Structural Assessment for Understanding and Guiding Dataset Distillation in Discrete Token Space

Yue Cao[1], Jianyang Gu[2], Vyacheslav Kungurtsev[3], Yu Hu[1], Jozsef Hamari[4], Zheng Liu[1]*, and Mohsen Zardadi[4]*

[1] The University of British Columbia, Canada
[2] The Ohio State University, USA
[3] Czech Technical University in Prague, Czech Republic
[4] TerraSense Analytics, Canada
{yue.cao, zheng.liu}@ubc.ca, mohsen.zardadi@terrasense.ca

Abstract. Dataset distillation (DD) has proven to reduce training cost while preserving accuracy. While promising, the factors that make one distilled dataset more effective than another remain poorly understood. In this work, we investigate this question through the lens of discrete visual tokenizers. Whereas many prior DD efforts emphasize matching global data distributions, we suggest that the effectiveness depends on which semantic concepts are captured and how they are composed. Discrete visual tokenizers provide a finite vocabulary that enables direct statistical analysis of such compositional structure. Through quantitative analysis of token-level statistics, we introduce the **structural score** to measure the adequacy of token compositions. We observe that distilled datasets with balanced token composition yield higher validation performance. On the other hand, divergence from the original data does not necessarily harm performance. We further show that samples with high structural scores in the discrete token space can effectively guide diffusion-based DD. Our findings highlight the importance of token composition in dataset effectiveness, offering a principled complement to distributional similarity considerations in DD.

Keywords: Dataset Distillation · Efficient Machine Learning

1 Introduction

Dataset distillation aims to replace a large training set with a surrogate while preserving training performance [19, 43, 45]. Despite rapid empirical progress, what makes a *good* distilled dataset remains poorly understood [23, 24]. Many prior efforts enforce surrogates to match the continuous distributions of real data via feature matching [22,47,52], optimal transport [21], or gradient alignment [1,48]. However, such distributional proximity alone does not indicate semantic structures or discriminate variations informative for learning. A surrogate may be

*Corresponding authors.

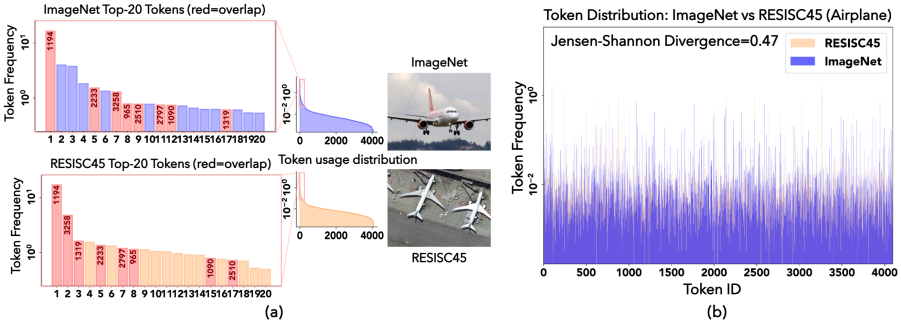

Fig. 1: The discrete token distribution of airplane images in standard and remote sensing imagery. (a) The airplanes frequently use similar top tokens across datasets. (b) At the same time, they exhibit a large divergence (JSD=0.47) in their overall token distribution due to structural differences.

close to real data in embedding space yet still collapse rare concepts, overrepresent trivial structures, or fail to span discriminative variations. In practice, the dataset most similar to the original is not always the most effective for training. This observation raises an open question: what properties of a surrogate, beyond distributional similarity, ultimately govern its training effectiveness?

We investigate this question through the lens of **discrete visual tokenizers**. Unlike high-dimensional continuous feature spaces, where semantics are heavily entangled, discrete tokenizers map each image into a sequence of indices from a finite vocabulary. This enables direct statistical analysis of how visual primitives are composed within a dataset. As illustrated in Figure 1, we use VQ-VAE [36, 39] to quantize airplane images from ImageNet [7] and remote sensing [5] into visual tokens. Although these two domains differ substantially in viewpoint, background, and style, they share a subset of high-frequency tokens that likely correspond to common airplane-related visual primitives. At the same time, their overall token distributions diverge due to the differences in context and domain-specific artifacts. This motivates us to move beyond matching global distributions and instead explicitly characterize token structure within a dataset.

To formalize this assessment, we represent each image as a discrete distribution over the shared visual vocabulary and quantify this representation through three complementary metrics: (1) Jensen–Shannon Divergence (JSD) [20] between an image's token distribution and the average distribution of its corresponding class. Lower JSD indicates samples with more representative compositional patterns. (2) Herfindahl–Hirschman Index (HHI) [15] over the token distribution, where a lower HHI reflects more balanced usage of codebook tokens. (3) Coverage rate (COV) over class-discriminative tokens identified from the real dataset, where higher coverage reflects stronger preservation of class-relevant primitives. Together, these metrics provide a holistic structural signature of each image. We then relate the aggregated statistics of surrogates generated

by diverse DD methods to their validation accuracy via regression analysis. Empirically, the combination of these three metrics, named the **structural score**, provides a solid prediction of validation accuracy. The resulting coefficients reveal that more balanced token composition primarily leads to higher accuracy. Intriguingly, merely minimizing token-level divergence from the original set does not reliably correlate with better performance.

Building on these insights, we move from diagnosis to design and explore how token-level statistics can guide surrogate generation. Unlike prior methods that rely on continuous feature centroids to guide diffusion [3, 34], we cluster data in the discrete token space and rank samples based on the structural score to acquire optimal guidance signals. Using the same steering technique as prior methods, the **Token-Guided Dataset Distillation (TGDD)** method demonstrates more effective information integration across diverse diffusion backbones and benchmarks. On ImageWoof, TGDD yields a 5.4% improvement compared to the state-of-the-art baseline [3].

To summarize, our work presents two contributions: (1) We introduce a training-free structural score that reliably assesses the quality of the distilled dataset in various DD methods and domains. (2) We develop a token-guided distillation framework that leverages this score to actively guide surrogate generation, proving the effectiveness of the discrete structural assessment.

2 Related Work

2.1 Dataset Distillation

Dataset distillation aims to synthesize a surrogate dataset such that training a model on it yields comparable performance to training on the full original dataset [19, 45]. Among previous efforts, optimization-based methods directly update surrogate images to mimic the training dynamics of the original dataset. [1,16,48] align the gradient or training trajectory, and [6,21,35,41,44,46] align feature or distribution statistics, while [38] applies distributionally robust optimization to enhance subgroup generalization. As optimization-based methods typically rely on a teacher model for supervision, their generalization across architectures is constrained. Another line of work synthesizes datasets using generative priors. For example, GLaD [2], D2M [31], and H-PD [51] leverage pretrained GANs to facilitate the optimization of the dataset. More recent works employ diffusion and visual autoregressive models as the backbone for DD, formulating synthesis as a guided generation process [11, 12, 34, 49, 50]. MGD^3 [3] and VLCP [53] improve intra-class diversity through mode guidance and vision–language prototypes. IGD [4] and CaO_2 [40] further introduce trajectory-influence and consistency-based guidance. Although representativeness, informativeness, and diversity are widely acknowledged to be important, prior work often relies on heuristic similarity scores as proxies. In this work, we aim to interpret datasets as compositions of visual concepts in certain contexts. Thereby, we turn the assessment of distilled data from distributional similarity to the structural score.

