CM^2: CROSS-MODAL MODE LEARNING FOR GENERATIVE MULTIMODAL DATASET DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal dataset distillation (MDD) compresses a large image–text corpus into a small synthetic set on which a vision–language model can be trained. Generative MDD methods cluster each modality separately and then pair image and text clusters by a post-hoc assignment; nothing in two independent clusterings forces cluster k to denote the same concept in both modalities, and the non-differentiable matching cannot correct this. We present CM^2, which makes cross-modal correspondence part of mode discovery rather than a step after it. CM^2 learns K image and K text modes, one pair per slot of the distillation budget, in a frozen vision–language embedding space. Its core is cross-modal swapped prediction: the image of a pair must predict the mode assignment of its caption and vice versa, with assignments balanced over the K slots, so that index k comes to denote one concept in both modalities. A mode-level contrastive term aligns same-index modes and keeps the two banks distinct, and a mode-anchoring term ties each mode to the raw-scale centroid of the samples it owns, so the learned modes are decoded directly and are themselves the distilled set. CM^2 requires no offline clustering, no post-hoc matching and no generator fine-tuning. It raises cross-modal index agreement by 30–85$\%$ over post-hoc matching and reduces mismatched image–caption pairs from 7.2% to 0.5% on Flickr30K at the largest budget. Across two benchmarks, four budgets and two evaluation backbones unseen during distillation, it improves mean retrieval recall over the strongest generative MDD baseline in all 16 settings, by $+3.60$ on Flickr30K and $+2.61$ on MS-COCO on average.

1 INTRODUCTION

Vision–language models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) learn transferable multimodal representations from image–text corpora whose scale buys semantic coverage at a high cost in storage, curation and training (Sharma et al., 2018; Lin et al., 2014; Schuhmann et al., 2022; Zhai et al., 2023), which raises the question of whether a competitive model can be trained on far fewer pairs (Wu et al., 2024; Xu et al., 2024). Selecting a subset by coreset methods, herding or filtering (Farahani & Hekmatfar, 2009; Welling, 2009; Gadre et al., 2023) covers the semantic diversity of the source poorly and is itself expensive at corpus scale (Zhao et al., 2025b; Choi et al., 2026).

Dataset distillation (DD) (Wang et al., 2018) instead synthesizes a small set that retains the training value of the original (Zhao et al., 2021; Cazenavette et al., 2022; Zhao & Bilen, 2023), but its conventional form is bi-level and reintroduces the computational cost it sets out to remove (Lee & Chung, 2025; Zhang et al., 2026; Wu et al., 2024; Xu et al., 2024). Generative DD (GDD) avoids the inner loop by conditioning a diffusion prior (Li et al., 2025; Zhou et al., 2025; Chan-Santiago et al., 2025; Zhao et al., 2025b; Choi et al., 2026): dense regions of the data distribution, *i.e.* modes, are obtained by k-means or a GMM and then used to guide generation (Chan-Santiago et al., 2025; Su et al., 2024). Multimodal DD (MDD) extends distillation to the image–text contrastive setting (Wu et al., 2024), where the single-modality design of these clustering procedures becomes the difficulty.

Choi et al. (2026) bring inter-modality information into GDD through a post-hoc linear assignment (Kuhn, 1955), but the clustering itself remains within one modality, so inter-modality relations stay underexplored (Zhao et al., 2025b), and the discrete matching stage is non-differentiable (Burkard & Cela, 1999); Fig. 1 (top). Zhao et al. (2025b) instead fold contrastive

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.

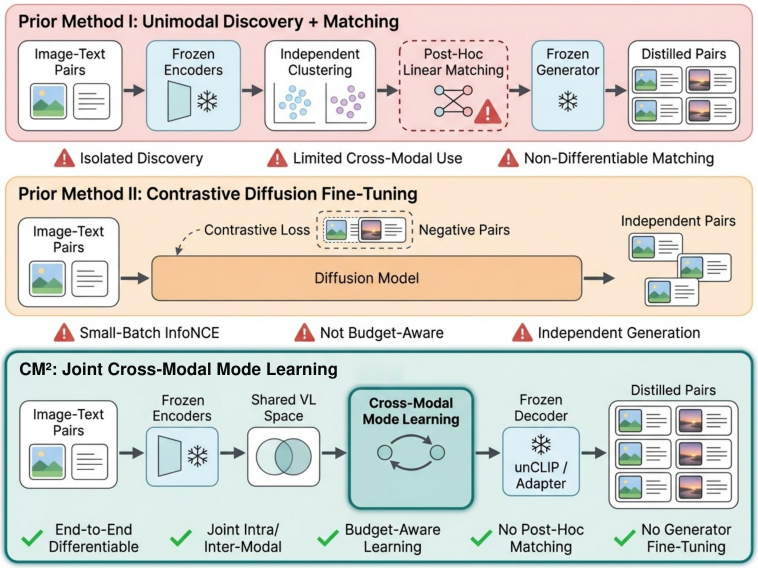

Figure 1: CM^2 compared with existing generative multimodal dataset-distillation paradigms. *Top:* Prototype-based approaches discover image and text modes independently and subsequently align them through a non-differentiable post-hoc matching step, limiting cross-modal interaction during mode discovery. *Middle:* Contrastive diffusion fine-tuning relies on in-batch negative pairs, is not explicitly tied to the distillation budget, and generates distilled pairs independently. *Bottom:* CM^2 instead learns K budget-indexed image–text modes jointly in a frozen vision–language embedding space, directly establishing cross-modal correspondence without post-hoc matching or generator fine-tuning. A frozen decoder or lightweight adapter then synthesizes one distilled pair from each learned mode.

training into diffusion fine-tuning (Radford et al., 2021; Oord et al., 2018), which aligns the modalities but drops the distillation budget from training and inference alike: the generator is never asked to cover the target distribution with a given number of pairs, and the result is a set of independently drawn samples rather than a compact one; Fig. 1 (middle). Folding InfoNCE into generator fine-tuning is also awkward in practice, since generative training caps the batch size while contrastive learning needs many in-batch negatives.[1]

In this paper we propose CM^2, a generative MDD framework that learns image and text modes jointly and differentiably under an explicit distillation budget, with no offline mode extraction, no post-hoc matching and no generative-model fine-tuning. Working in the shared embedding space of a frozen vision–language model (Radford et al., 2021; Jia et al., 2021), CM^2 casts MDD as a self-supervised problem: given a pair, the two modalities are treated as complementary realizations of one concept, and the mode assignment of each is predicted from the representation of the other. Modes are therefore discovered by prediction consistency rather than by an offline clustering procedure, and both intra- and inter-modality structure enter the discovery step.

Because the assignment is differentiable, a mode-level contrastive term can be added to it, matching each mode with its counterpart in the other modality while keeping the two banks from collapsing onto one another. A second term anchors each mode to the region its assigned samples occupy, which removes the last non-differentiable stage: the learned modes are the distilled samples, handed directly to the generator.

Our contributions are summarized as follows:

- We propose CM^2, a budget-aware generative MDD framework that learns K image and K text modes jointly and differentiably in a frozen vision–language embedding space, without offline clustering or generator fine-tuning. Each same-index pair becomes one distilled image–text pair, so K is both the optimization budget and the size of the distilled set.

[1]Zhao et al. (2025b) fine-tune with a batch size of 8, very small for InfoNCE (He et al., 2020; Wu et al., 2018).

- We introduce cross-modal swapped mode learning with an equipartition constraint and a mode-level contrastive objective, which establishes the correspondence between the two banks during optimization instead of recovering it afterwards, while keeping the budget slots evenly used and the two banks distinct.
- We introduce mode anchoring, a scale-free term that regresses each mode onto the raw-scale soft centroid of the samples the transport plan assigns to it. Its target is read off the real data rather than set by hand, so one weight transfers across datasets and budgets. It removes the last non-differentiable step of the pipeline and makes the learned modes the distilled samples, and it is what allows the distilled bank to be rendered by a generator other than the one the method was developed with.
- Across two benchmarks, four budgets and two evaluation backbones unseen during distillation, CM^2 improves mean recall over the strongest generative MDD baseline in all 16 evaluated settings, by $+3.60$ on Flickr30K and $+2.61$ on MS-COCO on average, and raises cross-modal index agreement in every setting. The advantage persists when the generator, or the encoder and the generator together, are replaced.

