ManifoldGD: Training-Free Hierarchical Manifold Guidance for Diffusion-Based Dataset Distillation

Supplementary Material

6. Algorithmic explanation of ManifoldGD

Given an IPC centroid c_s and its static latent neighborhood $\mathcal{N}_s$ (A detailed algorithmic explanation of the formation of $\mathcal{N}_s$ can be seen in Algorithm 1.), the time-aligned manifold patch at diffusion timestep t is obtained by forward-diffusing every $z \in \mathcal{N}_s$ using the DDPM noise schedule $\varepsilon_t^{(k)} \sim \mathcal{N}(0, (1-\bar{\alpha}_t)I)$:

$$\mathcal{M}_t^{(s)} = \left\{ x_t^{(k)} = \sqrt{\bar{\alpha}_t}\, z_k + \varepsilon_t^{(k)} \;\middle|\; z_k \in \mathcal{N}_s \right\}$$

Here the centroid c_s is fixed per synthetic sample: every reverse trajectory is conditioned on the specific c_s assigned at initialization.

At each timestep we compute local geometry relative to the current latent x_t by taking its K_t nearest neighbors ($K_t = 300$ empirically works the best) *inside its own* time-aligned patch $\mathcal{M}_t^{(s)}$:

$$\mathcal{N}_t = \mathrm{KNN}(x_t;\, \mathcal{M}_t^{(s)}, K_t).$$

This produces a time-dependent covariance

$$C_t = \frac{1}{|\mathcal{N}_t|} \sum_{x \in \mathcal{N}_t} (x - \bar{x})(x - \bar{x})^\top,$$

whose top-d ($d = 3$ empirically works the best) eigenvectors define the tangent and normal projectors,

$$P_{\mathcal{T}_t} = U_{1:d} U_{1:d}^\top, \qquad P_{\mathcal{N}_t} = I - P_{\mathcal{T}_t}.$$

The manifold-corrected guidance is obtained by canceling the normal component of the mode-guidance vector,

$$g^t_{\mathrm{manifold}}(x_t; c_s) = -\, P_{\mathcal{N}_t}\, g^t_{\mathrm{mode}}(x_t; c_s),$$

thereby restricting conditional attraction to the tangent directions of the evolving class manifold and preventing off-manifold drift.

Algorithm 1 Hierarchical IPC Selection & Static Neighborhoods

Require: Latents $Z = \{z_i\}_{i=1}^N$, IPC budget K, max depth L, start level s_{start}, radius r, tangent dim d, ridge γ
Ensure: Selected centroids $\{c_s, \mathcal{N}_s\}_{s=1}^K$

Build divisive binary tree (bisecting k-means, SSE split) until depth $\le L$
Let $\mathcal{L}_d$ be leaf nodes at depth d for $d = 0, \ldots, L$
$\mathcal{S} \leftarrow \varnothing$, $k \leftarrow 0$ ▷ Stage 1: repeated coarse→fine rounds
while $k < K$ **do**
  for $d = s_{\mathrm{start}}$ **downto** 0 **do**
    if $k < K$ and $\mathcal{L}_d \neq \varnothing$ **then**
      sample $n \sim \mathcal{L}_d$ uniformly; $\mathcal{S} \leftarrow \mathcal{S} \cup \{n\}$; remove n from $\mathcal{L}_d$
      $k \leftarrow k + 1$
    end if
  end for
  $s_{\mathrm{start}} \leftarrow s_{\mathrm{start}} + 1$ ▷ expand sweep outward
  if $s_{\mathrm{start}} > L$ **then break**
  end if
end while ▷ Stage 2: deep-fill remaining quota
if $k < K$ **then**
  $\mathcal{R} \leftarrow \bigcup_{d=0}^{L} \mathcal{L}_d$; sample $(K - k)$ nodes uniformly from $\mathcal{R}$; add to $\mathcal{S}$
  $k \leftarrow K$
end if ▷ Static neighborhoods (computed offline)
for each selected node $n_s \in \mathcal{S}$ **do**
  $c_s \leftarrow$ centroid of n_s (mean or medoid)
  $\mathcal{N}_s \leftarrow \{z \in Z : \|z - c_s\|_2 \le r\}$ ▷ static latent neighborhood
end for
return $\{c_s, \mathcal{N}_s\}_{s=1}^K$

