. The highest mean score occurs at $t_{\mathsf{stop}} = 0.5$; shorter or longer intervals are less effective.

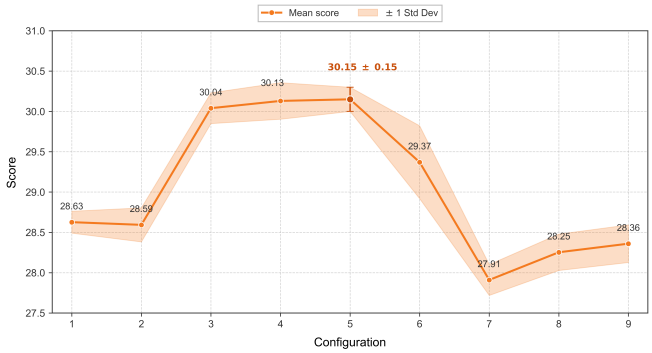


Figure 11: Stop-guidance timing on Flickr30K at $M = 100$ with NF-ResNet50. The fifth configuration uses $t_{\mathsf{stop}} = 0.5$. Curve points and shading show the mean and one sample standard deviation.

D.2 Clustering and Selection Controls

Robustness to clustering algorithms. Table 10 replaces MiniBatch K-Means with GMM (Dempster et al., 1977) or Ward clustering (Ward Jr, 1963) while keeping the subsequent selection, generation, and evaluation stages.

Table 10: Clustering algorithms on Flickr30K at $M = 100$. Values are R_{mean}; relative compute is normalized to MiniBatch K-Means.

Clustering	NF-ResNet50		ViT		Relative compute
	PDS	DPRG	PDS	DPRG	
MiniBatch K-Means	24.7	**30.2**	17.6	**19.5**	1.0×
GMM	25.7	29.9	17.5	17.7	163×
Ward	27.2	28.3	18.0	18.2	488×

Direct pair selection We compare DPRG with three non-learned selectors: random sampling, K-center (Gonzalez, 1985), and herding (Welling, 2009). The selected pairs serve as diffusion conditions, and downstream training uses the resulting generated pairs. The candidate pool contains all 145,000 Flickr30K training pairs. From frozen CLIP ViT-L/14 image and text features $\mathbf{v}_i$ and $\mathbf{t}_i$, we construct the joint selection feature $\mathbf{z}_i = \mathrm{norm}(0.5\,\mathrm{norm}(\mathbf{v}_i) + 0.5\,\mathrm{norm}(\mathbf{t}_i))$. Random draws 100 pairs uniformly without replacement (seed 0). K-center starts from a seed-0 random pair and applies farthest-first selection under cosine distance. Herding greedily minimizes the squared Euclidean distance between the selected-set mean and the corpus mean. Both greedy methods select without replacement and resolve ties by the smallest pair index.

Each selected pair contributes its original, unnormalized image and text features; no cluster average is formed. Its text feature is mapped to the nearest natural-language caption in the full training corpus. We then use the CLIP–unCLIP synthesis, Triple-CFG configuration, per-image noise seeds, and downstream evaluator specified in Table 3 and Section A.3.

Table 11: Direct pair selection and DPRG prototype construction on Flickr30K at $M = 100$.

Method	Image Retrieval			Text Retrieval			R_{mean}
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
NF-ResNet50							
Random	5.5 ± 0.2	16.9 ± 0.3	25.5 ± 0.2	6.3 ± 0.4	18.1 ± 0.6	26.2 ± 0.5	16.4 ± 0.2
K-center	4.7 ± 0.1	15.3 ± 0.2	23.1 ± 0.3	6.7 ± 0.7	19.3 ± 0.6	27.5 ± 0.2	16.1 ± 0.3
Herding	4.9 ± 0.2	15.2 ± 0.2	23.4 ± 0.3	4.9 ± 0.5	14.5 ± 0.1	22.1 ± 0.9	14.2 ± 0.2
DPRG (Ours)	**10.7** ± 0.1	**31.2** ± 0.5	**43.1** ± 0.2	**14.7** ± 0.7	**34.8** ± 0.9	**46.5** ± 1.3	**30.2** ± 0.2
ViT							
Random	2.5 ± 0.0	9.1 ± 0.2	14.7 ± 0.3	2.3 ± 0.2	7.9 ± 0.5	13.0 ± 0.6	8.3 ± 0.3
K-center	2.5 ± 0.1	8.5 ± 0.2	14.0 ± 0.1	3.4 ± 0.1	9.2 ± 0.5	14.1 ± 0.9	8.6 ± 0.2
Herding	1.9 ± 0.1	7.4 ± 0.0	12.1 ± 0.1	1.8 ± 0.1	6.7 ± 0.3	11.2 ± 0.3	6.9 ± 0.1
DPRG (Ours)	**7.6** ± 0.1	**21.4** ± 0.1	**31.6** ± 0.1	**7.6** ± 0.5	**20.0** ± 0.8	**28.7** ± 0.2	**19.5** ± 0.2