2 Related Work

2.1 Conventional Dataset Distillation

Dataset distillation (DD) compresses a training set $\mathcal{D}$ into a small synthetic set $\mathcal{S}$ on which a model trains to comparable accuracy (Wang et al., 2018). Early formulations are bi-level, optimizing $\mathcal{S}$ through the training dynamics of a model fitted to it (Lee et al., 2022b; Zhao & Bilen, 2021; Du et al., 2024; Kim et al., 2022); differentiating through that inner loop is the cost subsequent work removes, by matching gradients (Zhao et al., 2021; Zhao & Bilen, 2021), features or distributions (Wang et al., 2022; Zhao & Bilen, 2023; Liu et al., 2023; Zhang et al., 2024), or training trajectories (Cazenavette et al., 2022; Du et al., 2023; Cui et al., 2023), and more recently through decoupled optimization, compact parameterizations, frequency-domain representations and neural fields (Yin et al., 2023; Shao et al., 2024a; Sun et al., 2024; Tran et al., 2025; Shao et al., 2024b; Liu et al., 2022; Deng & Russakovsky, 2022; Wei et al., 2023; Shin et al., 2023; 2025; Zhang et al., 2023b; Feng et al., 2024). These methods stay tied to the architecture, optimizer and budget they were tuned for.

2.2 Generative Dataset Distillation

Generative DD (GDD) replaces pixel optimization with a pretrained generative prior. One line optimizes latent codes of GANs (Zhao & Bilen, 2022; Zhang et al., 2023a), StyleGAN (Cazenavette et al., 2023; Zhong et al., 2025) or latent diffusion models (Moser et al., 2025; Wang et al., 2025); a second distills the corpus into the parameters or the sampling behavior of the generator (Wang et al., 2024; Gu et al., 2024; Su et al., 2024); a third estimates dense regions of the data distribution, *i.e.* modes, and conditions generation on them (Su et al., 2024; Chan-Santiago et al., 2025; Zhao et al., 2025a). In MGD^3, representative of the third, class-wise modes are found by k-means or a GMM and sampling is steered toward distinct modes, raising intra-class diversity at no training cost (Chan-Santiago et al., 2025). All three lines target unimodal classification, and the third discovers its modes offline, within one modality, by a non-differentiable procedure.

2.3 Multi-modal Dataset Distillation

Multimodal DD (MDD) targets the joint image–text distribution rather than a class-conditional image distribution, so the distilled set must preserve alignment as well as diversity (Wu et al., 2024; Xu et al., 2024; Radford et al., 2021). Optimization-based methods carry the conventional machinery over: MTT-VL (Wu et al., 2024) optimizes synthetic images and text features by trajectory matching, TESLA-VL (Cui et al., 2023) bounds the memory this requires, LoRS (Xu et al., 2024) distills a low-rank image–text similarity matrix in place of a one-to-one correspondence, and later work addresses modality collapse and robustness (Zhang et al., 2026; Dang et al., 2025). All inherit the cost and the architecture dependence of the methods they extend (Zhong & Liu, 2023).

Generative MDD methods differ chiefly in *where* cross-modal correspondence is established. EDGE (Zhao et al., 2025b) imposes it on the generator, fine-tuning a diffusion model with an aux-

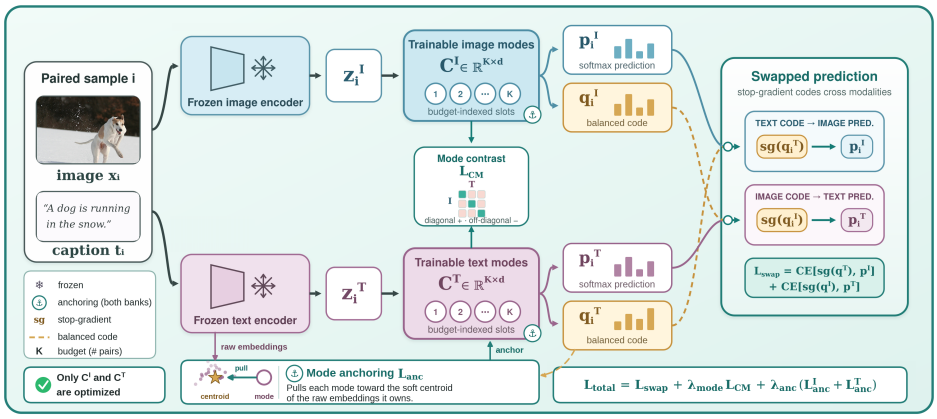

Figure 2: Overview of CM^2. Frozen encoders map a paired image and caption into a shared embedding space. In each modality, similarities to K trainable, budget-indexed modes give a softmax prediction p_i^m and a balanced code q_i^m satisfying the equipartition constraint of Eq. (4). The stop-gradient codes are swapped across modalities, so the text code supervises the image prediction and the image code the text prediction (Eq. (5)), while $\mathcal{L}_{\text{CM}}$ aligns same-index modes and separates the rest. Mode anchoring (Eq. (8)) pulls each mode toward the soft centroid of the raw embeddings it owns; it acts on both banks and is drawn for the text branch. Only C^I and C^T are optimized.

iliary contrastive objective; the distilled pairs are then drawn independently, so the budget enters only as the number of samples requested. Prototype-guided data synthesis (PDS) (Choi et al., 2026) keeps the generator frozen and recovers correspondence after the fact, clustering CLIP image and text embeddings separately and matching the two prototype sets by linear assignment (Sculley, 2010; Kuhn, 1955) before decoding with unCLIP (Ramesh et al., 2022); the budget fixes the number of clusters, but mode discovery in each modality proceeds blind to the other. Jeong et al. (2026) initialize synthetic pairs from clusters of the *joint* embedding space and update them against interpolated teachers, so the modes themselves are never optimized. These methods are not all evaluated on one scale: EDGE and Jeong et al. (2026) fine-tune the evaluation backbone and take sentence features from a frozen BERT, whereas Choi et al. (2026), whose protocol we adopt, keeps the backbone frozen and uses the CLIP text encoder. We therefore compare only against methods evaluated under a single protocol.

CM^2 departs from all three: correspondence is neither imposed on the generator nor recovered post hoc but is a property of the optimization, both banks being learned jointly under a shared index and an equipartition constraint defined over the budget. We share with Jeong et al. (2026) the premise that the joint space carries structure per-modality clustering discards, but they use it only to initialize pairs that are then matched against a teacher.

3 Method

3.1 Notation and Problem Formulation

Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{t}_i)\}_{i=1}^N$ be an image–text corpus. MDD seeks a compact synthetic set $\mathcal{S} = \{(\tilde{\mathbf{x}}_k, \tilde{\mathbf{t}}_k)\}_{k=1}^K$, $K \ll N$, on which a vision–language model trains to the behavior it would acquire from $\mathcal{D}$:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathbb{E}_{\mathcal{B}\sim\mathcal{D}} \left[|\mathcal{L}_{\text{ITC}}(\mathcal{B}; \theta_{\mathcal{S}}) - \mathcal{L}_{\text{ITC}}(\mathcal{B}; \theta_{\mathcal{D}})| \right], \qquad \theta_{\mathcal{S}} = \arg\min_{\theta} \mathcal{L}_{\text{ITC}}(\mathcal{S}; \theta), \quad (1)$$

where $\mathcal{B}$ is a batch of pairs and $\mathcal{L}_{\text{ITC}}$ is the bidirectional InfoNCE objective of Radford et al. (2021) at temperature τ, computed on the unit-normalized embeddings $\mathbf{z}_i^m = \mathbf{e}_i^m / \|\mathbf{e}_i^m\|_2$, $m \in \{I, T\}$, of the raw encoder outputs $\mathbf{e}_i^I = f_\phi^I(\mathbf{x}_i)$ and $\mathbf{e}_i^T = f_\phi^T(\mathbf{t}_i)$.

Solving Eq. (1) directly requires optimizing synthetic pairs through the training dynamics of a vision–language model. We instead distill in the embedding space of a frozen pretrained model, and learn K paired modes $\{(\mathbf{c}_1^I, \mathbf{c}_1^T), \ldots, (\mathbf{c}_K^I, \mathbf{c}_K^T)\}$, each pair a distilled image–text semantic unit, which a frozen generative model later renders into a viewable pair.

Table 1: Mean retrieval recall across multimodal dataset-distillation methods. Each entry is the mean of IR@$\{1, 5, 10\}$ and TR@$\{1, 5, 10\}$ on the real test split for a distilled set of K pairs, evaluated with a frozen backbone unseen during distillation under the cross-architecture protocol of Choi et al. (2026): the image backbone is frozen and the text encoder is the frozen CLIP text encoder with a learnable linear projection. Best per column in **bold**. †Reported by Choi et al. (2026). PDS is the published number at $K \in \{100, 300\}$ and our reproduction at $K = 500$; "–" marks a setting the source does not report.