7. Baselines used for comparison

We categorize existing dataset distillation methods into three primary groups: i) traditional coreset selection, ii) training-based diffusion methods, and iii) training-free diffusion methods. This categorization provides a structured landscape against which we position ManifoldGD. i)Coreset selection: Coreset selection techniques, such as **Herding** [49] and **K-Center** [38], aim to identify a representative subset of the original data that captures the underlying data distribution. While efficient, these methods often struggle to capture the full complexity of the data distribution and exhibit limited cross-architecture generalization. ii)Training-based diffusion methods: Training-based methods fine-tune or optimize within the generative model to produce synthetic datasets. **DM** [55] utilizes a latent diffusion model to distill datasets, showing strong generalization. **Min-Max Diffusion** [15] employs a min-max objective to enforce diversity and representativeness, though it requires generator optimization. Methods like **D^{4}M** [43] and **Zou et al.** [57] also fall into this category, achieving high performance but incurring the computational cost associated with model training or fine-tuning. **IDC-1** [22] and **GLAD** [4] are also training-based methods that optimize synthetic images through gradient matching and generative modeling respectively. iii)Training-free diffusion methods:

Table 6. **Class-wise Top-1 Accuracy (%) on ImageWoof for IPC=10 and IPC=50.** MGD [37] vs ManifoldGD across three classifiers. Columns list the 10 ImageWoof classes. Class abbreviations: Aust. Terrier (Australian Terrier), Border Terrier, Samoyed, Beagle, Shih-Tzu, Rhodesian Ridgeback, Dingo, Golden Retriever, Old English Sheepdog, German Shepherd.

Setting	Method	Aust. Terr.	Border Terr.	Samoyed	Beagle	Shih-Tzu	Ridgeback	Dingo	Golden Ret.	Old Eng. Sheep.	Germ. Shep.
IPC = 10											
ConvNet	MGD [37]	24.9	40.7	16.7	27.7	22.2	31.2	28.9	31.8	40.3	32.2
	ManifoldGD	24.7	36.8	21.5	40.6	31.2	36.1	28.7	34.6	34.5	35.6
ResNet-AP	MGD [37]	35.0	44.1	21.5	46.0	24.7	32.4	29.4	37.4	38.9	45.9
	ManifoldGD	36.4	39.3	22.6	49.1	35.2	39.6	38.4	48.6	38.2	50.0
ResNet-18	MGD [37]	28.9	43.4	18.2	37.5	37.2	31.7	36.2	34.4	45.2	35.6
	ManifoldGD	34.7	43.5	22.7	44.2	34.2	41.5	35.4	46.9	41.7	36.8
IPC = 50											
ConvNet	MGD [37]	47.4	58.3	39.2	50.4	54.9	45.5	43.9	50.7	53.6	56.8
	ManifoldGD	48.7	54.4	36.6	53.1	61.8	52.1	49.6	50.9	60.4	58.3
ResNet-AP	MGD [37]	57.0	69.4	50.2	52.7	57.6	52.1	55.1	49.8	63.4	56.1
	ManifoldGD	54.0	66.0	53.5	59.4	59.1	58.0	61.1	55.5	62.7	64.9
ResNet-18	MGD [37]	50.6	66.9	44.0	47.3	61.1	54.3	55.1	54.0	63.2	60.5
	ManifoldGD	53.3	63.7	42.5	61.2	64.3	60.2	57.1	58.1	62.9	61.2

Random [32] simply selects random real images, providing a lower-bound performance benchmark. **Latent Diffusion Models (LDM)** [34] and **DiT** [31] can be directly sampled from, but without guidance, the resulting datasets may lack semantic focus. **MGD** [37] proposes a mode-guided diffusion pipeline that identifies class prototypes via clustering and guides the sampling process towards them, significantly improving upon unguided sampling.

Several concurrent works have explored enhanced guidance strategies for diffusion-based dataset distillation. [52] employs an information-theoretic framework that requires a separately trained classifier to guide the sampling process. [7] utilizes a pre-trained classifier to compute influence functions for guiding the generation trajectory. Similarly, other methods leverage text encoders for enhanced semantic conditioning [57], additional classifiers for OOD detection [13], or other external guidance mechanisms. Unlike these approaches that rely on extra components beyond the diffusion model itself, our ManifoldGD framework operates with only a single pre-trained diffusion model and a VAE for feature extraction. To maintain a fair comparison, we do not directly compete with these more complex architectures in our main experiments. Nevertheless, as seen in Section Sec. 4.1, our method achieves comparable performance to the results reported in these papers.

8. Ablation Experiments

Agglomerative vs. divisive clustering for IPC centroid selection. Fig. 9 highlights a key structural difference between agglomerative clustering and divisive (bisecting) clustering when operating in the VAE latent space. Agglomerative clustering begins with individual samples and progressively merges them ("bottom up" approach), causing early merges to be dominated by peripheral or low-density points. As a result, the resulting IPC centroids tend to lie near the boundary of the feature cloud, often reflecting outlier- or noise-driven directions. In contrast, divisive clustering recursively splits clusters, producing progressively purer, high-density partitions. This top-down refinement naturally places IPC centroids closer to the mean or high-density core of the feature manifold, yielding more stable and semantically representative prototypes. For IPC=10 on ImageNette, this difference is visually evident: divisive-levelwise centroids align with class structure, whereas agglomerative ones drift toward noisy edges. *Takeaway:* Divisive clustering provides centrally anchored, density-aligned IPC centroids, while agglomerative clustering often produces edge-biased, noisier centroids.