The best direct selector is Random with NF-ResNet50 ($R_{\text{mean}} = 16.4$) and K-center with ViT (8.6). DPRG reaches 30.2 and 19.5, improvements of 13.8 and 10.9 points, respectively. The selector ranking changes across backbones.

Candidate selection: support ranking versus coverage selection. We compare direct ranking by distinct-image support, coverage-only selection, and their composition in Table 12. Coverage selection yields a substantially stronger distilled set than support ranking. Applying coverage after support filtering performs best for both downstream backbones: support filtering precedes farthest-first selection, which spreads the retained prototypes across the representation space.

Table 12: Candidate-selection ablation on Flickr30K at $M = 100$. All variants start from 600 candidates.

Backbone	Selection rule	Image Retrieval			Text Retrieval			R_{mean}
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
NF-ResNet50	Coverage only	9.4 ± 0.3	27.8 ± 0.4	39.2 ± 0.4	12.3 ± 0.6	31.0 ± 0.8	41.2 ± 0.8	26.8 ± 0.3
	Support ranking	5.1 ± 0.3	17.1 ± 0.6	26.2 ± 0.7	6.4 ± 0.9	19.6 ± 1.3	29.4 ± 1.7	17.3 ± 0.8
	Support + coverage	**10.7** ± 0.1	**31.2** ± 0.5	**43.1** ± 0.2	**14.7** ± 0.7	**34.8** ± 0.9	**46.5** ± 1.3	**30.2** ± 0.2
ViT	Coverage only	6.1 ± 0.4	18.7 ± 0.2	27.8 ± 0.3	5.8 ± 0.6	17.1 ± 0.8	25.9 ± 1.1	16.9 ± 0.3
	Support ranking	3.5 ± 0.2	12.5 ± 0.4	19.3 ± 0.5	3.2 ± 0.6	11.5 ± 0.6	18.1 ± 1.1	11.4 ± 0.4
	Support + coverage	**7.6** ± 0.1	**21.4** ± 0.1	**31.6** ± 0.1	**7.6** ± 0.5	**20.0** ± 0.8	**28.7** ± 0.2	**19.5** ± 0.2

D.3 Input-Pair Filtering

We remove the lowest-compatibility 20% of the original Flickr30K image–caption pairs using a frozen CLIP similarity score, then run DPRG on the remaining pairs.

At $M = 100$ with NF-ResNet50, the full corpus gives $R_{\text{mean}} = 30.2 \pm 0.2$. Removing the lowest-similarity 20% gives $R_{\text{mean}} = 29.9 \pm 0.3$; the six recall values are listed in Table 13.

Table 13: Input-pair filtering on Flickr30K ($M = 100$, NF-ResNet50).

Method	Image Retrieval			Text Retrieval			R_{mean}
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
DPRG (Ours)	**10.7** ± 0.1	**31.2** ± 0.5	**43.1** ± 0.2	**14.7** ± 0.7	**34.8** ± 0.9	**46.5** ± 1.3	**30.2** ± 0.2
DPRG + filtering	13.5 ± 0.1	31.6 ± 0.1	42.9 ± 0.2	14.9 ± 0.3	33.1 ± 0.6	43.6 ± 0.6	29.9 ± 0.3

D.4 Visualization Details

This section describes how the pair-assignment and prototype plots are constructed.

Pair assignment geometry Figure 3 uses one fixed Flickr30K population containing 21,414 image–caption pairs from 7,552 unique images. The non-pair and pair maps use the same frozen CLIP image-space t-SNE coordinates and randomized draw order; only the mode assigned to each pair, and hence its colour, changes. All 100 modes are coloured.