2.2 Feature-based Data Selection Criteria

To assess the usefulness of features and data samples, prior work employs a range of statistical and information-theoretic measures. The Herfindahl-Hirschman Index (HHI) [15], Shannon entropy [33], and the Gini coefficient [10] are used to characterize the concentration and diversity of feature distributions. To compare probability distributions induced by features or by selected subsets of data, Jensen–Shannon divergence and Kullback–Leibler divergence are widely used [18, 20]. For feature specificity, term frequency–inverse document frequency [32] highlights rare yet discriminative patterns. Collectively, these metrics provide quantitative tools for evaluating informativeness, coverage, and redundancy, which motivate the token-level criteria used in our framework.

3 Structural Assessment for Dataset Distillation

As motivated in §1, beyond overall proximity to the original data, we seek a more fine-grained understanding of surrogate quality. To this end, we propose explicitly assessing the structure of visual primitives by mapping continuous images into a discrete token space. Statistics over a finite vocabulary enable analyzing how these tokens are used at the dataset level. Building on this framework, we introduce the **structural score** as a robust signature of dataset quality to quantitatively evaluate the distilled surrogate.

3.1 Structural Representation in Token Space

We first employ a discrete visual tokenizer to map continuous images into a sequence of indices from a finite token vocabulary. In recent years, there have been extensive efforts in developing effective visual tokenizers [8, 27, 36]. This tokenizing step does not rely on a specific tokenizer type, and can be broadly applied. Without loss of generality, we first demonstrate an instantiation of the assessment with a multi-scale VQ-VAE [36] model.

Given a VQ-VAE codebook of size V and the encoder with L scales, each surrogate image is embedded with the encoder into discrete token maps $\{\boldsymbol{z}_i^{(1)}, \boldsymbol{z}_i^{(2)}, \ldots, \boldsymbol{z}_i^{(L)}\}$, where $\boldsymbol{z}_i^{(\ell)}$ contains indices from the codebook at the ℓ-th scale of the total L scales. For each scale, token occurrences are counted and normalized to form a probability vector over the codebook:

$$\boldsymbol{p}_i^{(\ell)} \in \mathbb{R}^V, \qquad \sum_{k=1}^{V} p_i^{(\ell)}(k) = 1,$$

where $p_i^{(\ell)}(k)$ denotes the relative frequency of the k-th token in image $\boldsymbol{x}_i$ at scale ℓ. While VQ-VAE tokens on different scales capture distinct visual information, analyzing them separately would yield fragmented scores that are difficult to reconcile. To synthesize these signals into a coherent statistical signature, we aggregate the distributions across all scales. Since the number of tokens grows

quadratically with scale, a direct summation causes an inherent density imbalance. Therefore, we fuse the token distributions using a weighted sum:

$$\boldsymbol{p}_i = \sum_{\ell=1}^{L} w_\ell \, \boldsymbol{p}_i^{(\ell)},$$

where $\boldsymbol{p}_i$ represents the fused token prior. We group the $L = 10$ scales into low (1–3), mid (4–7), and high (8–10) resolution bands. Intuitively, we assign them relative weights of $[3, 1, 0.5]$. This strategy effectively upweights the low-resolution bands to compensate for their sparsity. The validation of this weighting scheme is provided in Appendix §C.1.

Similar to words, each token in the codebook may convey a series of semantics. The semantics are not directly interpretable without being incorporated into a composition. We then examine the utilization structure of different visual tokens in addition to their presence. Specifically, we look into three properties of the distilled dataset that characterize its token-level composition:

Contextual fit. The surrogate dataset is expected to use a composition of tokens similar to that of the original dataset to reflect the corresponding contexts. Jensen–Shannon Divergence (JSD) [20] is incorporated to measure the contextual fit:

$$\mathrm{JSD}(\boldsymbol{p}_i, \boldsymbol{\mu}_c) = \tfrac{1}{2} \left(\mathrm{KL}(\boldsymbol{p}_i \,\|\, \boldsymbol{m}) + \mathrm{KL}(\boldsymbol{\mu}_c \,\|\, \boldsymbol{m}) \right),$$

where c is the class that $\boldsymbol{x}_i$ belongs to, and $\boldsymbol{\mu}_c$ is the average token distribution of the corresponding class, *i.e.*, the class centroid. $\mathrm{KL}(\cdot \,\|\, \cdot)$ calculates the KL divergence between two distributions, and $\boldsymbol{m}$ is the mixture distribution of $\boldsymbol{p}_i$ and $\boldsymbol{\mu}_c$. Lower JSD scores indicate that the samples have more typical compositional patterns of that class.

Compositional richness. The surrogate should convey meaning by combining a broad set of distinct primitives rather than reusing a few. Herfindahl-Hirschman Index (HHI) [15] is employed for this measurement:

$$\mathrm{HHI}(\boldsymbol{p}_i) = \sum_{k=1}^{V} \big(p_i(k)\big)^2,$$

where $k \in \{1, \ldots, V\}$ indexes codebook tokens, and $p_i(k)$ is the k-th entry of the fused per-image token distribution $\boldsymbol{p}_i$. Smaller HHI values correspond to more balanced token usage and thus richer visual information within the image.

Categorical presence. The usage of tokens should carry important class characteristics. We use the coverage rate of class-related tokens identified by TF-IDF [32] that appear in the original dataset:

$$\mathrm{COV}(\boldsymbol{p}_i) = \sum_{k \in \mathcal{T}_c} p_i(k),$$

where $\mathcal{T}_c \subseteq \{1, \ldots, V\}$ denotes the set of top TF-IDF tokens for class c. Higher coverage indicates that the sample uses more class-relevant visual primitives.

These three metrics reflect complementary properties of datasets. Note that the structural assessment can also be implemented with other discrete visual tokenizers, such as VQGAN [8] and BEiTv2 [27]. We then investigate the impact of these three properties on the validation performance.