	Flickr30K						**MS-COCO**					
	NF-ResNet50			ViT-T/16			NF-ResNet50			ViT-T/16		
Method	100	300	500	100	300	500	100	300	500	100	300	500
Frozen CLIP text encoder (protocol of Choi et al. (2026))												
TESLA-VL† (Cui et al., 2023)	15.55	29.93	–	8.00	16.20	–	6.60	13.27	–	1.55	6.68	–
LoRS† (Xu et al., 2024)	20.13	26.32	–	11.35	14.48	–	7.12	9.00	–	2.73	4.05	–
PDS (Choi et al., 2026)	24.73	37.57	40.91	17.58	24.68	28.00	11.67	18.00	19.75	8.08	12.47	13.79
CM^2 (ours)	**30.97**	**41.14**	**45.61**	**20.07**	**27.54**	**30.41**	**13.67**	**21.00**	**23.80**	**10.46**	**14.48**	**15.88**
Full dataset†		61.02			51.05			39.72			33.37	

The modes are trainable parameters. This is what separates CM^2 from the clustering-based pipelines it replaces (Choi et al., 2026; Li et al., 2025; Chan-Santiago et al., 2025; Zhou et al., 2025), which alternate between assigning single-modality features to clusters and recomputing each cluster from the features assigned to it, one modality at a time, and only afterwards pair the two mode sets. Following contrastive instance learning (Wu et al., 2018), we treat the image and the caption of a pair as two realizations of one concept and require the cluster assignment, or *code*, predicted from one modality to agree with the code of the other. Mode discovery becomes differentiable, which in turn admits an inter-modality contrastive term that aligns same-index modes and keeps the two banks from collapsing onto each other.

We use CLIP (Radford et al., 2021) as the frozen encoder pair and, following Choi et al. (2026), an unCLIP decoder, since off-the-shelf Stable Diffusion was not trained to consume a CLIP image embedding; Sec. 4.5 shows an adapter-based alternative (Ye et al., 2023) serves as well.

3.2 Cross-Modal Mode Learning

Mode-based generative MDD discovers image modes without knowing how the captions are organized, and text modes without knowing which images they co-occur with. CM^2 removes that separation by requiring each modality to predict the other's mode assignment (Fig. 2).

We parameterize the K image and K text modes as two learnable matrices $C^I = [\mathbf{c}_1^I, \dots, \mathbf{c}_K^I]^\top \in \mathbb{R}^{K\times d}$ and $C^T = [\mathbf{c}_1^T, \dots, \mathbf{c}_K^T]^\top \in \mathbb{R}^{K\times d}$, with rows $\mathbf{c}_k^I, \mathbf{c}_k^T \in \mathbb{R}^d$. The terms of this section and the next act on the row-normalized banks, and we write $\mathbf{c}_k^m$ for $\mathbf{c}_k^m / \|\mathbf{c}_k^m\|_2$ wherever a mode enters a similarity; the rows themselves are left free to carry the scale set in Sec. 3.4. For an image feature $\mathbf{z}_i^I$ and a text feature $\mathbf{z}_i^T$, the soft mode predictions are

$$p_i^I = \mathrm{softmax}\left(\frac{C^I \mathbf{z}_i^I}{\tau_p}\right), \qquad p_i^T = \mathrm{softmax}\left(\frac{C^T \mathbf{z}_i^T}{\tau_p}\right), \tag{2}$$

where τ_p controls assignment sharpness. To rule out the degenerate solution in which all samples collapse onto a few modes, the balanced codes q_i^I and q_i^T are computed under an equipartition constraint, following online swapped-assignment learning (Caron et al., 2020). Concretely, for modality $m \in \{I, T\}$, let $\mathrm{Sim}^m \in \mathbb{R}^{K\times|\mathcal{B}|}$ collect the mode–sample similarity scores of all $|\mathcal{B}|$ samples, $[\mathrm{Sim}^m]_{k,i} = (\mathbf{c}_k^m)^\top \mathbf{z}_i^m$. The assignment matrix Q^m is then obtained by solving an entropy-regularized optimal-transport problem,

$$Q^m = \arg\max_{Q\in\mathcal{Q}} \mathrm{Tr}\big(Q^\top \mathrm{Sim}^m\big) + \epsilon\, H(Q), \tag{3}$$

where $H(Q) = -\sum_{k,i} Q_{ki} \log Q_{ki}$ is the entropy, ϵ controls its strength, and

$$\mathcal{Q} = \left\{ Q \in \mathbb{R}_+^{K\times|\mathcal{B}|} \;:\; Q\mathbf{1}_{|\mathcal{B}|} = \tfrac{1}{K}\mathbf{1}_K, \quad Q^\top \mathbf{1}_K = \tfrac{1}{|\mathcal{B}|}\mathbf{1}_{|\mathcal{B}|} \right\} \tag{4}$$

is the transportation polytope whose marginals split the samples equally over the K modes, so no mode is unused and none absorbs the batch. Eq. (3) is solved by a few Sinkhorn–Knopp iterations (Caron et al., 2020), and q_i^m is the i-th column of Q^m rescaled to sum to one.

Treating $\mathbf{z}_i^I$ and $\mathbf{z}_i^T$ as complementary realizations of one instance, we require each to predict the other's assignment, giving the cross-modal swapped loss

$$\mathcal{L}_{\text{swap}} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left[(q_i^T)^\top \log p_i^I + (q_i^I)^\top \log p_i^T\right]. \tag{5}$$

Assignments are formed online and are shaped by cross-modal consistency, so each modality contributes to how the other partitions the corpus.

3.3 Cross-Modality Contrastive Alignment

Since the two banks are learned jointly, same-index modes are two views of one concept, and contrast can be applied at the level of the modes themselves: for image mode $\mathbf{c}_k^I$ the positive is $\mathbf{c}_k^T$ and the negatives are $\{\mathbf{c}_l^T\}_{l\neq k}$, and symmetrically for each text mode. With mode-level temperature τ_m and cross-modal similarity $s_{kl} = (\mathbf{c}_k^I)^\top \mathbf{c}_l^T/\tau_m$ between the k-th image mode and the l-th text mode, the bidirectional mode-level contrastive loss is

$$\mathcal{L}_{\text{CM}} = -\frac{1}{2K}\sum_{k=1}^{K}\left[\log\frac{\exp(s_{kk})}{\sum_{l=1}^{K}\exp(s_{kl})} + \log\frac{\exp(s_{kk})}{\sum_{l=1}^{K}\exp(s_{lk})}\right]. \tag{6}$$

The two terms align each mode with its counterpart in either direction. The negatives do work the equipartition constraint cannot: Eq. (4) keeps the samples spread over the K slots, whereas the negatives keep C^I and C^T from converging on a single shared matrix and losing the modality-specific structure the distilled set is meant to carry.

3.4 Mode Anchoring

Both terms so far are defined on row-normalized views of the banks — a softmax over cosine similarities, and inner products between unit directions — so neither constrains the radial coordinate of a mode, nor any component shared by every embedding in the corpus, which is common to all logits and cancels in the softmax. The objective fixes the direction of each mode and leaves its position within the encoder's output distribution undetermined. This does not matter for clustering, but it matters at synthesis time: the decoder is a conditional model estimated on that distribution, and its behavior is specified only where the estimate is supported.

Prior generative pipelines resolve this with a non-parametric pass after optimization that replaces each mode by the mean of the samples assigned to it (Appendix A). We instead constrain the modes during optimization, so that the quantity being optimized is also the quantity being decoded. Let Q^m be the transport plan of Eq. (3), held fixed. The plan assigns a distribution of mass over samples to each slot, whose weighted first moment

$$\boldsymbol{\mu}_k^m = \frac{\sum_i Q_{ki}^m \mathbf{e}_i^m}{\sum_i Q_{ki}^m} \tag{7}$$

is the location in the encoder's raw output space that slot k currently summarizes. We regularize each mode toward that location,

$$\mathcal{L}_{\text{anc}}^m = \frac{1}{K}\sum_{k=1}^{K}\frac{\left\|\mathbf{c}_k^m - \text{sg}[\boldsymbol{\mu}_k^m]\right\|_2^2}{\left\|\text{sg}[\boldsymbol{\mu}_k^m]\right\|_2^2}, \tag{8}$$

where $\text{sg}[\cdot]$ is the stop-gradient operator, so the term acts on the modes alone and does not propagate into the assignment. Two properties are deliberate. The target is a statistic of the real corpus rather than a quantity introduced by hand, so the constraint carries no free parameter beyond its weight; and normalizing the residual by the magnitude of that target makes the term invariant to the scale of the embedding space, so one weight applies across modalities, datasets and budgets. We fix $\lambda_{\text{anc}} = 1$ throughout and never select it against retrieval.