The four zoom windows in Figure 3 are selected by a deterministic rule. We scan a 60×60 grid with square windows whose side is 9.5% of the embedding-plane span, retain windows containing at least 200 pairs, and score each by the increase from non-pair to pair assignment in the share occupied by its three most frequent modes. We then keep the four highest-scoring non-overlapping windows. Their top-three shares change from 25% to 98%, 20% to 93%, 26% to 97%, and 27% to 98%, respectively.

For a global measure, let $\mathbf{D}_i^{(a)} \in \mathbb{R}^{100}$ be the distribution of image i's captions over modes under assignment $a \in \{\text{non-pair}, \text{pair}\}$, and let $\mathcal{N}_{30}(i)$ be its 30 nearest other images by cosine similarity in the original 768-dimensional CLIP image space. We compute

$$\ell_i^{(a)} = \frac{1}{30} \sum_{j \in \mathcal{N}_{30}(i)} \left(\mathbf{D}_i^{(a)}\right)^\top \mathbf{D}_j^{(a)}, \qquad P_i^{(a)} = \frac{\ell_i^{(a)} - c^{(a)}}{1 - c^{(a)}}, \qquad c^{(a)} = \sum_{m=1}^{100} \left(p_m^{(a)}\right)^2, \tag{16}$$

where $p_m^{(a)}$ is the image-balanced global frequency of mode m. The chance correction is computed separately for the two mode-frequency distributions ($c = 0.0123$ for non-pair and 0.0110 for pair assignment). A mode's score is the mean $P_i^{(a)}$ over images carrying that mode under assignment a. Pair assignment improves all 100 mode scores: the mean rises from 0.2 to 0.6, and the median per-mode gain is $+0.4$.

Coverage visualization. Figure 8 uses the fixed Flickr30K population and frozen CLIP image-space t-SNE coordinates from Figure 3. For each mode, we place its displayed prototype at the mean projected coordinate of the assigned pairs and draw the convex hull of the 100 displayed locations. Clustering and farthest-first selection use the original feature space; the hull is drawn in the projected coordinates.

E Additional Baseline Comparisons

We report multimodal, image-only, and correspondence-clustering comparisons.

E.1 External Distillers

We use the construction procedures of PTM-ST (Guo et al., 2026), EDGE (Zhao et al., 2026), and RepBlend (Zhang et al., 2026), then evaluate their delivered synthetic sets with the CLIP text encoder and the projection head described in Section A.3. A delivered cross-modal relation matrix is resolved once in CLIP space before student training.

Table 14: DPRG and additional baselines on Flickr30K at $M = 100$ under the CLIP-space projection-head protocol.

Method	Image Retrieval			Text Retrieval			R_{mean}
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
NF-ResNet50							
PTM-ST	7.7 ± 0.2	22.3 ± 0.9	32.9 ± 0.7	11.5 ± 0.6	27.9 ± 1.0	37.6 ± 1.1	23.3 ± 0.5
EDGE	6.0 ± 0.3	19.4 ± 0.2	28.5 ± 0.2	8.0 ± 0.3	20.7 ± 1.0	30.4 ± 0.2	18.8 ± 0.2
RepBlend	3.1 ± 0.1	11.4 ± 0.2	18.1 ± 0.1	6.2 ± 0.6	16.4 ± 0.5	22.9 ± 0.6	13.0 ± 0.1
DPRG (Ours)	**10.7** ± 0.1	**31.2** ± 0.5	**43.1** ± 0.2	**14.7** ± 0.7	**34.8** ± 0.9	**46.5** ± 1.3	**30.2** ± 0.2
ViT							
PTM-ST	3.2 ± 0.2	11.0 ± 0.3	17.3 ± 0.4	2.9 ± 0.2	8.2 ± 0.8	13.5 ± 0.5	9.3 ± 0.2
EDGE	3.9 ± 0.1	12.7 ± 0.3	20.2 ± 0.1	3.4 ± 0.1	11.8 ± 1.1	18.8 ± 0.3	11.8 ± 0.2
RepBlend	2.2 ± 0.1	8.3 ± 0.3	13.8 ± 0.3	3.2 ± 0.3	10.5 ± 0.5	15.9 ± 0.3	9.0 ± 0.2
DPRG (Ours)	**7.6** ± 0.1	**21.4** ± 0.1	**31.6** ± 0.1	**7.6** ± 0.5	**20.0** ± 0.8	**28.7** ± 0.2	**19.5** ± 0.2

E.2 Image-Only Distillers Adapted to Paired Data

We transfer MGD^3 (Chan-Santiago et al., 2025) and CoDA (Zhou et al., 2026) to Flickr30K at $M = 100$ by attaching a caption to each discovered visual mode. Each method uses its native mode-discovery and diffusion pipeline. For each mode, we choose a representative real image and the highest-CLIP-similarity caption among its five paired captions. This caption conditions generation and remains paired with the generated image.