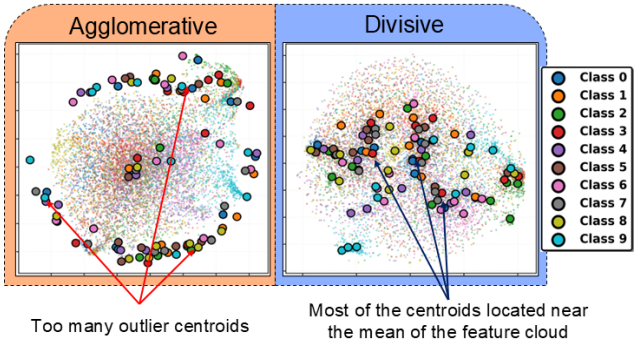


Figure 9. **VAE feature space and IPC illustration for Nette IPC=10.** Agglomerative clustering outputs IPC cluster centroids that near the edge of the feature cloud whereas Divisive clustering outputs IPC centroids that are near the mean of the feature cloud.

Level analysis for Divisive-levelwise clustering. As shown in Fig. 10, ImageNette selects a higher s_{start} compared to ImageNet-100, indicating that datasets with fewer classes benefit from starting IPC selection deeper in the hi-

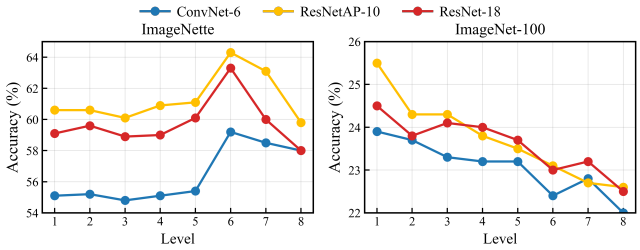


Figure 10. **Ablation experiments to find out the level of divisive-layerwise clustering.** It can be seen that for datasets with less class (ImageNette) the start level (s_{start}) is more whereas for datasets with more class (ImageNet-100) the level is less.

Table 7. **Effect of radius r on downstream classification accuracy (%).** Best result is highlighted with **bold**.

	ImageNette (IPC=10)			ImageNet-100 (IPC=10)		
Classifier	r=0.05	r=0.1	r=0.2	r=0.05	r=0.1	r=0.2
ConvNet-6	**60.5**	59.7	59.5	23.7	**24.5**	24.0
ResNetAP-10	**63.7**	63.2	63.0	26.5	**27.4**	27.2
ResNet-18	**62.3**	62.0	61.2	23.1	**23.8**	23.4

erarchy. With fewer classes, the VAE latent space is less congested, and class clusters occupy more separated, fine-grained regions, thus making leaf-level centroids reliable representatives. In contrast, ImageNet-100 exhibits substantial feature overlap due to its larger number of classes, causing leaf nodes to contain noisy, outlier-like samples that blur class boundaries. Starting closer to the root therefore avoids oversampling these unstable leaf clusters and yields more semantically coherent centroids. This validates the design of divisive-levelwise IPC selection, i.e., controlling s_{start} naturally adapts the granularity of the centroid set to the intrinsic structure of each dataset's feature manifold. *Takeaway:* The behavior of s_{start} serves as a diagnostic for class-specific feature overlap. i) Low-overlap datasets (e.g., ImageNette): deeper, fine-level picks are effective. ii) High-overlap datasets (e.g., ImageNet-100): shallower levels preserve separability and avoid noise-dominated leaves.