3.2 Empirical Results

We collect surrogates for ImageWoof generated by a variety of dataset distillation methods, including both optimization-based (DM [47], SRe^{2}L [44]), selection-based (RDED [35]), and generative-based approaches (MGD^3 [3], Minimax [11], VAR [36], Stable Diffusion [28]). For each method except SRe^{2}L, we collect multiple surrogates from at least 5 runs to reduce randomness. We compute the token statistics of each surrogate with respect to the original ImageWoof [9] training set, where each metric is computed by averaging the scores of all individual samples within the distilled dataset. We also train a ResNet-10 [13] classifier from scratch on each surrogate to obtain their validation accuracy.

We fit a linear regression model with a Lasso regularization coefficient of 0.5 that maps the token statistics to the resulting validation accuracy. As shown in Figure 2, the predicted validation accuracy exhibits a small mean-squared error (MSE) from the ground-truth accuracy. This strong correlation indicates that the token statistics provide essential information for assessing the quality of surrogates.

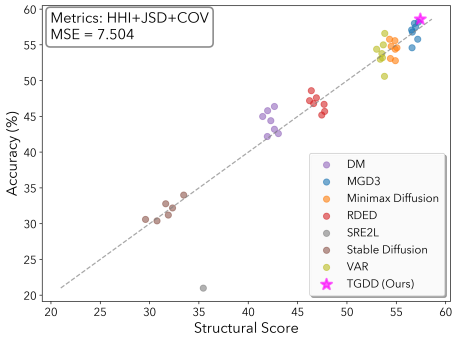

Fig. 2: Structural score versus ground-truth validation accuracy on ImageWoof across distilled datasets generated by multiple methods. We report the mean-squared error (MSE) over all data points. Note that TGDD (Ours) is only used for test but not for linear regression.

We further ablate this linear regression with different groups of metrics, and the results are shown in Figure 3. When only one metric is adopted to fit the validation accuracy, HHI illustrates a smaller MSE compared with JSD and COV. Combining two metrics yields a smaller MSE compared with each of them alone. The best explanation is produced from the combination of all three metrics, which is our proposed assessment in Figure 2. The acquired coefficients for each metric and the intercept β are:

$$w_{\text{JSD}} = 2.96, w_{\text{HHI}} = -10.99, w_{\text{COV}} = 2.07, \beta = 47.75.$$

While they cannot be evaluated in isolation, the validation performance tends to have a strong correlation with lower HHI values, *i.e.*, more balanced token compositions. We refer to this combination of statistics as the **structural score**,

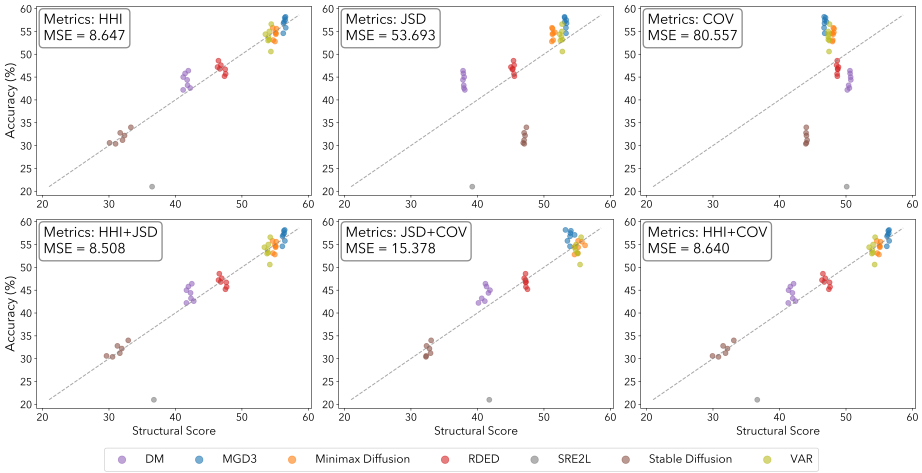

Fig. 3: The ablation study of different combinations of token statistics (HHI, JSD, COV) in predicting model accuracy via linear regression. Each subplot shows the relationship between the adopted token statistics and model accuracy across various distilled datasets, with Mean Squared Error (MSE) reported.

which provides more comprehensive assessments of distilled datasets compared with the distributional similarity alone.

We list the token statistics and validation accuracy of surrogates distilled by Stable Diffusion [28], DM [47], RDED [35], and MGD^3 [3] in Table 1 for a more detailed case study. Aligned with the coefficient, HHI has the strongest influence on the validation accuracy, while the other two metrics cannot lead to a strong correlation by themselves. Notably, DM and RDED illustrate higher JSD compared with the images generated by Stable Diffusion, but still yield higher accuracy. This result corresponds to our claim that the distributional similarity alone cannot indicate the data effectiveness.

Table 1: Detailed metric values and validation accuracy of 50-IPC surrogates for ImageWoof.

Method	JSD	HHI	COV	Accuracy
Stable Diffusion [28]	0.122	4.608×10^{-3}	0.299	31.2
DM [47]	0.186	3.334×10^{-3}	0.207	42.2
RDED [35]	0.136	2.616×10^{-3}	0.228	47.2
MGD^3 [3]	0.077	1.185×10^{-3}	0.261	55.8

Generalizability of the scoring model. With the initial metric coefficients acquired on ImageWoof, we further conduct the same linear regression with surrogates from ImageNette [9]. The obtained weights for the three metrics are: $w_{\text{JSD}} = 2.43$, $w_{\text{HHI}} = -9.59$, $w_{\text{COV}} = 1.41$. While the absolute numbers vary, the relative importance of the metrics remains consistent, with HHI suggesting

Fig. 4: Qualitative comparison of images with high and low values of JSD, COV, and HHI metrics. Each row corresponds to one metric, with images on the left exhibiting high values and those on the right showing low values.

the most significant influence on the validation performance. This result confirms the effectiveness of the derived conclusions from the structural assessments.

Table 2: Application of the proposed structural score to the EuroSAT dataset, demonstrating that our metric generalizes to visually distinct domains without retraining.

Iters	0	200	400	600	800	1000	1200	1400	1600	1800
Structural Score	-76.7	-34.2	-22.1	-15.5	-13.7	-11.1	-9.9	-4.9	-1.6	2.2
Accuracy (%)	11.5	40.6	44.9	47.1	47.8	51.7	53.0	55.8	57.2	61.1

Direct application to other domains. Although the relative relationship stands across datasets, it is infeasible to calculate the coefficients for each new application. Therefore, we demonstrate the direct generalization of the acquired coefficients on remote sensing imagery [14]. We employ Distribution Matching (DM), an optimization-based method, to perform the dataset distillation for EuroSAT. During the distillation process, we collect 10 surrogate datasets at different timesteps. The structural score and corresponding validation accuracy of each surrogate are reported in Table 2. As the images are progressively updated, the structural scores increase monotonically alongside the downstream validation accuracy. Due to the large domain gap, the predicted score does not equal the accuracy. However, it is still able to compare the effectiveness of different surrogates based on relative relationships. This trend confirms that our token-level metrics are robust indicators of dataset quality, even when applied to domains significantly different from ImageNet.