MGD^3-Transfer uses full-data K-means on flattened SD 2.1 VAE latents at 512×512 ($n_{\text{init}} = 1$, at most 300 Lloyd iterations, tolerance 10^{-4}), followed by SD 2.1-base sampling with 50 steps, CFG 8.0, mode weight $\lambda = 0.1$, and mode guidance above timestep 500. The nearest real image supplies the caption while the centroid remains the generation target. CoDA-Transfer uses standardized SDXL-VAE latents at 1024×1024, UMAP (40 neighbours, 50 components), HDBSCAN (minimum cluster size 55, `min_samples` $= 3$), and its native split/outlier processing to obtain 100 representatives. Generation uses SDXL-base with DPM-Solver++, Karras sigmas, 25 steps, CFG 5.0, distribution-alignment weight $\gamma = 0.05$, and guidance fraction 0.9; the representative's original latent is the alignment target.

For each method, we use its encoder, generator, resolution, and guidance rule. We use the MGD^3 configuration above and the strongest CoDA setting from its clustering-parameter sweep.

Table 15: Transfer of image-only diffusion distillers to Flickr30K at $M = 100$. MGD^3 and CoDA retain their native mode-discovery and generation pipelines.

Method	Image Retrieval			Text Retrieval			R_{mean}
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
NF-ResNet50							
MGD^3-Transfer	5.0 ± 0.0	15.3 ± 0.1	23.5 ± 0.3	6.6 ± 0.5	17.2 ± 1.2	26.3 ± 0.6	15.7 ± 0.4
CoDA-Transfer	6.6 ± 0.0	20.1 ± 0.2	29.7 ± 0.1	9.4 ± 0.6	24.2 ± 0.6	33.9 ± 0.5	20.7 ± 0.1
DPRG (Ours)	**10.7** ± 0.1	**31.2** ± 0.5	**43.1** ± 0.2	**14.7** ± 0.7	**34.8** ± 0.9	**46.5** ± 1.3	**30.2** ± 0.2
ViT							
MGD^3-Transfer	3.2 ± 0.1	11.0 ± 0.2	17.4 ± 0.4	2.9 ± 0.4	9.4 ± 0.3	15.6 ± 0.7	9.9 ± 0.2
CoDA-Transfer	4.1 ± 0.1	13.4 ± 0.1	21.3 ± 0.1	3.5 ± 0.1	12.8 ± 0.4	20.3 ± 0.3	12.6 ± 0.1
DPRG (Ours)	**7.6** ± 0.1	**21.4** ± 0.1	**31.6** ± 0.1	**7.6** ± 0.5	**20.0** ± 0.8	**28.7** ± 0.2	**19.5** ± 0.2

CoDA-Transfer is stronger than MGD^3-Transfer and trails DPRG by 9.5 and 6.9 R_{mean} points with NF-ResNet50 and ViT, respectively.

E.3 ProCo-Style Correspondence Clustering

ProCo (Dang et al., 2026) uses correspondence patterns to construct clusters. We replace DPRG's prototype-construction stage with correspondence-pattern clustering and call the variant *ProCo-clustering*. Full ProCo additionally optimizes the distilled data through trajectory matching and cluster-structure regularization with a conditional neural-field parameterization.