3.5 Overall Objective and Distilled Dataset

The complete objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{swap}} + \lambda_{\text{mode}}\mathcal{L}_{\text{CM}} + \lambda_{\text{anc}}\big(\mathcal{L}_{\text{anc}}^I + \mathcal{L}_{\text{anc}}^T\big), \tag{9}$$

and is minimized over C^I and C^T alone; the encoders stay frozen, so the procedure never differentiates through them and is independent of the downstream architecture.

Under Eq. (9) each bank is fixed in direction by the cross-modal terms and in position by Eq. (8), so it is a valid decoder condition when training ends. The distilled dataset is therefore (C^I, C^T) itself: the optimized parameters are the compressed corpus, and nothing is estimated between optimization and use. The frozen decoder that renders $\mathbf{c}_k^I$ is a fixed readout, not a stage of the method. We retain the agreement cells Ω_k of Appendix A only as a diagnostic of index agreement, and the post-hoc means only as the construction anchoring is compared against.

Synthesis. Each text mode $\mathbf{c}_k^T$ is mapped to its nearest caption in the training corpus, and the image mode $\mathbf{c}_k^I$ is decoded by unCLIP with that caption as the text condition. The released pair is (decoded image, $\mathbf{c}_k^T$): the retrieved caption conditions the decoder and is published with the image, but the downstream projection head is trained on $\mathbf{c}_k^T$ directly.

4 Experiments

4.1 Experimental Setup

Datasets and budgets. We use the benchmarks adopted by prior MDD work (Wu et al., 2024; Xu et al., 2024; Choi et al., 2026): Flickr30K (Plummer et al., 2015) and MS-COCO (Lin et al., 2014) with the Karpathy splits (Karpathy & Fei-Fei, 2015), distilled from the training split and evaluated on the real test split, at budgets $K \in \{100, 300, 500, 1000\}$. A fixed K is not a fixed compression ratio — $K = 1000$ retains 3.45% of Flickr30K against 0.88% of MS-COCO — and in CM^2 it is not a sampling parameter but the number of modes the optimization is defined over.

Evaluation protocol. Following prior work (Wu et al., 2024; Xu et al., 2024; Choi et al., 2026), quality is measured by bidirectional retrieval on the real test split after training a downstream model on the distilled set: Recall@$\{1, 5, 10\}$ in both directions, denoted IR@k and TR@k, and their mean. The *same* distilled set is evaluated with two frozen image backbones never involved in distillation, NF-ResNet50 and ViT-T/16, and every number is a mean over five independently initialized runs.

Two aspects of this protocol, shared by every method compared here, affect how the numbers should be read. The evaluation loop of Choi et al. (2026) selects its checkpoint on the test split; we reuse it unmodified for comparability and repeat the anchoring comparison under a validation-selected protocol, under which its sign is unchanged. The reported dispersion covers downstream initialization alone, while mode-bank initialization and decoder sampling each contribute more, so differences of about one recall point are not resolvable from the printed interval; where such a difference is claimed it is a paired change at a fixed decoder seed, reproduced with both stages resampled (Appendix C).

Implementation details. We use a frozen CLIP ViT-L/14 (Radford et al., 2021) as f_ϕ^I and f_ϕ^T, giving $d = 768$; embeddings are extracted once and cached, so the encoders are evaluated a single time over the corpus. The two mode matrices are optimized with AdamW for $T = 1000$ full-batch iterations, and images are synthesized with a frozen unCLIP decoder (Ramesh et al., 2022) at 224×224. Distillation requires no gradient through either the encoders or the decoder. Appendix B lists every remaining hyperparameter.

4.2 Comparison with State-of-the-Art

Table 1 compares CM^2 with prior MDD methods under the cross-architecture protocol of Choi et al. (2026); Appendix D gives the per-metric breakdown against PDS (Choi et al., 2026), the strongest generative baseline under that protocol. On Flickr30K, CM^2 is ahead at every budget and on both backbones; at $K = 300$ with NF-ResNet50 it reaches 41.14 mean recall against 29.93 for TESLA-VL and 26.32 for LoRS. Over the eight settings of each dataset in Appendix D, the margin over PDS averages $+3.60$ on Flickr30K and $+2.61$ on MS-COCO and is positive in all sixteen. The smaller MS-COCO margin tracks the compression ratio, four times tighter at a fixed K, under which absolute recall is also lower for every method. Following Choi et al. (2026) we exclude MTT-VL (Wu et al., 2024), whose memory requirement at these budgets exceeds that of TESLA-VL and LoRS without a corresponding gain.

Table 2: Component ablation at $K = 500$. Each intermediate row removes or replaces exactly one component of the full method; SWAP ONLY replaces the post-hoc matching of Choi et al. (2026) with the swapped-prediction objective of Eq. (5). All rows involving the decoder share a decoder seed.

	Flickr30K		MS-COCO	
Objective	NF-RN50	ViT-T	NF-RN50	ViT-T
PDS: k-means + Hungarian (Choi et al., 2026)	40.91	28.00	19.75	13.79
Learned modes + Hungarian (no swapped prediction)	43.28	29.53	20.36	14.69
Swap only ($\lambda_{\text{mode}} = 0$)	43.31	29.90	20.98	15.86
Swap + $\mathcal{L}_{\text{CM}}$, positives only	43.10	29.28	19.94	14.57
Swap + $\mathcal{L}_{\text{CM}}$, post-hoc agreement-cell means	43.21	30.09	21.23	15.29
Swap + $\mathcal{L}_{\text{CM}}$, direct mode decoding, no $\mathcal{L}_{\text{anc}}$	43.13	28.29	20.93	14.63
Full CM^2: Swap + $\mathcal{L}_{\text{CM}}$ + $\mathcal{L}_{\text{anc}}$, modes decoded directly	**45.61**	**30.41**	**23.80**	**15.88**

4.3 ABLATION

Table 2 isolates each term at $K = 500$, starting from the PDS pipeline of k-means and Hungarian matching. The full objective is best in all four settings, improving over PDS by $+4.70$, $+2.41$, $+4.05$ and $+2.09$ recall. Replacing k-means with a learned bank while keeping the post-hoc assignment leaves the method well short of the full objective, so the gain is not a property of the clustering procedure; removing the negatives of $\mathcal{L}_{\text{CM}}$ is harmful in all four settings. Swapped prediction fixes which index a concept occupies, and $\mathcal{L}_{\text{CM}}$ aligns same-index modes while keeping the rest apart. With NF-ResNet50, every rung without $\mathcal{L}_{\text{anc}}$ trails the full method by at least 2.3 recall on either dataset; with ViT-T/16 the rungs are closer, and swap-only is within 0.02 on MS-COCO.

What anchoring changes. The effect of $\mathcal{L}_{\text{anc}}$ is measurable before any downstream training. On Flickr30K at $K = 100$ the agreement-cell means of the post-hoc construction (Appendix A) sit close to the real embeddings (nearest-real cosine $0.90/0.87$ for the image and text bank) but pair the banks weakly (pair margin $+0.161$) and are mutually redundant (between-mode cosine 0.800), while unanchored modes invert that trade-off ($+0.855$, -0.007, $0.38/0.49$). Anchoring optimizes neither property directly, lands between the two ($+0.710$, 0.214, $0.73/0.78$), and is the construction that decodes best (Table 8). It also requires a bank initialized at the real embedding scale with a learning rate scaled to match; applying that parameterization without the term leaves the mode norm unconstrained, and it grows well outside the range on which the decoder was estimated (Table 10), so the gain is attributable to $\mathcal{L}_{\text{anc}}$ and not to the parameterization it requires.