3.3 Example Visualization

We compare representative image groups with high and low metric values to better understand how each token statistics in the structural assessment reflects dataset quality, as illustrated in Figure 4. We select examples generated by MGD^3 [3] and Stable Diffusion [28], and analyze how each metric correlates with visual quality. For JSD, both sets of images are generated by MGD^3, and the JSD is computed by comparing each image's token distribution with the average token distribution of the corresponding class in the original ImageWoof dataset. Images with higher JSD values tend to exhibit unrealistic features, such as disordered object shapes or incorrect combinations of semantic elements, making them deviate from the original class distribution. For the coverage metric, lower values are often associated with images where the target object occupies a smaller region of the image or is missing entirely. This indicates a weaker representation of class-specific content. Finally, we visualize images based on their HHI values. We use samples generated by MGD^3 for the low-HHI group, which generally exhibit higher diversity. By contrast, samples generated by Stable Diffusion show similar object poses and more repetitive patterns, reflected by higher HHI values.

4 Token-Guided Dataset Distillation

The preceding analysis demonstrates that token statistics reliably indicate the quality of distilled datasets. Building on this insight, we further show that these statistics can also be incorporated to guide diffusion denoising. Following [3], we cluster each class to identify representative modes, but based on the token distribution instead of continuous features. We then rank and select anchors from each cluster using the proposed structural score, and use the selected anchors to guide a diffusion model in generating synthetic samples as in [3].

4.1 Anchor Selection

Given a real dataset $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{N_T}$, the discrete visual tokenizer is employed to embed each image to capture the underlying distribution of visual concepts $\{\boldsymbol{p}_i\}$. For numerical stability, principal component analysis (PCA) is first applied to reduce the token space to d dimensions, and the projected vectors are L2-normalized. This normalization makes Euclidean distances a monotonic transformation of cosine similarities, ensuring that clustering captures differences in token composition rather than in feature magnitude. k-means is then run on the normalized vectors to obtain K_c clusters, where K_c is set by the IPC target. This step is the same as that in [3], except in the discrete token space, where distances reflect composition differences over visual concepts. Therefore, each cluster corresponds to a pattern of token use within the class.

Token-level clustering reveals intra-class token modes but leaves many candidates per mode. Instead of using the centroid as in [3], we select a small number of

anchors from each cluster based on the token statistics identified in §3.1. Specifically, we use the linear regression model in §3.2 to predict a structure score for each sample. The only difference from the previous JSD calculation is that samples are compared with cluster centers instead of class centers. The top M images with the highest scores are selected as anchors. We set $M = 20$ anchors per cluster for IPC < 50 and $M = 10$ for higher IPC settings. We summarize the anchor selection procedure in algorithm 1 of Appendix §A.

4.2 Token-Guided Generation

The selected anchors from each cluster are used to guide synthetic data generation. We average the anchor embeddings of one cluster into one latent, representing the corresponding mode. Following the denoising procedure of [3], we guide the denoising process with the acquired modes:

$$\boldsymbol{z}_{t-1} = \boldsymbol{z}_{t-1}^{\text{base}} + \mathbb{1}_{\{t>t_{\text{stop}}\}}\, \lambda \left(\bar{\boldsymbol{z}}_{c,m} - \hat{\boldsymbol{z}}_0\right),$$

where λ denotes the mode-guidance strength, $\boldsymbol{z}_t$ is the latent variable at step t, and $\hat{\boldsymbol{z}}_0$ is the predicted clean latent. The guidance is applied only for timesteps $t > t_{\text{stop}}$ to maintain sample diversity. We demonstrate that by simply summarizing modes in the discrete token space with the help of token statistics, the acquired guidance can be more effective than that of continuous cluster centroids. Moreover, in §5.3, we demonstrate that the method can be applied to a broad range of discrete tokenizers and diffusion architectures to provide more effective guidance than continuous models.

5 Experiments

5.1 Implementation Details

For the visual tokenizer, we adopt the multi-scale VQ-VAE model developed by [36], while the diffusion generation framework employs a pre-trained DiT [26]. Following the literature [3, 26], we use 256×256 image resolution with 50 sampling steps and apply stop guidance at step 25. All experiments are conducted on a single RTX 3080 GPU. Further implementation details are provided in Appendix §B.

Datasets. We evaluate our method on multiple high-resolution benchmarks with 256×256 images to thoroughly validate its performance. The dataset used in this experiment includes ImageNette, ImageWoof [9], ImageIDC [16], ImageNet-100 [37], and ImageNet-1k [7]. ImageNette and ImageIDC each contain 10 classes, and ImageWoof comprises 10 different dog breeds, making it particularly challenging due to high inter-class similarity. For evaluation, we adopt the same metrics as [3, 11].

Table 3: Comparison of performance across state-of-the-art methods on ImageWoof. The best results are marked in **bold**.