For each of the $N = 145{,}000$ Flickr30K training pairs, frozen CLIP ViT-L/14 defines image-to-text and text-to-image retrieval distributions over the complete training gallery,

$$\mathbf{h}_i^{\mathsf{I2T}} = \mathrm{softmax}\big(\{\cos(\mathbf{v}_i, \mathbf{t}_j)/\tau\}_{j=1}^{N}\big), \qquad \mathbf{h}_i^{\mathsf{T2I}} = \mathrm{softmax}\big(\{\cos(\mathbf{t}_i, \mathbf{v}_j)/\tau\}_{j=1}^{N}\big). \tag{17}$$

We set $\tau = 0.07$ and define the pairwise dissimilarity as

$$d(i,j) = \mathrm{JSD}(\mathbf{h}_i^{\mathsf{I2T}}, \mathbf{h}_j^{\mathsf{I2T}}) + \mathrm{JSD}(\mathbf{h}_i^{\mathsf{T2I}}, \mathbf{h}_j^{\mathsf{T2I}}). \tag{18}$$

Average-linkage agglomerative clustering produces exactly 100 groups. From each group, we retain the real pair minimizing the same summed divergence to the two cluster-mean distributions, resolving ties by the smallest original pair index. Its original, unnormalized CLIP features form the prototype, and the text feature is mapped to its nearest caption in the full training corpus. We use temperature $\tau = 0.07$, average linkage, and the representative rule above. The variant constructs one real-pair representative per correspondence cluster.

Both variants then use Stable Diffusion 2.1 unCLIP-small, matched per-image noise seeds, and Triple CFG with $(w^{\mathsf{IR}}, w^{\mathsf{TR}}, t_{\mathsf{stop}}) = (1.0, 1.25, 0.5)$ under Table 3. Evaluation uses the fixed final epoch under Section A.3.

Table 16: Prototype construction with ProCo retrieval-distribution clustering on Flickr30K at $M = 100$. ProCo-clustering denotes the correspondence-clustering variant.

Method	Image Retrieval			Text Retrieval			R_{mean}
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
NF-ResNet50							
ProCo-clustering	5.3 ± 0.3	17.5 ± 0.1	26.9 ± 0.1	6.4 ± 0.1	17.9 ± 0.4	25.5 ± 0.6	16.6 ± 0.1
DPRG (Ours)	**10.7** ± 0.1	**31.2** ± 0.5	**43.1** ± 0.2	**14.7** ± 0.7	**34.8** ± 0.9	**46.5** ± 1.3	**30.2** ± 0.2
ViT							
ProCo-clustering	1.7 ± 0.0	7.9 ± 0.2	13.0 ± 0.2	2.0 ± 0.3	7.6 ± 0.2	10.8 ± 0.2	7.2 ± 0.1
DPRG (Ours)	**7.6** ± 0.1	**21.4** ± 0.1	**31.6** ± 0.1	**7.6** ± 0.5	**20.0** ± 0.8	**28.7** ± 0.2	**19.5** ± 0.2

ProCo-clustering reaches $R_{\mathsf{mean}} = 16.6$ with NF-ResNet50 and 7.2 with ViT. The table reports the correspondence-clustering variant; full ProCo adds trajectory matching and neural-field optimization.

F Full-Dataset Retrieval Performance

We train the projection-head evaluator on each complete training split. The image and text encoders remain frozen, and the projection head is trained for 100 epochs with the settings in Section A.3. With no generation stage, results are averaged over three projection-head initializations.

Table 17: Full-dataset retrieval performance under the projection-head evaluator. Values are percentages reported as mean ± standard deviation.

Dataset	Backbone	Image Retrieval			Text Retrieval			R_{mean}
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
Flickr30K	NF-ResNet50	28.5 ± 0.2	59.6 ± 0.1	71.4 ± 0.1	46.0 ± 0.6	76.2 ± 0.3	84.4 ± 0.2	61.0 ± 0.1
	ViT	22.7 ± 0.3	49.9 ± 0.3	62.3 ± 0.3	35.5 ± 0.4	62.9 ± 0.3	73.0 ± 0.4	51.0 ± 0.1
MS-COCO	NF-ResNet50	12.6 ± 0.1	33.4 ± 0.1	46.5 ± 0.2	26.6 ± 0.4	53.1 ± 0.2	66.1 ± 0.1	39.7 ± 0.1
	ViT	11.5 ± 0.1	29.8 ± 0.1	41.7 ± 0.1	19.5 ± 0.2	42.4 ± 0.4	55.3 ± 0.3	33.4 ± 0.1
Flickr8K	NF-ResNet50	24.7 ± 0.0	54.4 ± 0.1	67.4 ± 0.2	39.5 ± 0.4	67.7 ± 0.9	79.1 ± 0.6	55.5 ± 0.3
	ViT	16.9 ± 0.6	41.7 ± 0.4	54.8 ± 0.4	22.6 ± 0.7	50.3 ± 0.4	62.6 ± 1.2	41.5 ± 0.3
Pascal1K	NF-ResNet50	54.1 ± 0.9	90.7 ± 0.8	95.9 ± 0.5	53.3 ± 1.5	87.7 ± 1.5	94.7 ± 1.5	79.4 ± 0.9
	ViT	42.9 ± 1.1	80.1 ± 1.6	88.7 ± 1.5	50.3 ± 3.2	76.7 ± 0.6	87.7 ± 2.3	71.1 ± 1.2

