A Appendix

A.1 Related Work (Recap and Expansion)

A.1.1 Out-of-distribution (OOD) Detection

Many early OOD detection methods rely on post-hoc processing of classifier outputs, such as logits [19, 18, 33, 36], model features [32, 40, 43, 50], or gradients [3, 26]. Other approaches directly modify classifier training by employing adjusted loss functions such as energy-based objectives [35], confidence-based rejection [9, 25, 31], or auxiliary self-supervised tasks [1, 54]. Recently, Outlier Exposure (OE) has significantly improved performance by leveraging large auxiliary datasets as pseudo-OOD samples, shaping conservative and robust decision boundaries; however, its effectiveness is limited by the availability and relevance of external datasets, thus motivating synthesized outlier methods. Approaches like VOS [14] and NPOS [51] generate synthetic outliers in feature space, while Dream-OOD [13] introduces a two-stage diffusion-based method that aligns and perturbs embeddings in Stable Diffusion's text-conditional space to produce pixel-level OOD images. Subsequent works such as BOOD [34] and NCIS [11] adopt this framework but employ adversarial perturbations and learned nonlinear invariants, respectively. In contrast, our method eliminates embedding alignment by directly guiding the diffusion sampling trajectory toward OOD regions via gradient-based pixel-level adjustments. This single-stage strategy significantly improves efficiency, robustness, and the diversity of synthesized OOD images.

A.1.2 Guided Sampling with Diffusion Models

Diffusion models generate samples through iterative denoising, effectively estimating data gradients [46, 44], and naturally support post-hoc conditioning via optimization signals. Classifier-based guidance [47, 10] uses a noise-conditional classifier to provide gradients during sampling, while classifier-free guidance (CFG) [23] avoids extra classifiers by interpolating between conditional and unconditional predictions. However, both approaches require task-specific training, which limits scalability. In contrast, training-free guidance (TFG) guides sampling using differentiable target functions—such as classifiers or loss—without additional training. While several TFG methods have been proposed [17, 45, 6, 2, 61], most remain task-specific and lack unified theoretical foundations. Recently, Ye et al. [59] formalized a general TFG framework to unify such strategies. Building on this, our work extends training-free guidance to OOD samples generation by defining a target function that captures OOD signals via pre-trained classifiers and guiding the diffusion process to synthesize informative outliers.

A.2 Theoretical Analysis

A.2.1 Integration of GOOD into the TFG Framework

The key contribution of the paper *"TFG: Unified Training-Free Guidance for Diffusion Models"* [59] is the proposal of a unified and extensible algorithmic framework that enables conditional generation from unconditional diffusion models using off-the-shelf, differentiable target functions $f(x)$—without requiring any retraining. TFG generalizes and unifies several prior training-free sampling methods under a common formalism.

At each denoising step t, the diffusion model estimates the clean signal as:

$$x_{0|t} = \frac{x_t - \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}.$$

To guide sampling, TFG applies gradients of a smoothed objective function:

$$\tilde{f}(x) = \mathbb{E}_{\delta \sim \mathcal{N}(0,I)} \left[f\left(x + \bar{\gamma}\sqrt{1-\bar{\alpha}_t} \cdot \delta\right) \right],$$

which stabilizes gradients by smoothing the predictor landscape via Gaussian convolution.

Guidance is introduced via two complementary strategies:

- **Variance guidance**, a second-order signal, uses gradients with respect to the noisy sample x_t:

$$\Delta_t = \rho_t \cdot \nabla_{x_t} \log \tilde{f}(x_{0|t}),$$

  leveraging the covariance between x_t and $x_{0|t}$ to adjust the sampling trajectory.

- **Mean guidance**, a first-order signal, applies gradients with respect to the predicted clean sample $x_{0|t}$:

$$\Delta_0 = \sum_{i=1}^{N_{\text{iter}}} \mu_t \cdot \nabla_{x_{0|t}} \log \tilde{f}(x_{0|t} + \Delta_0),$$

  steering samples directly in data space toward higher scoring regions.