IPC (Ratio)	Test Model	Random	Herding [42]	DiT [26]	DM [47]	MiniMax [11]	MGD^3 [3]	**TGDD**	Full
10 (0.8%)	ConvNet-6	$24.3_{\pm 1.1}$	$26.7_{\pm 0.5}$	$34.2_{\pm 1.1}$	$26.9_{\pm 1.2}$	$\mathbf{37.0}_{\pm 1.0}$	$34.7_{\pm 1.1}$	$34.8_{\pm 1.1}$	$86.4_{\pm 0.2}$
	ResNetAP-10	$29.4_{\pm 0.8}$	$32.0_{\pm 0.3}$	$34.7_{\pm 0.5}$	$30.3_{\pm 1.2}$	$39.2_{\pm 1.3}$	$40.4_{\pm 1.9}$	$\mathbf{41.2}_{\pm 2.6}$	$87.5_{\pm 0.5}$
	ResNet-18	$27.7_{\pm 0.9}$	$30.2_{\pm 1.2}$	$34.7_{\pm 0.4}$	$33.4_{\pm 0.7}$	$37.6_{\pm 0.9}$	$38.5_{\pm 2.5}$	$\mathbf{38.8}_{\pm 0.9}$	$89.3_{\pm 1.2}$
20 (1.6%)	ConvNet-6	$29.1_{\pm 0.7}$	$29.5_{\pm 0.3}$	$36.1_{\pm 0.8}$	$29.9_{\pm 1.0}$	$37.6_{\pm 0.2}$	$39.0_{\pm 3.5}$	$\mathbf{39.1}_{\pm 0.6}$	$86.4_{\pm 0.2}$
	ResNetAP-10	$32.7_{\pm 0.4}$	$34.9_{\pm 0.1}$	$41.1_{\pm 0.8}$	$35.2_{\pm 0.6}$	$45.8_{\pm 0.5}$	$43.6_{\pm 1.6}$	$\mathbf{46.3}_{\pm 0.8}$	$87.5_{\pm 0.5}$
	ResNet-18	$29.7_{\pm 0.5}$	$32.2_{\pm 0.6}$	$40.5_{\pm 0.5}$	$29.8_{\pm 1.7}$	$42.5_{\pm 0.6}$	$41.9_{\pm 2.1}$	$\mathbf{42.7}_{\pm 0.5}$	$89.3_{\pm 1.2}$
50 (3.8%)	ConvNet-6	$41.3_{\pm 0.6}$	$40.3_{\pm 0.7}$	$46.5_{\pm 0.8}$	$44.4_{\pm 1.0}$	$53.9_{\pm 0.6}$	$54.5_{\pm 1.6}$	$\mathbf{54.9}_{\pm 0.7}$	$86.4_{\pm 0.2}$
	ResNetAP-10	$47.2_{\pm 1.3}$	$49.1_{\pm 0.7}$	$49.3_{\pm 0.2}$	$47.1_{\pm 1.1}$	$56.3_{\pm 1.0}$	$56.5_{\pm 1.9}$	$\mathbf{60.3}_{\pm 0.9}$	$87.5_{\pm 0.5}$
	ResNet-18	$47.9_{\pm 1.8}$	$48.3_{\pm 1.2}$	$50.1_{\pm 0.5}$	$46.2_{\pm 0.6}$	$57.1_{\pm 0.6}$	$58.3_{\pm 1.4}$	$\mathbf{61.5}_{\pm 0.4}$	$89.3_{\pm 1.2}$
70 (5.4%)	ConvNet-6	$46.3_{\pm 0.6}$	$46.2_{\pm 0.6}$	$50.1_{\pm 1.2}$	$47.5_{\pm 0.8}$	$55.7_{\pm 0.9}$	$55.1_{\pm 2.5}$	$\mathbf{58.1}_{\pm 1.5}$	$86.4_{\pm 0.2}$
	ResNetAP-10	$50.8_{\pm 0.6}$	$53.4_{\pm 1.4}$	$53.4_{\pm 0.9}$	$51.7_{\pm 0.8}$	$58.3_{\pm 0.2}$	$60.2_{\pm 2.4}$	$\mathbf{63.4}_{\pm 1.0}$	$87.5_{\pm 0.5}$
	ResNet-18	$52.1_{\pm 1.0}$	$49.7_{\pm 0.8}$	$51.5_{\pm 1.0}$	$51.9_{\pm 0.8}$	$58.8_{\pm 0.7}$	$59.7_{\pm 2.7}$	$\mathbf{65.1}_{\pm 1.2}$	$89.3_{\pm 1.2}$
100 (7.7%)	ConvNet-6	$52.2_{\pm 0.4}$	$54.4_{\pm 1.1}$	$53.4_{\pm 0.3}$	$55.0_{\pm 1.3}$	$61.1_{\pm 0.7}$	$60.1_{\pm 1.2}$	$\mathbf{63.6}_{\pm 1.6}$	$86.4_{\pm 0.2}$
	ResNetAP-10	$59.4_{\pm 1.0}$	$61.7_{\pm 0.9}$	$58.3_{\pm 0.8}$	$56.4_{\pm 0.8}$	$64.5_{\pm 0.2}$	$66.5_{\pm 1.0}$	$\mathbf{67.3}_{\pm 0.8}$	$87.5_{\pm 0.5}$
	ResNet-18	$61.5_{\pm 1.3}$	$59.3_{\pm 0.7}$	$58.9_{\pm 1.3}$	$60.2_{\pm 1.0}$	$65.7_{\pm 0.4}$	$68.8_{\pm 0.7}$	$\mathbf{70.1}_{\pm 0.6}$	$89.3_{\pm 1.2}$

5.2 Comparison with State-of-the-art Methods

We compare our method against multiple state-of-the-art baselines using a consistent evaluation protocol. The baselines include both generative-based methods: Minimax Diffusion [11], DiT [26], and MGD^3 [3], and optimization-based methods: DM [47], RDED [35], SRe^{2}L [44] and IDC [16]. All methods are evaluated across multiple images-per-class (IPC) settings and classification architectures to validate the generalizability.

ImageWoof is selected for initial analysis due to its fine-grained classification structure, where all classes correspond to different dog breeds. The high visual similarity among these classes requires models to rely on subtle, localized features for accurate discrimination. In such scenarios, methods that focus solely on global distribution alignment often struggle to capture meaningful intra-class variation, leading to trivial or overlapping representations. TGDD addresses this challenge by leveraging discrete token statistics to identify structurally informative anchors, offering complementary perspectives on dataset composition. As shown in Table 3, TGDD consistently achieves the best classification accuracy across nearly all configurations. Furthermore, we show in Figure 2 that the token-level statistics of TGDD surrogate accurately predict its validation accuracy via the structural score.

Extending the evaluation to more challenging benchmarks, Table 4 and Table 5 report distillation results on ImageNet-100 and ImageNet-1k in two IPC settings. TGDD consistently achieves top-tier accuracy across these configurations. These results highlight the scalability of TGDD to a more diverse and complex dataset with greater class variability. In addition, the method performs consistently well across both lightweight and deeper architectures, indicating strong generalization across model architectures.

The effectiveness of TGDD is further evaluated on the ImageNette and ImageIDC subsets across varying IPC settings using a ResNet-10 backbone. These

Table 4: Comparison of dataset distillation performance on ImageNet-100. The best results are marked in **bold**.

IPC (Ratio)	Test Model	Random	Herding [42]	IDC-1 [16]	MiniMax [11]	MGD^3 [3]	**TGDD**	Full
10 (0.8%)	ConvNet-6	$17.0_{\pm 0.3}$	$17.2_{\pm 0.3}$	$24.3_{\pm 0.5}$	$22.3_{\pm 0.5}$	$23.4_{\pm 0.9}$	$\mathbf{24.8}_{\pm 0.4}$	$79.9_{\pm 0.4}$
	ResNetAP-10	$19.1_{\pm 0.4}$	$19.8_{\pm 0.3}$	$25.7_{\pm 0.1}$	$24.8_{\pm 0.2}$	$25.8_{\pm 0.5}$	$\mathbf{27.0}_{\pm 0.4}$	$80.3_{\pm 0.2}$
	ResNet-18	$17.5_{\pm 0.5}$	$16.1_{\pm 0.2}$	$25.1_{\pm 0.2}$	$22.5_{\pm 0.3}$	$23.6_{\pm 0.4}$	$\mathbf{24.7}_{\pm 0.9}$	$81.8_{\pm 0.7}$
20 (1.6%)	ConvNet-6	$24.8_{\pm 0.2}$	$24.3_{\pm 0.4}$	$28.8_{\pm 0.3}$	$29.3_{\pm 0.4}$	$30.6_{\pm 0.4}$	$\mathbf{31.8}_{\pm 0.5}$	$79.9_{\pm 0.4}$
	ResNetAP-10	$26.7_{\pm 0.5}$	$27.6_{\pm 0.1}$	$29.9_{\pm 0.2}$	$32.3_{\pm 0.1}$	$33.9_{\pm 1.1}$	$\mathbf{35.2}_{\pm 0.3}$	$80.3_{\pm 0.2}$
	ResNet-18	$25.5_{\pm 0.3}$	$24.7_{\pm 0.1}$	$30.2_{\pm 0.2}$	$31.2_{\pm 0.1}$	$32.6_{\pm 0.4}$	$\mathbf{33.4}_{\pm 0.3}$	$81.8_{\pm 0.7}$

Table 5: Comparison of dataset distillation performance on ImageNet-1k. The best results are marked in **bold**.