Effect of Neighborhood Radius r in Local Manifold Estimation. In our method, each IPC centroid c_s defines a local latent neighborhood $\mathcal{N}_s = \{z \in Z : \|z - c_s\|_2 \leq r\}$, where r is a radius in the *VAE latent space*. Intuitively, a smaller r yields a tightly localized neighborhood capturing fine-grained geometric structure, while a larger r aggregates a broader region of the latent feature cloud, mixing together points with potentially different local curvature or class-specific variations. Since the tangent space and normal projector are computed from the covariance of $\mathcal{N}_s$, the choice of r directly affects the quality of the estimated local diffusion manifold $\mathcal{M}_t^{(s)}$. For ImageNette (10-class subset), we observe that a small radius $r = 0.05$ performs best, while performance degrades for $r = 0.1$ and $r = 0.2$. Conversely, for ImageNet-100 (100 classes), the trend reverses: $r = 0.05$ tends to fail, and moderate radii $r = 0.1$ or $r = 0.2$ yield better performance. This pattern is expected because r interacts with *dataset granularity*, *class overlap*, and *latent density*. In ImageNette, classes occupy well-separated, compact regions of the VAE space, so a small radius correctly isolates coherent local geometry. In ImageNet-100, however, the latent space becomes more crowded (classes have heavier overlap and higher intrinsic variation) so $r = 0.05$ often yields too few neighbors to estimate a stable covariance, causing unreliable tangent spaces. Moderate radii provide enough samples to obtain statistically stable, noise-robust local geometry. *Takeaway:* r must match the density and geometry of the feature space. Small r excels when class clusters are compact and well-separated (e.g., ImageNette). Larger r becomes necessary in denser, higher-class settings (e.g., ImageNet-100) to obtain stable covariance and reliable tangent estimation. This reinforces that effective manifold guidance depends on capturing *locally coherent* structure, i.e., not too narrow (noisy), not too broad (geometrically inconsistent).

Table 8. **Performance comparison using DDIM.** Using ImageNette IPC=10 setting for comparison.

Method	ResNet-18	ResNet-10	ConvNet-6
MGD [37]	64.6	62.7	60.8
ManifoldGD	66.5 (+1.9)	63.6 (+0.9)	62.1 (+1.3)

Table 9. Effect of d on ConvNet-6 performance (IPC=10), transposed.

d	1	2	3	4	5
Nette	58.6	58.5	**59.5**	58.2	57.9
ImageNet-100	24.0	24.1	**24.5**	24.4	23.8

Effect of Tangent space dimension d in Local Manifold Estimation. Tab. Tab. 9 analyzes the effect of tangent subspace dimension d. We observe a consistent peak at d=3. $d < 3$ under-represents the local manifold, restricting valid tangent directions. $d > 3$ introduces noisy/off-manifold directions. *Takeaway:* Performance is stable across a reasonable range of d, indicating low sensitivity to this hyperparameter. This is expected since manifold correction is local and only requires principal variation directions (not full geometric reconstruction).

Comparison using different schedulers. From Tab. 8 we can see that ManifoldGD outperforms MGD [37] while using a DDIM noise scheduler as well. It was already seen in Tabs. 1 and 2 that ManifoldGD achieves better performance with DDPM scheduler as well. *Takeaway:* Superior performance of ManifoldGD over training-free methods is scheduler agnostic.

Time complexity/comparison. We measure inference time

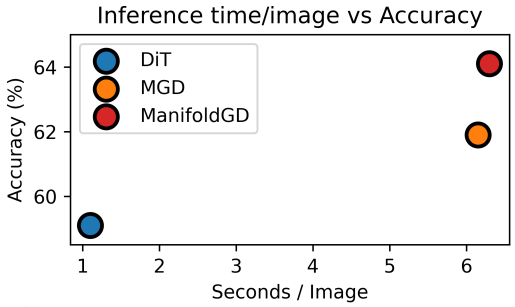


Figure 11. **Inference time vs Accuracy of MGD [37], DiT [31] and ManifoldGD for ImageNette IPC 10.** We see that ManifoldGD has a higher inference compared to MGD [37] and DiT [31] while also producing superior performance.

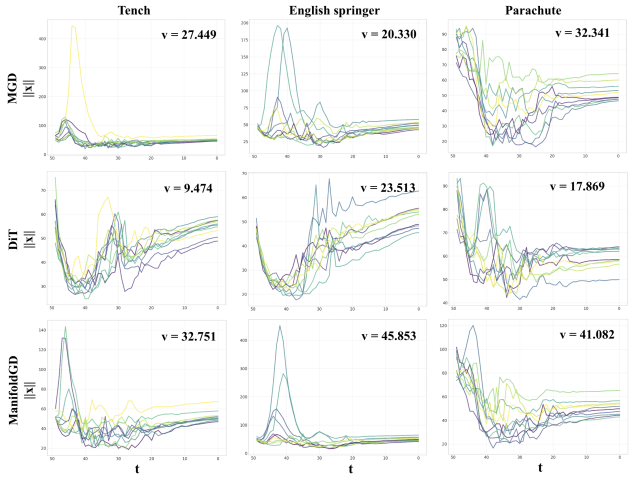


Figure 12. **Trajectory comparison of MGD [37], DiT [31] and ManifoldGD for ImageNette IPC 10.** We see that ManifoldGD has a higher final trajectory variance of the norm of the latent vector $||x||$ compared to MGD [37] and DiT [31].

for training-free methods when generating one ImageNette sample (IPC=10) on an NVIDIA A6000 (Fig. Fig. 11) ManifoldGD adds modest overhead compared to pure inference-time guidance, yet remains fully training-free, which would otherwise have incurred substantially higher compute and memory costs. In practice, ManifoldGD trades a small increase in sampling time (-$\Delta 0.15s$ vs. MGD [34], -$\Delta 5.2s$ vs. DiT [28]) for improved data quality and downstream performance (+$\Delta 2.2$ vs. MGD [34], +$\Delta 5.0$ vs. DiT [28]).