4.4 ANALYSIS

Cross-modal index agreement. CM^2 is built so that index k denotes the same concept in both modalities, and that is measurable on the distilled set itself, with nothing trained downstream. Assigning every training pair to its nearest image mode and, independently, to its nearest text mode, the *agreement rate* is the fraction of pairs whose two indices coincide, the population of the agreement cell Ω_k (Appendix A). CM^2 raises it over post-hoc matching in all eight dataset–budget settings, by 30–85% relative, and leaves fewer budget slots empty in every setting (at most 1.1%, against up to 7.6%). The consequence is visible in the released pairs. Scored against their own captions by an independent CLIP ViT-B/32, the two methods coincide over most of the distribution and separate in the lower tail: on Flickr30K at $K = 1000$, 0.5% of CM^2 pairs fall below a cosine of 0.25 against 7.2% for post-hoc matching, which pairs a synthesized baseball scene with a caption describing a fruit and vegetable stand. Such mismatches arise when independently discovered modes are assigned to each other despite describing different semantic regions, and their rate grows with the budget under post-hoc matching (1.0% at $K = 100$ to 7.2% at $K = 1000$), as the number of correspondences it must recover grows, while CM^2 stays at or below 0.5%. Appendix E gives the full diagnostics (Tables 6 and 7) and the score distributions with ranked examples (Fig. 3).

Cost. Making correspondence part of the optimization is also cheaper than recovering it afterwards. Encoding, decoding and downstream training are identical for both methods; only mode mining differs, and there CM^2 replaces two clusterings and an $O(K^3)$ assignment with a single

Table 3: Transfer to other encoders and generators. Flickr30K, $K = 100$, mean recall; each row re-runs the full pipeline with λ_{anc} and τ_m fixed at the values of Appendix B. Karlo (Lee et al., 2022a) takes the same CLIP ViT-L/14 embedding as unCLIP and is a decoder substitution alone; IP-Adapter (Ye et al., 2023) also changes the encoder.

		NF-ResNet50		ViT-T/16	
Encoder + generator	Method	Mean recall	Δ	Mean recall	Δ
CLIP ViT-L/14 + Karlo	PDS (Choi et al., 2026)	25.17	–	16.67	–
	CM^2 (ours)	**29.61**	+4.44	**19.61**	+2.94
OpenCLIP ViT-H/14 + SD1.5 + IP-Adapter	PDS (Choi et al., 2026)	25.30	–	17.16	–
	CM^2 (ours)	**32.35**	+7.04	**20.93**	+3.77

optimization over $2Kd$ parameters. On Flickr30K it takes 13 s against 172 s at $K = 100$ and 79 s against 1667 s at $K = 1000$, a margin that widens with the budget as the accounting predicts. Both are dominated by decoding, at $\approx$4.8 s per released pair, so mode learning is under two per cent of the pipeline. Appendix G gives the accounting and the full measurements.

4.5 Transfer to Other Encoders and Generators

The generator enters CM^2 through one assumption: that it consumes embeddings produced by the encoder used for distillation and behaves as intended where those embeddings lie. We test it at two further points in (encoder, generator) space, with the objective unmodified and λ_{anc} and τ_m fixed at the values of Appendix B (Table 3, Flickr30K at $K = 100$).

Substituting the generator, and the encoder with it. Karlo (Lee et al., 2022a) is an independently trained unCLIP model, built by a different group on a different corpus, that consumes the same CLIP ViT-L/14 embedding as the decoder of Sec. 4.1; substituting it changes the readout and nothing else, since the bank is neither re-optimized nor rescaled. This isolates the generator in a way an image-prompt adapter cannot, since such adapters also replace the encoder. The margins over PDS are $+4.44$ with NF-ResNet50 and $+2.94$ with ViT-T/16. The second configuration replaces both components: modes are learned in OpenCLIP ViT-H/14 (Cherti et al., 2023), so the banks are 1024- rather than 768-dimensional, and are rendered by Stable Diffusion (Rombach et al., 2022) through IP-Adapter (Ye et al., 2023), which conditions on an image embedding by decoupled cross-attention rather than by inverting it. The margin is $+7.04$ with NF-ResNet50 and $+3.77$ with ViT-T/16, so the advantage survives a change of both the space the modes are learned in and the mechanism that renders them, with $\lambda_{\text{anc}} = 1$ carried across the change of dimension unadjusted, which the scale-free form of Eq. (8) permits.

Anchoring is what makes the bank portable. Without $\mathcal{L}_{\text{anc}}$, Karlo collapses the distilled set, rendering every mode as a near-identical generic image: with NF-ResNet50, mean recall falls from 29.61 to 1.60, against a drop of 1.14 under unCLIP in the same caption-free setting. The collapse is not a matter of scale, since it persists when the unanchored bank is rescaled to the radius of the real embeddings. Without anchoring, the distilled set cannot be rendered by an unfamiliar generator (Appendix F.2).

5 Conclusion

We presented CM^2, which learns paired image and text modes by differentiable optimization in a frozen vision–language embedding space. Correspondence is established during mode learning rather than recovered afterwards, and mode anchoring places each mode where the generator can render it, so the optimized parameters are the distilled set. CM^2 improves mean recall over the strongest generative MDD baseline in all 16 evaluated settings, and the advantage persists under a substituted generator and encoder.

AI use statement

Generative AI tools were used for writing assistance, including editing for clarity and concision; for coding assistance in experimental and analysis scripts; and for generating illustrative figures used to

visualize the proposed method. They were not used to generate research ideas, formulate or refine hypotheses, design the research methodology, select or interpret results, or produce reported numerical results. All AI-assisted text and figures were reviewed and revised, and all AI-assisted code was inspected and tested. Reported results were verified against the outputs of the corresponding experiments. The authors take responsibility for the final content of this work.

REFERENCES

Rainer E Burkard and Eranda Cela. Linear assignment problems and extensions. In *Handbook of combinatorial optimization: Supplement volume A*, pp. 75–149. Springer, 1999.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

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. 4750–4759, 2022.

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

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.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

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, Xiaojun Chang, and Ivor W Tsang. Multi-modal dataset distillation in the wild. *arXiv preprint arXiv:2506.01586*, 2025.

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

Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3758, 2023.

Jiawei Du, Qin Shi, and Joey Tianyi Zhou. Sequential subset matching for dataset distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

Reza Zanjirani Farahani and Masoud Hekmatfar. *Facility location: concepts, models, algorithms and case studies*. Springer Science & Business Media, 2009.

Yunzhen Feng, Shanmukha Ramakrishna Vedantam, and Julia Kempe. Embarrassingly simple dataset distillation. In *The Twelfth International Conference on Learning Representations*, 2024.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

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*, 2026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 11102–11118, 2022.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Donghoon Lee, Jiseob Kim, Jisu Choi, Jongmin Kim, Minwoo Byeon, Woonhyuk Baek, and Saehoon Kim. Karlo-v1.0.alpha on coyo-100m and cc15m. `https://github.com/kakaobrain/karlo`, 2022a.

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

Yongmin Lee and Hye Won Chung. Covmatch: Cross-covariance guided multimodal dataset distillation with trainable text encoder. *Advances in Neural Information Processing Systems*, 38: 101490–101518, 2025.

Longzhen Li, Guang Li, Ren Togo, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Generative dataset distillation based on self-knowledge distillation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

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.

Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. In *NeurIPS*, pp. 1100–1113, 2022.

Yanqing Liu, Jianyang Gu, Kai Wang, Zheng Zhu, Wei Jiang, and Yang You. DREAM: Efficient dataset distillation by representative matching. In *ICCV*, pp. 17314–17324, 2023.

Brian B Moser, Federico Raue, Sebastian Palacio, Stanislav Frolov, and Andreas Dengel. Latent dataset distillation with diffusion models. *WACV*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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 *ICCV*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 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.

Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. In *CVPR*, pp. 16709–16718, 2024a.

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Donghyeok Shin, Seungjae Shin, and Il-Chul Moon. Frequency domain-based dataset distillation. In *NeurIPS*, pp. 70033–70044, 2023.

Donghyeok Shin, HeeSun Bae, Gyuwon Sim, Wanmo Kang, and Il-Chul Moon. Distilling dataset into neural field. In *ICLR*, 2025.

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

Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *CVPR*, pp. 9390–9399, 2024.

Minh-Tuan Tran, Trung Le, Xuan-May Le, Thanh-Toan Do, and Dinh Phung. Enhancing dataset distillation via non-critical region refinement. In *CVPR*, pp. 10015–10024, 2025.

Haoxuan Wang, Zhenghao Zhao, Junyi Wu, Yuzhang Shang, Gaowen Liu, and Yan Yan. CaO_2: Rectifying inconsistencies in diffusion-based dataset distillation. In *ICCV*, 2025.