IPC	SRe^{2}L [44]	RDED [35]	MiniMax [11]	MGD^3 [3]	**TGDD**
10	$21.3_{\pm 0.6}$	$42.0_{\pm 0.1}$	$44.3_{\pm 0.5}$	$45.6_{\pm 0.1}$	$\mathbf{45.8}_{\pm 0.2}$
50	$46.8_{\pm 0.2}$	$56.5_{\pm 0.1}$	$58.6_{\pm 0.3}$	$60.2_{\pm 0.1}$	$\mathbf{60.3}_{\pm 0.1}$

datasets consist of coarser-grained, well-separated classes with lower inter-class ambiguity compared to fine-grained datasets like ImageWoof. Experimental results in Table 6 show that TGDD achieves the highest accuracy across all configurations. In low-IPC settings, limited data makes it harder to recover semantic diversity. TGDD maintains a clear advantage by selectively choosing structurally sufficient anchors instead of using all cluster members. This reduces noise and leads to cleaner and more representative cluster centers in the low IPC setting. In contrast, methods such as MGD^3 [3], which include all members of the cluster when computing centers, are more likely to introduce noisy samples, especially under tight data budgets. Beyond quantitative performance, we further provide qualitative comparisons and efficiency analysis in Appendix §C. These additional evaluations confirm that TGDD maintains competitiveness in both sample quality and computational overhead.

5.3 Ablation Study

Contributions of TGDD components. Table 7 presents an ablation study evaluating the contributions of three key components in the proposed guided diffusion pipeline: discrete space, PCA, and anchor selection. Each module contributes to a distinct aspect of the framework's overall effectiveness. Clustering in the discrete space provides a symbolic representation of image content, which facilitates guidance with more specific visual concepts. PCA reduces the dimensionality of token features, making subsequent computations more robust and highlighting the most informative axes of variation. Anchor selection further enhances performance by filtering out noisy or redundant samples and retaining only structurally sufficient examples to guide synthesis. To verify that this gain comes from the structural score itself, we further evaluate two control settings on ImageWoof IPC=20. Replacing score-based selection with random-M sampling within each discrete cluster drops accuracy to 43.1%, even below the

Table 6: Performance on the ImageNet subsets (Nette and IDC) across multiple images-per-class (IPC) settings. All results are obtained on a ResNetAP-10. The best results are marked in **bold**.

	IPC	Random	DiT [26]	MiniMax [11]	MGD^3 [3]	**TGDD**	Full
Nette	10	$54.2_{\pm 1.6}$	$59.1_{\pm 0.7}$	$62.0_{\pm 0.2}$	$66.4_{\pm 2.4}$	$\mathbf{67.8}_{\pm 0.6}$	$93.3_{\pm 0.1}$
	20	$63.5_{\pm 0.5}$	$64.8_{\pm 1.2}$	$66.8_{\pm 0.4}$	$71.2_{\pm 0.5}$	$\mathbf{73.6}_{\pm 0.5}$	
	50	$76.1_{\pm 1.1}$	$73.3_{\pm 0.9}$	$76.6_{\pm 0.2}$	$79.5_{\pm 1.3}$	$\mathbf{81.3}_{\pm 0.6}$	
IDC	10	$48.1_{\pm 0.8}$	$54.1_{\pm 0.4}$	$53.1_{\pm 0.2}$	$55.9_{\pm 2.1}$	$\mathbf{57.1}_{\pm 1.6}$	$92.1_{\pm 0.4}$
	20	$52.5_{\pm 0.9}$	$58.9_{\pm 0.2}$	$59.0_{\pm 0.4}$	$61.9_{\pm 0.9}$	$\mathbf{63.4}_{\pm 0.5}$	
	50	$68.1_{\pm 0.7}$	$64.3_{\pm 0.6}$	$69.6_{\pm 0.2}$	$72.1_{\pm 0.8}$	$\mathbf{73.1}_{\pm 0.8}$	

Table 7: The ablation study of the proposed token-guided dataset distillation scheme. Results are reported on the ImageWoof dataset with 20 and 50 images per class.

Discrete Space	PCA	Anchor Selection	IPC=20	IPC=50
–	–	–	$43.6_{\pm 1.6}$	$56.5_{\pm 1.9}$
✓	–	–	$44.3_{\pm 1.2}$	$57.4_{\pm 1.3}$
✓	✓	–	$45.2_{\pm 1.3}$	$58.9_{\pm 1.0}$
✓	✓	✓	$\mathbf{46.3}_{\pm 0.8}$	$\mathbf{60.3}_{\pm 0.9}$

centroid baseline of 45.2% in Table 7. Conversely, applying score-based ranking on continuous-feature clustering yields 44.5%, which improves over MGD^3 at 43.6% but remains below the full discrete pipeline at 46.3%. These results confirm that each component contributes incremental gains, that the score identifies meaningfully better anchors, and that discrete clustering and score-based selection are complementary.

Structural score component. To analyze how each structural score component in the anchor selection module contributes to dataset distillation, an ablation study is conducted, with results summarized in Table 8. The results indicate that relying on any single metric provides limited benefit, as each captures only a partial aspect of the structural score. JSD measures distributional divergence within each class, but offers limited benefit when intra-class similarity is high. COV captures class-specific compositional information, which is valuable in low-IPC settings, where preserving class identity is even more critical. HHI reflects information concentration and balance, contributing consistently across different configurations. Combining metrics in pairs leads to moderate improvements, suggesting partial complementarity. The full integration of JSD, HHI, and COV yields the best results across all the settings. These findings highlight the complementary nature of the three components and underscore the importance of their joint use for effective anchor selection in token-guided dataset distillation.