Moreover, TFG allows **Recurrence**—repeated application of guidance and denoising—to further refine sampling over N_{recur} cycles, improving convergence and robustness to local optima.

This design space is formalized as:

$$\mathcal{H}_{\text{TFG}} = \{(N_{\text{recur}}, N_{\text{iter}}, \bar{\gamma}, \rho, \mu)\},$$

which subsumes prior works such as DPS [6], LGD [45], MPGD [17], FreeDoM [61], and UGD [2] as special cases, enabling unified theoretical analysis and practical design.

Extension to Our Method. *Our method GOOD (Guided OOD sampling) fits naturally into the TFG framework by instantiating its guidance procedure with two novel, task-driven target predictors $f(x)$ tailored for out-of-distribution sample generation.*

Target Predictor: In line with the TFG formulation, we define a differentiable target function $f_c(x) : \mathcal{X} \to \mathbb{R}_+ \cup \{0\}$ that evaluates how well a sample x aligns with an OOD objective conditioned on c. Following the conditional sampling framework:

$$p_0(x \mid c) = \frac{p_0(x) f_c(x)}{\int_{\tilde{x}} p_0(\tilde{x}) f_c(\tilde{x}) d\tilde{x}},$$

we seek to guide diffusion trajectories toward low-likelihood, feature-sparse regions representing diverse and informative OOD samples.

Inspired by post-hoc OOD scoring methods [19, 18, 33, 36, 50], we propose two concrete instantiations of $f_c(x)$:

- **Image-Level Predictor (GOOD$_{\text{img}}$)**: Based on the *free energy* [35] of a pretrained classifier f, we define

$$G_{ood}^{img}(x) := exp(E(x; f)) = \sum_{k=1}^{C} \exp(f_k(x)),$$

  which approximates the negative log-likelihood of x under the classifier's predictive distribution. Its gradient $\nabla_x \mathcal{E}(x; f)$ pushes samples from high-density ID regions to low-density OOD regions in pixel space.
- **Feature-Level Predictor (GOOD$_{\text{feat}}$)**: To capture structural sparsity, we compute the k-nearest-neighbor distance in the normalized feature space:

$$G_{ood}^{feat}(x) := D_k(x; f) = \|z(x) - z^{(k)}\|_2, \quad z(x) = \frac{f^{(1:L)}(x)}{\|f^{(1:L)}(x)\|_2},$$

  where $z^{(k)}$ is the k-th nearest feature vector from an in-distribution memory bank. Gradients of this score encourage exploration of semantically novel and underrepresented directions in feature space.

These target predictors define complementary guidance signals and are differentiable, enabling their seamless integration into the TFG framework.

Instantiation within TFG. We instantiate GOOD using the following configurations in TFG:

- **(a) Implicit Dynamic:** Gaussian smoothing is applied to f via

$$\tilde{G}_{ood}(x) = \mathbb{E}_{\delta \sim \mathcal{N}(0,I)}[G_{ood}(x + \bar{\gamma}\sqrt{1 - \bar{\alpha}_t}\delta)]$$

  to ensure smooth and stable gradient fields;
- **(b) Variance Guidance:** Guidance on the noisy sample x_t via second-order signal $\nabla_{x_t} \log \tilde{G}_{ood}(x_{0|t})$;

Figure 7: Energy distributions of 500 OOD samples generated under each $(\bar{\rho}, \bar{\mu})$ configuration on ImageNet ($\bar{\gamma} = 0.1$). Each subplot shows the histogram of negative energy scores (i.e., $\log \sum_k \exp f_k(x)$) under a fixed pair of variance guidance strength $\bar{\rho}$ (columns) and mean guidance strength $\bar{\mu}$ (rows). Higher $\bar{\rho}$ and $\bar{\mu}$ tend to produce lower-energy (more outlying) samples, while lower values yield samples closer to the in-distribution manifold. This grid highlights the tradeoff between sample extremeness and diversity, supporting our use of balanced sampling across parameter combinations.