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

Kai Wang, Jianyang Gu, Jianyang Gu, Hansong Zhang, Daquan Zhou, Zheng Zhu, Wei Jiang, and Yang You. DiM: Distilling dataset into generative model. In *Proc. European Conference on Computer Vision*, pp. 42–59. Springer, 2024.

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

Xing Wei, Anjia Cao, Funing Yang, and Zhiheng Ma. Sparse parameterization for epitomic dataset distillation. In *NeurIPS*, pp. 50570–50596, 2023.

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

Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation. *Transact. mach. learn. res.*, 2024.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

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

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In *NeurIPS*, pp. 73582–73603, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

David Junhao Zhang, Heng Wang, Chuhui Xue, Rui Yan, Wenqing Zhang, Song Bai, and Mike Zheng Shou. Dataset condensation via generative model. *arXiv preprint arXiv:2309.07698*, 2023a.

Hansong Zhang, Shikun Li, Pengju Wang, Dan Zeng, and Shiming Ge. M3d: Dataset condensation by minimizing maximum mean discrepancy. In *AAAI*, pp. 9314–9322, 2024.

Lei Zhang, Jie Zhang, Bowen Lei, Subhabrata Mukherjee, Xiang Pan, Bo Zhao, Caiwen Ding, Yao Li, and Dongkuan Xu. Accelerating dataset distillation via model augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11950–11959, 2023b.

Xin Zhang, Ziruo Zhang, Jiawei Du, Zuozhu Liu, and Joey Tianyi Zhou. 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. pp. 12674–12685. PMLR, 2021.

Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan. *arXiv preprint arXiv:2204.07513*, 2022.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proc. Winter Conference on Applications of Computer Vision*, pp. 6514–6523, 2023.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *ICLR*, 2021.

Lin Zhao, Yushu Wu, Xinru Jiang, Jianyang Gu, Yanzhi Wang, Xiaolin Xu, Pu Zhao, and Xue Lin. Taming diffusion for dataset distillation with high representativeness. 2025a.

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*, 2025b. arXiv:2509.15472.

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

Xuyang Zhong and Chen Liu. Towards mitigating architecture overfitting in dataset distillation. *arXiv preprint arXiv:2309.04195*, 2023.

Letian Zhou, Songhua Liu, and Xinchao Wang. Coda: From text-to-image diffusion models to training-free dataset distillation. *arXiv preprint arXiv:2512.03844*, 2025.

A THE POST-HOC CONSTRUCTION WE REPLACE

Prior generative pipelines, and the ablation rungs of Sec. 4.3, close the gap between unit-norm modes and the decoder's input distribution with a non-parametric pass applied after optimization. Each sample is hard-assigned to its nearest mode in both modalities, and the samples on which the two modalities agree form the agreement cell

$$\Omega_k = \left\{ i \in \mathcal{D}_\rho \;:\; \arg\max_l \left(\mathbf{c}_l^I\right)^\top \mathbf{z}_i^I = k \;\;\wedge\;\; \arg\max_l \left(\mathbf{c}_l^T\right)^\top \mathbf{z}_i^T = k \right\}, \tag{10}$$

after which each mode is replaced by the mean raw embedding of its cell,

$$\hat{c}_k^m = \begin{cases} \frac{1}{|\Omega_k|} \sum_{i \in \Omega_k} \mathbf{e}_i^m, & \Omega_k \neq \emptyset, \\ c_k^m, & \text{otherwise}, \end{cases} \qquad m \in \{I, T\}. \tag{11}$$

Ω_k is well defined only because index k denotes the same concept in both modalities. With Eq. (8) in place this pass is unnecessary, and we retain Eq. (10) solely as a diagnostic of index agreement (Table 6) and Eq. (11) as the construction against which anchoring is compared (Tables 2 and 8). An empty cell has no agreeing sample and falls back to an uninformative center, which is one reason the pass is undesirable: at $K = 1000$ it affects 7.6% of the budget.

B HYPERPARAMETERS

The mode banks are optimized with AdamW, without weight decay, for $T = 1000$ full-batch iterations. Both banks start from one shared random initialization, rescaled to the mean norm of the real embeddings of each modality, and the base learning rate of 10^{-2} is multiplied by the mean image-embedding norm to match. Rows are not renormalized; the swap and mode-level terms act on their unit-normalized views (Sec. 3). The prediction temperature is $\tau_p = 0.1$, the Sinkhorn regularizer $\epsilon = 0.05$ with three normalization iterations, the mode-level temperature $\tau_m = 0.1$, the alignment weight $\lambda_{\text{mode}} = 1.0$ with warm-up $T_{\text{warm}} = 10$, and $\lambda_{\text{anc}} = 1.0$. Synthesis uses 100 denoising steps, classifier-free guidance scale 5 and noise level 10. No hyperparameter is tuned per dataset or per budget.

C EVALUATION PROTOCOL AND SOURCES OF VARIANCE

Checkpoint selection. The evaluation loop we inherit from Choi et al. (2026) trains a downstream model on the distilled set, scores a fixed schedule of checkpoints, and reports the best of them; in the released implementation that score is computed on the test split, so selection and reporting draw on the same data. We reuse the loop unmodified, because changing it would make our numbers incomparable with the published ones, and we repeat the comparisons under a protocol that instead selects on the validation split and evaluates the selected checkpoint once on test. The conclusions do not depend on the choice: at $K \in \{100, 500\}$, the eight paired cells for which both protocols were run, the effect of $\mathcal{L}_{\text{anc}}$ is positive in all eight under validation selection, with mean $+2.18$ on NF-ResNet50 and $+1.20$ on ViT-T/16, against $+2.23$ and $+1.12$ on the same cells under the inherited loop.

Which stages carry the variance. The intervals printed in Tables 4 and 5 are standard deviations over five downstream initializations, the last of three stochastic stages. The two earlier ones are larger: resampling the decoder seed alone moves mean recall by up to 1.16, and resampling the mode-bank seed by roughly 0.5, both exceeding the downstream standard deviation. A difference of about one recall point between two rows of the same table is therefore not resolvable from the printed interval. Where we claim such a difference we report it as a paired change at a fixed decoder seed, and verify it in a repeat that resamples the mode-bank seed and the decoder seed together; those repeats are the right-hand block of Table 9, in which the effect remains positive in eight of eight configurations with mean $+1.40$.

D PER-METRIC RETRIEVAL RESULTS

Table 1 reports mean recall, which is the quantity compared throughout the body. Tables 4 and 5 give the underlying per-metric breakdown against PDS (Choi et al., 2026), the strongest generative

Table 4: Image–text retrieval on Flickr30K. All methods distill the same training split into K pairs; the same distilled set is then evaluated with two frozen backbones. Best per column within each budget in **bold**. PDS rows at $K \in \{100, 300\}$ are the published results of Choi et al. (2026), which report no standard deviation; all other entries are means ± standard deviation over five downstream runs.

		Text-to-Image			Image-to-Text			
Method	K	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	Mean
Evaluation backbone: NF-ResNet50								
PDS (Choi et al., 2026)	100	7.90	25.80	37.30	10.20	28.20	39.00	24.73
CM^2 (ours)	100	**12.01**±0.32	**32.91**±0.16	**44.85**±0.17	**15.32**±0.48	**34.50**±0.40	**46.26**±0.94	**30.97**±0.26
PDS (Choi et al., 2026)	300	14.40	38.10	51.40	18.70	45.00	57.80	37.57
CM^2 (ours)	300	**17.79**±0.22	**42.82**±0.31	**56.16**±0.18	**22.62**±0.64	**47.74**±0.14	**59.72**±0.60	**41.14**±0.18
PDS (Choi et al., 2026)	500	16.54±0.09	42.13±0.24	55.06±0.29	22.52±0.52	48.36±0.39	60.84±0.50	40.91±0.13
CM^2 (ours)	500	**20.66**±0.32	**47.16**±0.37	**59.36**±0.13	**26.60**±0.76	**53.98**±0.64	**65.92**±0.41	**45.61**±0.15
PDS (Choi et al., 2026)	1000	18.40±0.23	44.23±0.24	57.93±0.37	25.72±0.71	54.04±0.56	66.48±0.28	44.47±0.17
CM^2 (ours)	1000	**22.60**±0.36	**49.57**±0.56	**62.38**±0.37	**30.70**±1.06	**59.80**±0.32	**71.86**±0.79	**49.48**±0.16
Evaluation backbone: ViT-T/16								
PDS (Choi et al., 2026)	100	6.80	19.20	28.50	6.60	17.50	26.90	17.58
CM^2 (ours)	100	**8.98**±0.20	**23.65**±0.22	**32.23**±0.19	**7.42**±0.32	**19.40**±0.43	**27.76**±0.65	**20.07**±0.14
PDS (Choi et al., 2026)	300	9.10	27.30	38.50	9.60	26.10	37.50	24.68
CM^2 (ours)	300	**11.94**±0.25	**30.43**±0.27	**41.86**±0.34	**12.20**±0.30	**29.36**±0.71	**39.44**±0.77	**27.54**±0.22
PDS (Choi et al., 2026)	500	11.82±0.18	29.94±0.26	41.00±0.45	12.44±0.63	31.30±0.59	41.52±0.46	28.00±0.30
CM^2 (ours)	500	**13.98**±0.26	**34.44**±0.35	**45.61**±0.55	**13.98**±0.12	**31.72**±0.80	**42.74**±0.72	**30.41**±0.20
PDS (Choi et al., 2026)	1000	13.30±0.40	34.40±0.53	45.67±0.62	13.94±0.33	34.72±0.59	46.40±0.64	31.41±0.25
CM^2 (ours)	1000	**15.29**±0.17	**36.00**±0.21	**47.20**±0.26	**15.90**±0.45	**35.80**±0.50	**47.36**±1.32	**32.92**±0.42