Analysis of denoising trajectories. From Tab. 8. We visualize the per-step denoising trajectories by tracking the Euclidean norm of the latent prediction $||\hat{x}||$ (the VAE-latent decoded estimate returned by the diffusion step) for multiple IPC centroid-conditioned samples (10 trajectories for IPC=10) and computing the variance of those norms across samples at each timestep. Concretely, for each conditioned sample we record $||\hat{x}(t)||$ at every denoising step, then plot the sample-wise trajectories (left) and the timestep-wise variance of those norms (right). The plotted final variance (v) refers to the variance of $||\hat{x}(50)||$ at the end of the reverse process (t=50). Fig. 12 shows these trajectories for DiT, MGD, and ManifoldGD on ImageNette (IPC=10). ManifoldGD consistently produces a larger final trajectory variance than both MGD [37] and DiT [31]. This is an expression of meaningful intra-mode variability preserved by our manifold correction. Intuitively, MGD's [37] plain Euclidean mode-pull strongly concentrates samples toward the centroid direction and therefore reduces final variability (mode-collapse within a prototype). DiT [31] being unguided often yields lower variance as well because it samples broadly from the prior but lacks targeted, class-anchored diversity. By contrast, ManifoldGD combines semantic attraction (mode guidance) with an explicit cancellation of the off-manifold normal component, i.e., we remove the normal projection and keep tangent-aligned variations. This lets samples remain close to class modes while preserving the valid degrees of freedom along the manifold (higher tangent-aligned spread), which increases the measured final variance. *Takeaway:* ManifoldGD's larger final trajectory variance reflects useful, manifold-consistent intra-class diversity rather than noise. By eliminating the off-manifold normal component while retaining tangent-aligned variations, ManifoldGD avoids the two failure modes of plain mode-pulling (i) collapse onto a tight prototype and (ii) drift off the data manifold, so samples remain class-relevant yet diverse. Because FID and downstream accuracy also improve, the increased variance serves as a compact diagnostic of the desired balance between representativeness and geometric fidelity.

Class-wise analysis. Tab. 6 presents a detailed class-wise comparison between MGD and ManifoldGD on ImageWoof for IPC = 10 and IPC = 50 across three architectures. ImageWoof is intentionally challenging for dataset distillation because all classes are dog breeds with highly overlapping visual structure (textures, fur patterns, pose similarity), making the latent space significantly more homogeneous than ImageNette or ImageNet-100. In such settings, mode-collapse and insufficient intra-class variation are more likely, and class separation depends heavily on preserving subtle geometric cues. This makes ImageWoof an especially sensitive benchmark for evaluating whether a distillation method can retain fine-grained, within-class diversity. Across IPC = 10, ManifoldGD improves over MGD on a majority of classes and architectures. For example, on ResNet-AP, ManifoldGD achieves notable gains on Beagle (+3.1), Shih-Tzu (+10.5), Ridgeback (+7.2), and Golden Retriever (+11.2), showing that manifold-aligned variation helps recover class-specific fine-grained details that Euclidean mode-pulling suppresses. Similar trends appear for ConvNet (Beagle +12.9, Samoyed +4.8) and ResNet-18 (Samoyed +4.5, Beagle +6.7, Ridgeback +9.8). These improvements emerge specifically in classes with subtle visual boundaries, where preserving tangent directions of the manifold is more critical than strict centroid at-

traction. At IPC = 50, both methods improve (as expected with increased supervision) but ManifoldGD continues to outperform MGD on most architectures. In particular, for ResNet-18, large gains appear for Beagle (+13.9), Shih-Tzu (+3.2), Ridgeback (+5.9), and Golden Retriever (+4.1). ConvNet and ResNet-AP show similar improvements, especially on texture-heavy breeds such as Shih-Tzu, Ridgeback, and Golden Retriever. These gains indicate that even with a larger IPC budget, enforcing manifold-consistent guidance preserves meaningful degrees of freedom within each breed's local geometry, preventing over-regularization and enabling better fine-grained discrimination. *Takeaway:* ManifoldGD is especially effective on fine-grained, homogeneous datasets like ImageWoof, where class distinctions rely on subtle manifold-level variations rather than coarse semantic separation. The consistent improvements across architectures and IPC budgets demonstrate that combining semantic attraction with manifold-consistent tangent preservation yields more discriminative and diverse synthetic data, ultimately improving downstream classification accuracy.