G Accuracy, Compute, and Memory Accounting

This appendix reports the retrieval scores, construction cost, and peak memory represented in Figure 1. All methods use Flickr30K with $M = 100$ and the NF-ResNet50 evaluator. Construction cost covers set construction and excludes the shared projection-head evaluation.

G.1 Retrieval scores

Table 18 reports all six recall measures used to compute R_{mean}. TESLA-VL, LoRS, and PDS use the source-reported values described in Section A; the remaining rows use the evaluations in Tables 14 and 15.

Table 18: Detailed retrieval scores for the Flickr30K, $M = 100$ points in Figure 1. Values are reported as mean ± standard deviation.

Method	Image Retrieval			Text Retrieval			R_{mean}
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
Training-free construction							
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	24.7 ± 0.2
MGD^3-Transfer	5.0 ± 0.0	15.3 ± 0.1	23.5 ± 0.3	6.6 ± 0.5	17.2 ± 1.2	26.3 ± 0.6	15.7 ± 0.4
CoDA-Transfer	6.6 ± 0.0	20.1 ± 0.2	29.7 ± 0.1	9.4 ± 0.6	24.2 ± 0.6	33.9 ± 0.5	20.7 ± 0.1
DPRG (Ours)	**10.7** ± 0.1	**31.2** ± 0.5	**43.1** ± 0.2	**14.7** ± 0.7	**34.8** ± 0.9	**46.5** ± 1.3	**30.2** ± 0.2
Optimization-based construction							
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	15.6 ± 0.4
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	20.1 ± 0.2
PTM-ST	7.7 ± 0.2	22.3 ± 0.9	32.9 ± 0.7	11.5 ± 0.6	27.9 ± 1.0	37.6 ± 1.1	23.3 ± 0.5
EDGE	6.0 ± 0.3	19.4 ± 0.2	28.5 ± 0.2	8.0 ± 0.3	20.7 ± 1.0	30.4 ± 0.2	18.8 ± 0.2
RepBlend	3.1 ± 0.1	11.4 ± 0.2	18.1 ± 0.1	6.2 ± 0.6	16.4 ± 0.5	22.9 ± 0.6	13.0 ± 0.1

G.2 Construction cost and peak memory

Table 19 reports construction GPU-hours and peak single-process memory on NVIDIA A100 80GB GPUs. Completed stages are timed end to end; for fixed-length optimization loops, synchronized per-iteration time is multiplied by the configured iteration count.

The timed stages are feature extraction and synthesis for PDS and DPRG, expert or trajectory preparation and optimization for the optimization-based methods, and representation extraction, prototype discovery, and generation for MGD^3 and CoDA.

Table 19: Construction cost and peak memory for Flickr30K at $M = 100$. The final column lists the timed construction stages.

Method	GPU-hour↓	Peak mem. (GiB)↓	Accounted construction stages
Training-free construction			
PDS	0.289	4.35	CLIP feature extraction and 100-image unCLIP synthesis
MGD^3-Transfer	0.304	9.27	VAE feature pass, clustering, and 100-image SD 2.1 synthesis
CoDA-Transfer	0.829	16.24	VAE feature pass, clustering, and 100-image SDXL synthesis
DPRG (Ours)	**0.329**	**4.35**	CLIP feature extraction and 100-image Triple-CFG synthesis
Optimization-based construction			
TESLA-VL	16.56	11.30	20-expert preparation and 3,000-step distillation
LoRS	16.62	11.30	20-expert preparation and 3,000-step distillation
PTM-ST	18.38	23.94	CLIP-space expert-buffer construction and distillation
EDGE	15.72	11.08	EDGE training and $M = 100$ sampling
RepBlend	6.60	15.39	20 expert trajectories and 3,000-step distillation

G.3 PDS AND DPRG STAGE BREAKDOWN

PDS and DPRG share the encoder, generator, resolution, denoising schedule, and feature cache. Table 20 separates their construction stages. Counterfactual guidance raises synthesis time from 0.131 to 0.171 GPU-hour and total cost from 0.289 to 0.329 GPU-hour (13.84%), while peak memory remains unchanged and R_{mean} rises from 24.7 to 30.2.