Table 5: Image–text retrieval on MS-COCO. All methods distill the same training split into K pairs; the same distilled set is then evaluated with two frozen backbones. Best per column within each budget in **bold**. PDS rows at $K \in \{100, 300\}$ are the published results of Choi et al. (2026), which report no standard deviation; all other entries are means ± standard deviation over five downstream runs.

		Text-to-Image			Image-to-Text			
Method	K	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	Mean
Evaluation backbone: NF-ResNet50								
PDS (Choi et al., 2026)	100	2.80	10.00	17.30	4.50	14.00	21.40	11.67
CM^2 (ours)	100	**4.13**±0.03	**12.97**±0.04	**21.09**±0.08	**5.58**±0.13	**15.15**±0.38	**23.10**±0.30	**13.67**±0.13
PDS (Choi et al., 2026)	300	5.30	17.20	27.20	7.40	20.70	30.20	18.00
CM^2 (ours)	300	**7.29**±0.09	**20.59**±0.09	**31.01**±0.15	**9.19**±0.24	**23.73**±0.48	**34.17**±0.33	**21.00**±0.17
PDS (Choi et al., 2026)	500	6.36±0.09	19.43±0.09	29.34±0.10	8.07±0.13	22.58±0.19	32.74±0.18	19.75±0.06
CM^2 (ours)	500	**8.46**±0.13	**23.30**±0.17	**34.49**±0.20	**10.72**±0.27	**27.40**±0.44	**38.41**±0.29	**23.80**±0.13
PDS (Choi et al., 2026)	1000	7.42±0.03	21.75±0.16	32.66±0.16	9.80±0.25	26.45±0.45	37.60±0.29	22.61±0.22
CM^2 (ours)	1000	**9.68**±0.11	**25.59**±0.16	**36.92**±0.13	**13.16**±0.23	**30.89**±0.26	**42.65**±0.49	**26.48**±0.14
Evaluation backbone: ViT-T/16								
PDS (Choi et al., 2026)	100	2.30	8.60	14.50	2.20	7.70	13.20	8.08
CM^2 (ours)	100	**3.78**±0.03	**11.07**±0.15	**17.93**±0.25	**3.50**±0.17	**10.39**±0.24	**16.10**±0.43	**10.46**±0.12
PDS (Choi et al., 2026)	300	4.10	13.40	21.20	3.70	12.60	19.80	12.47
CM^2 (ours)	300	**5.68**±0.06	**16.51**±0.11	**24.80**±0.12	**4.78**±0.16	**13.79**±0.27	**21.35**±0.11	**14.48**±0.05
PDS (Choi et al., 2026)	500	4.59±0.09	14.67±0.29	22.60±0.31	4.15±0.12	14.42±0.39	22.28±0.29	13.79±0.11
CM^2 (ours)	500	**6.48**±0.09	**17.77**±0.17	**26.54**±0.21	**5.54**±0.15	**15.43**±0.27	**23.51**±0.53	**15.88**±0.15
PDS (Choi et al., 2026)	1000	5.19±0.02	16.65±0.17	25.43±0.12	5.24±0.09	**17.68**±0.42	**26.54**±0.40	16.12±0.09
CM^2 (ours)	1000	**7.23**±0.07	**19.61**±0.20	**28.87**±0.30	**6.12**±0.27	17.60±0.73	26.24±0.75	**17.61**±0.30

baseline under this protocol, at every budget and for both evaluation backbones. Every entry is the mean over five independently initialized downstream runs.

Table 6: Intrinsic properties of the distilled set, Flickr30K. *Agreement*: fraction of training pairs whose nearest image mode and nearest text mode carry the same index, i.e. the population of Ω_k in Eq. (10), measured on the learned bank (C^I, C^T) for CM^2 and on the matched prototypes for PDS. *Slot coverage*: fraction of the K budget slots with a non-empty agreement set. *Mismatched pairs*: fraction of distilled pairs whose image and caption score below a CLIP ViT-B/32 cosine of 0.25, measured with a backbone independent of the ViT-L/14 the modes are built in. All three are measured on the distilled set, not on the retrieval task. The MS-COCO counterpart is Table 7.

			Budget K			
Dataset	Metric	Method	100	300	500	1000
Flickr30K	Agreement ↑	PDS	0.331	0.311	0.295	0.273
		CM^2	**0.601**	**0.547**	**0.527**	**0.506**
	Slot coverage ↑	PDS	0.990	0.973	0.956	0.924
		CM^2	**1.000**	**1.000**	**1.000**	**1.000**
	Mismatched pairs ↓	PDS	1.0%	2.3%	3.2%	7.2%
		CM^2	**0.0%**	**0.0%**	**0.2%**	**0.5%**

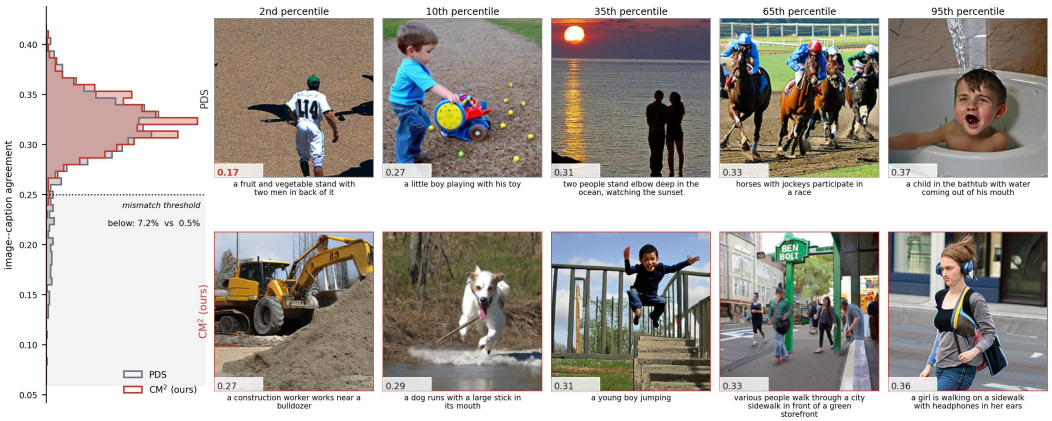

Figure 3: Distilled image–caption pairs on Flickr30K at $K = 1000$. *Left:* distribution over the K distilled pairs of the CLIP ViT-B/32 cosine between each synthesized image and the caption it is paired with; the scoring backbone is independent of the ViT-L/14 in which the modes are learned. The two methods coincide in the bulk and separate in the lower tail, where 7.2% of the pairs obtained by post-hoc matching fall below the mismatch threshold against 0.5% for CM^2. *Right:* for each method, the pair lying at a fixed set of percentiles of *its own* score distribution, annotated with that score; the comparison is therefore matched in rank rather than selected by hand. The two methods are indistinguishable above the median, and differ in what their weakest pairs look like: at the 2nd percentile, post-hoc matching pairs a baseball scene with a caption describing a fruit and vegetable stand.

E Intrinsic diagnostics of the distilled set

Table 6 reports the diagnostics of Sec. 4.4 on Flickr30K and Table 7 on MS-COCO; Fig. 3 shows the image–caption score distributions from which the mismatch rates are computed, together with pairs drawn at matched percentiles of each method's own distribution.