9. Qualitative Analysis

Figs. 13 and 14 shows a qualitative comparison between MGD [37], DiT [31] and ManifoldGD. GPT-4.0 was used as an unbiased evaluator, prompted to assess conceptual clarity, visual sharpness, and background separation without revealing method identities. The responses indicate that ManifoldGD generates sharper, more coherent images where the concept "fish" and "monkey" are clearly distinguished from the background, exhibiting better texture and lighting consistency. In contrast, MGD [37] samples display moderate sharpness and weaker subject–background separation. This supports that g_t^{man} promotes denoising along intrinsic manifold directions, yielding semantically faithful yet geometrically consistent generations. Furthermore, ManifoldGD achieves variations (See Fig. 14) without compromising image quality which is lacking for DiT [31] and MGD [37].

Figs. 15 and 16 compare DiT [31], MGD [37], and ManifoldGD samples on ImageNette and ImageNet-100. For ImageNette, the visual gap between methods is more pronounced: MGD often exhibits over-smoothed textures or centroid-induced collapse, while DiT occasionally drifts off-class or produces weaker semantic detail. ManifoldGD, by contrast, consistently produces sharper boundaries, richer textures, and more stable object geometry reflecting the benefit of tangent-aligned guidance and the removal of off-manifold drift. For ImageNet-100, all methods generate reasonably strong images because the underlying data distribution is more visually diverse and less constrained, making minor errors less perceptually salient. Nevertheless, ManifoldGD still improves FID (see Fig. 5) and yields slightly cleaner, more coherent samples, even though the perceptual gap appears narrower. Overall, the qualitative results mirror our quantitative trends: on simpler, low-entropy datasets (ImageNette), manifold-corrected guidance yields visibly stronger structure and detail, while on high-entropy datasets (ImageNet-100), improvements remain measurable but more subtle.

Table 10. **Comparison on ImageNet-1k using ConvNet-6.** Best results are in **bold**, second-best are underlined. * are results of methods achieved after reimplementation

IPC	DiT*[31]	MGD*[37]	ManifoldGD
1	$2.6_{\pm 0.3}$	$\underline{2.8_{\pm 0.9}}$	$\mathbf{3.1}_{\pm 0.9}$ (+0.3%)
50	$18.5_{\pm 1.3}$	$\underline{20.3_{\pm 1.1}}$	$\mathbf{21.4}_{\pm 1.5}$ (+1.1%)

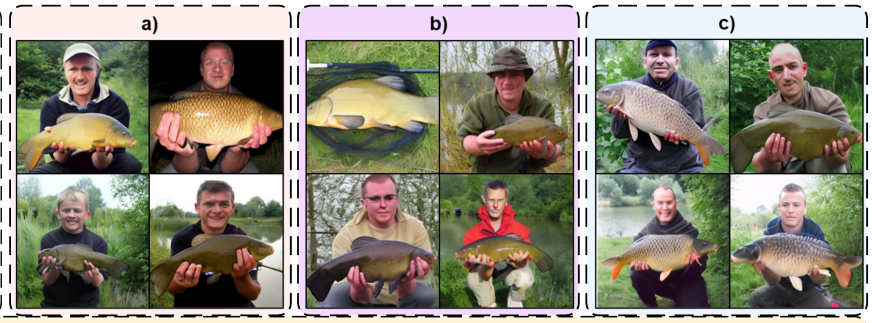
10. Results on ImageNet-1k

We additionally benchmark DiT [31], MGD [37], and ManifoldGD on the full ImageNet-1k dataset under the most challenging hard-label evaluation protocol for IPC = 1 and 50. As shown in Tab. 10, ManifoldGD consistently outperforms both baselines across both IPC regimes. These gains reflect the effectiveness of our manifold-aligned guidance in preserving class-consistent structure even when only a single distilled exemplar per class is available.

Furthermore, ImageNet-1k represents a significantly more challenging testbed compared to ImageNette or ImageNet-100. The dataset contains 1000 classes, many of which are fine-grained, visually similar, and highly imbalanced in the VAE latent space. This dramatically increases mode density and inter-class overlap, making it difficult for training-free methods to avoid collapsing onto overly generic prototypes. Methods relying solely on Euclidean mode attraction (e.g., MGD [37]) tend to overconcentrate samples around coarse class directions, while unguided diffusion sampling (DiT [31]) lacks the semantic anchoring required to generate discriminative exemplars. In contrast, ManifoldGD's tangent-aligned guidance preserves intra-class variability while suppressing off-manifold drift, leading to more faithful and well-separated prototypes across a large number of classes.