Dimensions of PCA. Table 9a investigates the impact of varying the output dimensionality of PCA within the proposed framework. Results are reported on ImageWoof, with ResNetAP-10 trained at IPC 10 and 50. The baseline model without PCA achieves 57.4% accuracy. Applying PCA leads to consistent improvements, with the best performance observed at 512 dimensions. However,

Table 8: Ablation study of structural score components in TGDD. Top-1 test accuracies (%) on ImageWoof and ImageNette under IPC = 10 and 50 are reported.

JSD	HHI	COV	ImageWoof		ImageNette	
			IPC=10	IPC=50	IPC=10	IPC=50
✓	–	–	$38.2_{\pm 2.4}$	$58.7_{\pm 0.6}$	$65.1_{\pm 1.5}$	$80.3_{\pm 0.6}$
–	✓	–	$37.8_{\pm 2.3}$	$58.7_{\pm 1.0}$	$65.9_{\pm 1.0}$	$80.5_{\pm 0.8}$
–	–	✓	$37.1_{\pm 2.6}$	$57.8_{\pm 2.1}$	$65.6_{\pm 1.7}$	$80.0_{\pm 1.2}$
✓	✓	–	$38.7_{\pm 3.4}$	$59.3_{\pm 0.5}$	$65.9_{\pm 0.6}$	$81.2_{\pm 0.5}$
✓	–	✓	$38.9_{\pm 1.4}$	$59.0_{\pm 1.1}$	$65.8_{\pm 0.3}$	$80.7_{\pm 1.1}$
–	✓	✓	$38.9_{\pm 1.7}$	$59.5_{\pm 0.6}$	$66.0_{\pm 0.7}$	$80.8_{\pm 0.5}$
✓	✓	✓	$\mathbf{41.2}_{\pm 2.6}$	$\mathbf{60.3}_{\pm 0.9}$	$\mathbf{67.8}_{\pm 0.6}$	$\mathbf{81.3}_{\pm 0.6}$

Table 9: (a) The ablation study on the effectiveness of PCA. Results are reported on ImageWoof with IPC= 10 and 50. (b) Impact of anchor set size on distillation performance. The experiment is conducted on ImageNette using ResNetAP-10 under IPC = 10 and 50.

(a)

IPC	PCA Dimension				
	None	1024	512	256	128
10	$38.7_{\pm 0.9}$	$40.1_{\pm 1.7}$	$\mathbf{40.2}_{\pm 0.9}$	$39.5_{\pm 1.1}$	$38.5_{\pm 0.9}$
50	$57.4_{\pm 1.3}$	$57.5_{\pm 1.7}$	$\mathbf{58.9}_{\pm 1.0}$	$58.7_{\pm 0.4}$	$58.5_{\pm 1.9}$

(b)

IPC	Number of anchors				
	1	5	10	20	30
10	$61.3_{\pm 2.4}$	$63.8_{\pm 1.8}$	$64.2_{\pm 0.9}$	$\mathbf{67.8}_{\pm 0.6}$	$64.0_{\pm 1.4}$
50	$78.5_{\pm 1.0}$	$79.7_{\pm 0.6}$	$\mathbf{81.3}_{\pm 0.6}$	$79.9_{\pm 0.1}$	$80.2_{\pm 1.0}$

further compressing the feature dimensionality to 256-D and 128-D results in a slight performance decline, suggesting that excessive compression can remove important semantic information. These results confirm the effectiveness of PCA as a feature compression mechanism and indicate that a 512-dimensional representation offers the best trade-off between compactness and task performance.

Effect of anchor quantity on generation. To identify the optimal number of anchors for guiding the distillation process, we study how varying the anchor set size affects generation quality. Results are shown in Table 9b. When IPC = 10, using 20 anchors yields the best performance, while at higher IPC levels, fewer anchors (*e.g.*, 10) are sufficient. This trend aligns with expectations: under limited data conditions, using more guiding anchors helps enrich the semantic coverage of the generated samples. However, when excessive anchors are used, they may introduce noise or redundancy, leading to a drop in performance.

Generalization across spaces and diffusion architectures. To validate the generalization and effectiveness of the proposed TGDD framework, we conduct a comprehensive cross-architecture evaluation. We evaluate our method using three distinct discrete visual tokenizers: VQ-VAE [36], VQGAN [8], and BEiTv2 [27], and compare them with widely used continuous embedding models including VAE encoder [17] from [30], CLIP [29], and DINOv2 [25]. These representations are paired with two diffusion architectures, specifically the transformer-based DiT [26] and the U-Net-based LDM [30]. For continuous models, we follow the

Table 10: Performance comparison across various representation spaces and diffusion architectures. Using our structural score for anchor selection, discrete tokenizers consistently outperform continuous models in guiding the generation process on ImageWoof.

Diffusion Arch.	No Guidance	Continuous Models			Discrete Tokenizers		
		VAE [17]	CLIP [29]	DINOv2 [25]	VQGAN [8]	BEiTv2 [27]	VQ-VAE [36]
DiT [26]	$49.3_{\pm 0.2}$	$56.5_{\pm 1.9}$	$57.5_{\pm 0.4}$	$58.1_{\pm 0.6}$	$60.0_{\pm 0.4}$	$59.3_{\pm 0.7}$	$60.3_{\pm 0.9}$
LDM [30]	$48.6_{\pm 1.6}$	$52.9_{\pm 2.2}$	$53.4_{\pm 1.3}$	$53.7_{\pm 0.8}$	$55.8_{\pm 0.7}$	$54.2_{\pm 0.5}$	$55.1_{\pm 0.6}$

standard practice of clustering dense features and selecting samples closest to the centroids as anchors [3]. The results in Table 10 demonstrate the robustness of our token-guided approach. While applying guidance with advanced continuous models like DINOv2 and CLIP improves upon the unguided baselines, they are consistently outperformed by our TGDD framework across all tested discrete tokenizers and generative architectures. This empirical evidence confirms that our structural score metrics actively drive performance gain and provide a robust anchor selection strategy that generalizes across diverse discrete tokenizers and offers a novel compositional perspective to the dataset distillation task.

6 Conclusion

In this work, we revisit dataset distillation through the compositional perspective of discrete visual tokens. By mapping image representations to discrete tokens via visual tokenizers, we reframe the assessment of distilled datasets as a matter of structural score rather than merely distributional similarity. We find that the statistics of visual tokens can provide a reliable predictor of validation performance. Building on this insight, we further leverage token statistics to guide diffusion denoising. The proposed Token-Guided Dataset Distillation (TGDD) achieves state-of-the-art performance across multiple benchmarks. These results suggest that discrete token analysis provides principled value for understanding and guiding dataset distillation.

Acknowledgements

This work was supported by Mitacs through the Mitacs Accelerate Program (Grant No. IT42711), by TerraSense Analytics, and in part by the U.S. National Science Foundation (OAC-2118240, HDR Institute: Imageomics).