Table 20: Stage-level construction cost for PDS and DPRG on Flickr30K at $M = 100$.

Method	Feature extraction (GPU-hour)	Synthesis (GPU-hour)	Total GPU-hour	CPU prototype time (min)	Peak mem. (GiB)
PDS	0.158	0.131	0.289	1.15	4.35
DPRG (Ours)	**0.158**	**0.171**	**0.329**	**3.16**	**4.35**

H QUALITATIVE EXAMPLES

Figure 12 shows representative pairs from the Flickr30K $M = 100$ distilled set. Each caption is the condition supplied to unCLIP for the corresponding prototype.

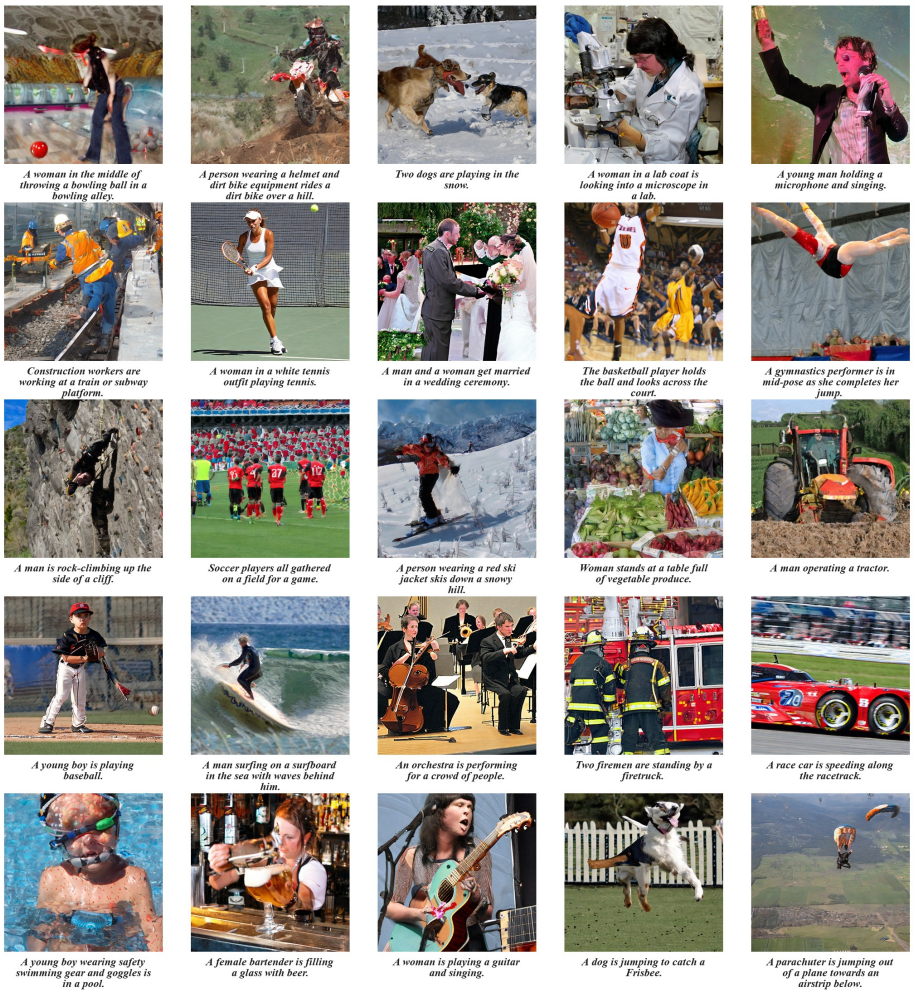


Figure 12: Synthetic images and their corresponding unCLIP conditioning captions for DPRG on Flickr30K with $M = 100$.