Overall, these results demonstrate that ManifoldGD is not only effective in small-scale settings but also scales robustly to the full ImageNet-1k challenge, maintaining strong sample fidelity and discriminative power even at extremely low IPC budgets.

Query: The concepts of all the images are fish. We have two methods (a), b)). Give paragraphs for a) and b) on what you see covering the following points: i)Is the concept from all the sample images understandable? ii)What features do you see of the concepts in the images and how confident are you? iii)How is the image quality, distinction from background, contrast, etc. of the images?



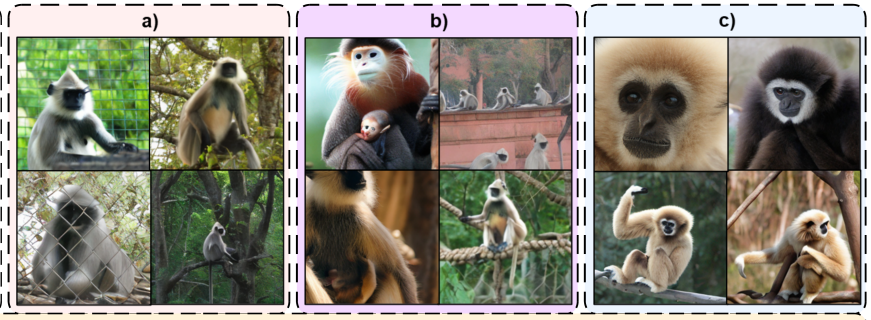
a) The concept of fish is **clearly conveyed in all samples**, with each image showing a person holding a recognizable fish. The fish shapes and features are distinct, **though lighting varies across images—one is captured at night with a flash, introducing harsh contrast**. Backgrounds are somewhat busy but do not obscure the main subject. Overall, image quality is acceptable, with moderate variation in brightness and contrast.

b) The fish concept is evident, though **slightly less consistent in clarity due to varying angles and lighting**. Some images **blend more with the background**, especially where the fish coloration matches natural surroundings. The overall **contrast and sharpness are moderate**, and while the **concept remains clear**, visual distinction could be improved.

c) All images clearly **represent the fish concept** with **well-defined shapes, textures, and patterns**. The **lighting and color balance are strong, producing sharp, high-quality visuals**. **The fish stand out distinctly from the background**, and overall presentation is the **most coherent and visually appealing among the three sets**.

Figure 13. **Qualitative comparison of MGD [37], DiT [31] and ManifoldGD.** We provide the Query to instruct GPT 4.o to evaluate the images (a) → DiT, b) → MGD, c) → ManifoldGD). We see from the answers of GPT 4.o that ManifoldGD achieves better images that discriminate the concept (fish) from other information (background) better. Green highlighted text indicates positive attributes whereas Red highlighted text indicates negative attributes.

Query: The concepts of all the images are fish. We have two methods (a), b)). Give paragraphs for a) and b) on what you see covering the following points: i)Is the concept from all the sample images understandable? ii)What features do you see of the concepts in the images and how confident are you? iii)How is the image quality, distinction from background, contrast, etc. of the images?



a) The primate concept is **clearly recognizable across all four samples**, with each image depicting langur-like monkeys in natural or zoo environments. The animals' long limbs, light fur, and dark facial features are consistently visible, making the **concept easy to interpret**. Image quality varies: two images show fence patterns that **partially obscure the subject**, and **lighting is uneven across the set**. **Backgrounds are somewhat cluttered with foliage or mesh, but the primates remain distinguishable**.

b) **The concept is understandable** throughout, showing a mix of monkey species with characteristic facial coloration, body posture, and group behavior. Some images focus closely on facial details or limbs, while others capture full-body shots or groups, offering **varied but coherent representations of primates**. The image quality is generally good, though **backgrounds can be visually busy**, and one image has a **softer, lower-contrast appearance**. Overall subject–background separation is adequate.

c) **All images clearly represent** gibbon-like primates with **distinctive light or dark fur, expressive faces, and long arms typical of arboreal species**. The subjects are **well-captured with clean compositions, making features like facial structure and body posture easy to see**. **Image quality is strong across the set, with good lighting, sharpness, and natural contrast**. The **primates stand out clearly against the backgrounds, offering the most visually polished set among the three**.

Figure 14. **Qualitative comparison of MGD [37], DiT [31] and ManifoldGD.** We provide the Query to instruct GPT 4.o to evaluate the images (a) → DiT, b) → MGD, c) → ManifoldGD). We see from the answers of GPT 4.o that ManifoldGD achieves better images that discriminate the concept (monkey) from other information (background) better. Green highlighted text indicates positive attributes whereas Red highlighted text indicates negative attributes.