References

1. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Dataset distillation by matching training trajectories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4750–4759 (2022)

2. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Generalizing dataset distillation via deep generative prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3739–3748 (2023)
3. Chan Santiago, J.A., Tirupattur, P., Nayak, G.K., Liu, G., Shah, M.: MGD^3: Mode-guided dataset distillation using diffusion models. In: Proceedings of the 42nd International Conference on Machine Learning (ICML) (2025)
4. Chen, M., Du, J., Huang, B., Wang, Y., Zhang, X., Wang, W.: Influence-guided diffusion for dataset distillation. In: The Thirteenth International Conference on Learning Representations (2025)
5. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)
6. Cui, X., Qin, Y., Zhou, W., Li, H., Li, H.: Optical: Leveraging optimal transport for contribution allocation in dataset distillation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15245–15254 (2025)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
9. Fastai: fastai/imagenette: A smaller subset of 10 easily classified classes from imagenet, and a little more french. `https://github.com/fastai/imagenette` (2019), last accessed on 2026-02-20
10. Gini, C.: Measurement of inequality of incomes. The economic journal **31**(121), 124–125 (1921)
11. Gu, J., Vahidian, S., Kungurtsev, V., Wang, H., Jiang, W., You, Y., Chen, Y.: Efficient dataset distillation via minimax diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15793–15803 (2024)
12. Gu, J., Wang, H., Jia, R., Vahidian, S., Kungurtsev, V., Jiang, W., Chen, Y.: Concord: Concept-informed diffusion for dataset distillation. arXiv preprint arXiv:2505.18358 (2025)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019)
15. Hirschman, A.O.: National power and the structure of foreign trade, vol. 105. Univ of California Press (1980)
16. Kim, J.H., Kim, J., Oh, S.J., Yun, S., Song, H., Jeong, J., Ha, J.W., Song, H.O.: Dataset condensation via efficient synthetic-data parameterization. In: International Conference on Machine Learning. pp. 11102–11118. PMLR (2022)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
18. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951)
19. Lei, S., Tao, D.: A comprehensive survey of dataset distillation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(1), 17–32 (2023)
20. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information theory **37**(1), 145–151 (1991)

21. Liu, H., Li, Y., Xing, T., Dalal, V., Li, L., He, J., Wang, H.: Dataset distillation via the wasserstein metric. arXiv preprint arXiv:2311.18531 (2023)
22. Liu, Y., Gu, J., Wang, K., Zhu, Z., Jiang, W., You, Y.: Dream: Efficient dataset distillation by representative matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17314–17324 (2023)
23. Lu, Y., Chen, X., Gu, J., Zhang, Y., Xuan, Q., Zhu, Z.: Dataset distillation with pre-trained models: A contrastive approach. Neurocomputing p. 132015 (2025)
24. Lu, Y., Chen, X., Zhang, Y., Gu, J., Zhang, T., Zhang, Y., Yang, X., Xuan, Q., Wang, K., You, Y.: Can pre-trained models assist in dataset distillation? arXiv preprint arXiv:2310.03295 (2023)
25. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
26. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
27. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F.: Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366 (2022)
28. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
31. Sajedi, A., Khaki, S., Liu, L.Z., Amjadian, E., Lawryshyn, Y.A., Plataniotis, K.N.: Data-to-model distillation: Data-efficient learning framework. In: European Conference on Computer Vision. pp. 438–457. Springer (2024)
32. Sammut, C., Webb, G.I. (eds.): TF–IDF, pp. 986–987. Springer US, Boston, MA (2010). https://doi.org/10.1007/978-0-387-30164-8_832
33. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)
34. Su, D., Hou, J., Gao, W., Tian, Y., Tang, B.: D^4m: Dataset distillation via disentangled diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5809–5818 (June 2024)
35. Sun, P., Shi, B., Yu, D., Lin, T.: On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9390–9399 (2024)
36. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
37. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: European conference on computer vision. pp. 776–794. Springer (2020)
38. Vahidian, S., Wang, M., Gu, J., Kungurtsev, V., Jiang, W., Chen, Y.: Group distributionally robust dataset distillation with risk minimization. In: The Thirteenth International Conference on Learning Representations (2025)
39. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)

40. Wang, H., Zhao, Z., Wu, J., Shang, Y., Liu, G., Yan, Y.: Cao$_2$: Rectifying inconsistencies in diffusion-based dataset distillation. arXiv preprint arXiv:2506.22637 (2025)
41. Wang, S., Yang, Y., Liu, Z., Sun, C., Hu, X., He, C., Zhang, L.: Dataset distillation with neural characteristic function: A minmax perspective. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25570–25580 (2025)
42. Welling, M.: Herding dynamical weights to learn. In: Proceedings of the 26th annual international conference on machine learning. pp. 1121–1128 (2009)
43. Yang, W., Zhu, Y., Deng, Z., Russakovsky, O.: What is dataset distillation learning? arXiv preprint arXiv:2406.04284 (2024)
44. Yin, Z., Xing, E., Shen, Z.: Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. Advances in Neural Information Processing Systems **36**, 73582–73603 (2023)
45. Yu, R., Liu, S., Wang, X.: Dataset distillation: A comprehensive review. IEEE transactions on pattern analysis and machine intelligence **46**(1), 150–170 (2023)
46. Zhang, H., Li, S., Lin, F., Wang, W., Qian, Z., Ge, S.: DANCE: Dual-view distribution alignment for dataset condensation. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (2024)
47. Zhao, B., Bilen, H.: Dataset condensation with distribution matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6514–6523 (2023)
48. Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. In: International Conference on Learning Representations (2021), `https://openreview.net/forum?id=mSAKhLYLSsl`, last accessed on 2026-02-20
49. Zhao, L., Jiang, X., Xiao, X., Fan, Q., Lu, L., Wang, Y., Lin, X., Camps, O., Zhao, P., Gu, J.: Hieramp: Coarse-to-fine autoregressive amplification for generative dataset distillation. arXiv preprint arXiv:2603.06932 (2026)
50. Zhao, L., Wu, Y., Jiang, X., Gu, J., Wang, Y., Xu, X., Zhao, P., Lin, X.: Taming diffusion for dataset distillation with high representativeness. In: Forty-second International Conference on Machine Learning (2025)
51. Zhong, X., Fang, H., Chen, B., Gu, X., Qiu, M., Qi, S., Xia, S.T.: Hierarchical features matter: A deep exploration of progressive parameterization method for dataset distillation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30462–30471 (2025)
52. Zhou, Y., Nezhadarya, E., Ba, J.: Dataset distillation using neural feature regression. Advances in Neural Information Processing Systems **35**, 9813–9827 (2022)
53. Zou, Y., Li, G., Su, D., Wang, Z., Yu, J., Zhang, C.: Dataset distillation via vision-language category prototype. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)