F Mode anchoring

Table 8 reports what anchoring changes about the geometry of the bank at a single budget. Table 9 reports the resulting change in retrieval over the full grid of datasets, budgets and evaluation backbones.

Table 7: Intrinsic properties of the distilled set, MS-COCO. Columns and metrics are those of Table 6.

Dataset	Metric	Method	Budget K 100	300	500	1000
MS-COCO	Agreement ↑	PDS	0.568	0.391	0.332	0.278
		CM^2	**0.740**	**0.584**	**0.528**	**0.466**
	Slot coverage ↑	PDS	0.970	0.970	0.956	0.954
		CM^2	**1.000**	**1.000**	**1.000**	**0.989**
	Mismatched pairs ↓	PDS	1.0%	4.0%	4.0%	4.6%
		CM^2	1.0%	**0.3%**	**0.2%**	**0.4%**

Table 8: What mode anchoring changes about the distilled bank. Flickr30K, $K = 100$. *Pair margin* is an image mode's cosine to its own text mode less its mean cosine to the other text modes, and measures how specifically the two banks are paired. *Between-mode* is the mean cosine between distinct image modes, so a lower value means the budget is spent on less mutually redundant samples. *Nearest-real* is the mean cosine from each mode to its closest real training embedding, reported for the image and text bank respectively, and indicates whether a mode lies where the decoder was estimated. The two limiting constructions each forfeit one property; anchoring optimizes neither directly and attains an intermediate value of both. Table 9 gives the change in retrieval over the full budget grid.

Construction	Pair margin ↑	Between-mode ↓	Nearest-real ↑	NF-RN50	ViT-T
Agreement-cell means (Eq. 11)	+0.161	0.800	0.90 / 0.87	28.43	17.88
Learned modes, no anchoring	+0.855	-0.007	0.38 / 0.49	29.50	19.43
Learned modes $+ \mathcal{L}_{\text{anc}}$ (ours)	+0.710	0.214	0.73 / 0.78	**30.97**	**20.07**

Table 9: Change in mean recall from $\mathcal{L}_{\text{anc}}$, over the full grid. The left block fixes the decoder seed so that decoder sampling does not enter the comparison; the right block resamples the mode-bank seed and the decoder seed together, these being the two stages that contribute more variability than downstream training does. "−" marks a configuration for which the independent repeat was not run.

		Shared bank, paired decoder		Independent bank and decoder	
Dataset	K	NF-RN50	ViT-T	NF-RN50	ViT-T
Flickr30K	100	+2.54	+2.19	+1.79	+2.13
Flickr30K	300	+0.62	+0.25	–	–
Flickr30K	500	+2.40	+0.32	+1.17	+0.07
Flickr30K	1000	+0.69	-1.32	–	–
MS-COCO	100	+1.41	+1.37	+1.56	+1.10
MS-COCO	300	+2.69	+1.19	–	–
MS-COCO	500	+2.57	+0.59	+2.99	+0.39
MS-COCO	1000	+2.68	+0.28	–	–
Positive in 15 of 16 paired comparisons (mean $+1.28$) and 8 of 8 independent repeats (mean $+1.40$).					

F.1 Optimizer-only control

Anchoring requires a bank initialized at the real embedding scale together with a learning rate scaled to match. The control below applies that parameterization without the term, and isolates the term from the parameterization it requires (Table 10).

F.2 Anchoring under a substituted decoder

Karlo (Lee et al., 2022a) consumes the same CLIP ViT-L/14 image embedding as unCLIP but takes no text condition, so the matched control is unCLIP without a caption. Table 11 removes $\mathcal{L}_{\text{anc}}$ under both decoders. Under unCLIP the unanchored bank loses 1.14 mean recall; under Karlo it loses 28.01.

Table 10: Optimizer-only control for $\mathcal{L}_{\text{anc}}$. The anchored bank is initialized at the real embedding scale and uses a learning rate scaled accordingly. Applying both without the anchoring term leaves the mode norm unconstrained: it grows to roughly three times the real scale, outside the range on which the decoder was estimated, and the resulting distilled set is unusable. The scaled parameterization therefore does not account for the improvement reported in Table 8; it is a requirement of the term rather than an independent cause.

Dataset	K	Scaled bank, no $\mathcal{L}_{\text{anc}}$	Full method	$\|C^I\|$ / real
Flickr30K	100	1.26	**30.97**	2.70
Flickr30K	500	1.09	**45.61**	2.89
MS-COCO	100	0.20	**13.67**	3.02
MS-COCO	500	0.22	**23.80**	3.02

Table 11: Effect of $\mathcal{L}_{\text{anc}}$ under two decoders. Flickr30K, $K = 100$, NF-ResNet50; neither decoder receives a caption. *Round-trip*: cosine between the CLIP embedding of a rendered image and the image mode it was rendered from. *Between-image*: mean cosine between distinct rendered images (lower means less redundant). *Pairing*: fraction of rendered images whose nearest text mode among all K is their own. *Measured with the caption.

Decoder	Bank	Mean recall	Round-trip	Between-image	Pairing
unCLIP	w/o $\mathcal{L}_{\text{anc}}$	28.93	0.374	0.511	1.000*
	CM^2	**30.07**	0.717	0.544	1.000*
Karlo	w/o $\mathcal{L}_{\text{anc}}$	1.60	0.018	0.797	0.110
	CM^2	**29.61**	0.710	0.541	1.000

The intrinsic measurements show that this is a collapse of the decoder rather than a failure of the bank. Without anchoring, the cosine between a rendered image and the mode it was rendered from falls to 0.018 under Karlo, and the mean cosine between distinct rendered images rises to 0.797: Karlo returns near-identical images regardless of its input. Only 11% of the rendered images remain nearest to their own text mode, against all of them with anchoring. unCLIP normalizes its conditioning internally and keeps the pairing intact without the term.

Radius does not explain the collapse. The anchored bank has a mean norm of 10.30, below the 18.51 of the real embeddings, and decodes correctly; the unanchored bank collapses even when rescaled to 18.55. What Karlo requires is that each mode point in a direction occupied by real embeddings, and this is the property $\mathcal{L}_{\text{anc}}$ enforces.

G Cost of the pipeline

Generative distillation has four stages: encoding the corpus once with the frozen CLIP pair, mining the modes, decoding one image per mode, and training the downstream model. Only mode mining differs between CM^2 and post-hoc matching; the other three run the same encoder, decoder, number of denoising steps and evaluation loop on a set of the same size.

Complexity. Each CM^2 iteration forms two $N \times K$ similarity matrices at $O(NKd)$, runs s Sinkhorn–Knopp normalizations at $O(sNK)$ and evaluates $\mathcal{L}_{\text{CM}}$ at $O(K^2 d)$, for $O(T(NKd + sNK + K^2 d))$ in total and $O(NK)$ memory; no matching stage follows. Post-hoc matching runs two clusterings at $O(INKd)$ each for I passes over the corpus, one further pass for the co-occurrence table, and an $O(K^3)$ Hungarian assignment (Kuhn, 1955). Both pay the NKd term; the baseline pays it once per modality and adds a term cubic in the budget.

Measurements. Table 12 reports mode-mining wall-clock time on one A100 with the released implementation of each method: CM^2 runs on the GPU, and the baseline's k-means is the CPU implementation its authors use. CM^2 is faster at every budget, by $13\times$ to $21\times$ on Flickr30K and $4.4\times$ to $5.7\times$ on MS-COCO, where the larger N weighs more on full-batch optimization than on mini-batch k-means. Neither method dominates the pipeline. Decoding takes ≈ 4.8 s per pair at 100 denoising steps, about 80 minutes at $K = 1000$, against 79 s for CM^2 on Flickr30K (under 2%) and

Table 12: Mode-mining cost, the one stage of the pipeline in which the two methods differ, measured on one A100 with the released implementation of each. N is the number of image–text rows surviving the alignment filter. Encoding, decoding and downstream training are identical for the two methods and are excluded.

		mode-mining time (s)			
Dataset	Method	$K=100$	$K=300$	$K=500$	$K=1000$
Flickr30K ($N=116{,}000$)	post-hoc matching (Choi et al., 2026)	172	498	839	1667
	CM^2 (ours)	**13**	**26**	**42**	**79**
MS-COCO ($N=453{,}398$)	post-hoc matching (Choi et al., 2026)	170	499	835	1658
	CM^2 (ours)	**39**	**95**	**157**	**292**

28 minutes for post-hoc matching. Establishing correspondence during optimization therefore adds no cost to the pipeline.