Figure 15. **Qualitative comparison of MGD [37], DiT [31] and ManifoldGD for ImageNette.** ManifoldGD produces sharper, more semantically aligned and structurally coherent samples compared to MGD, while also avoiding the occasional blurring or texture flattening observed in DiT. Differences are especially prominent in edges, fine textures (fur, feathers), and object–background boundaries.

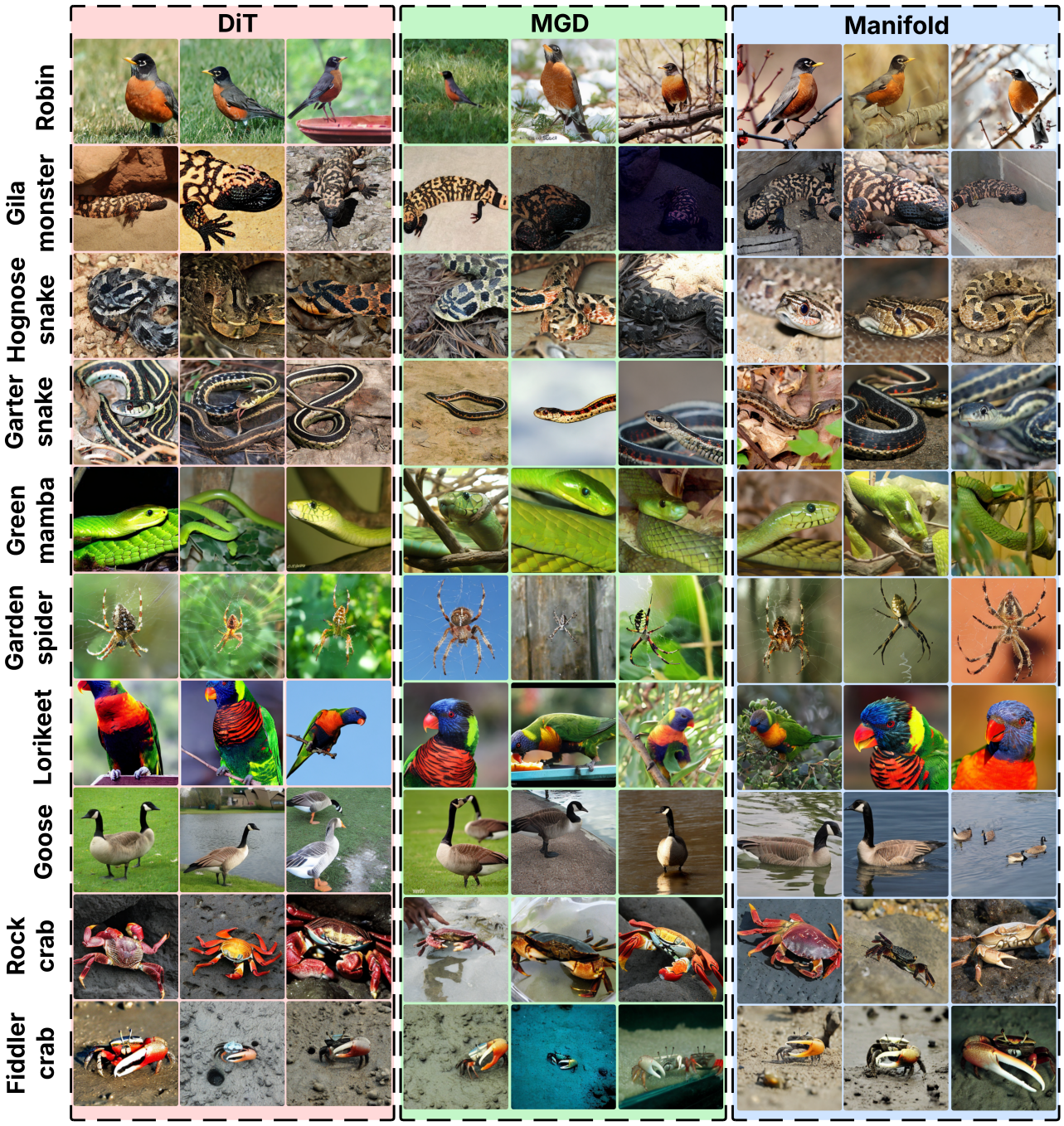


Figure 16. **Qualitative comparison of MGD [37], DiT [31] and ManifoldGD for ImageNet-100.** All three training-free methods generate visually plausible samples for this larger and more diverse dataset, though ManifoldGD still achieves cleaner geometry, fewer artifacts, and improved object–background separation. Differences are more subtle than in ImageNette due to the higher visual entropy and class variability of ImageNet-100.