- **(c) Mean Guidance:** Guidance on the denoised estimate $x_{0|t}$ via first-order optimization $\nabla_{x_{0|t}} \log \tilde{G}_{ood}(x_{0|t})$.

For computational efficiency, we omit recurrence and set $N_{\text{recur}} = N_{\text{iter}} = 1$, since outlier training typically requires generating a large number of samples, and iterative guidance would significantly slow down the sampling process.

By plugging our OOD-specific target predictors into TFG's modular framework, GOOD enables principled, training-free guidance of diffusion sampling toward low-density and classifier-sensitive regions. This design allows us to generate boundary-adjacent OOD samples that are diverse, semantically aligned, and highly effective for outlier exposure training.

A.2.2 Hyperparameter Selection

The TFG framework introduces a structured design space:

$$\mathcal{H}_{\text{TFG}} = \{(N_{\text{recur}}, N_{\text{iter}}, \bar{\gamma}, \rho, \mu)\},$$

where each hyperparameter governs a specific guidance behavior. We now detail our choices for each component in the context of OOD sampling, balancing generation quality, diversity, and efficiency.

Recurrence (N_{recur}) and Mean Iteration (N_{iter}). Recurrence amplifies the guidance signal by repeating the guidance–denoise–reinject cycle, while N_{iter} controls the number of inner steps used to refine the clean estimate $x_{0|t}$. Although these operations can improve sample quality, they are computationally expensive. Since outlier exposure training typically requires synthesizing thousands of OOD samples, we prioritize efficiency and fix both to one step: $N_{\text{recur}} = N_{\text{iter}} = 1$.

Smoothing Strength ($\bar{\gamma}$). The Gaussian smoothing in implicit dynamics stabilizes the gradient field of $f(x)$, especially in low-density regions, by averaging local variations. We select $\bar{\gamma}$ based on

resolution: 0.1 for low-resolution CIFAR and 0.001 for high-resolution ImageNet. These values were determined via a lightweight beam search focused on the energy distribution of generated images.

Variance Guidance Strength (ρ). The coefficient ρ_t scales the gradient $\nabla_{x_t} \log \tilde{f}(x_{0|t})$ applied in the noisy space. As variance guidance encodes second-order information, it is particularly useful during early diffusion steps when x_t is heavily corrupted and carries limited semantic content.

Following the design in the original TFG paper [59], we adopt an increasing time-dependent schedule:

$$\rho_t = \bar{\rho} \cdot \frac{\alpha_t}{\sum_{s=1}^{T} \alpha_s},$$

where $\bar{\rho}$ controls the overall strength of variance guidance. This increasing structure has been empirically validated across multiple tasks—including Gaussian deblurring, super-resolution, label guidance, and style transfer—demonstrating its effectiveness in gradually shifting the focus from low-level pixel alignment to higher-level structural refinement as the denoising process unfolds.

Mean Guidance Strength (μ). Similarly, mean guidance operates on the denoised estimate $x_{0|t}$ and becomes more impactful in later steps, when the sample becomes semantically clearer and closer to the data manifold. We therefore adopt the same increasing schedule:

$$\mu_t = \bar{\mu} \cdot \frac{\alpha_t}{\sum_{s=1}^{T} \alpha_s}.$$

This schedule ensures that mean guidance remains weak during early steps, where gradients are noisy, and grows stronger as the sample becomes more refined.

Sampling with Diverse Parameter Combinations Rather than Single Optimal. Instead of searching for a single optimal $(\bar{\rho}, \bar{\mu})$ pair, we adopt a balanced sampling strategy using a fixed grid of parameter combinations. Specifically, we sample from the Cartesian product $\bar{\rho}, \bar{\mu} \in \{0.1, 0.5, 1.0, 2.0, 5.0\}$, resulting in 25 configurations. Each configuration yields OOD samples with different anomaly intensities and semantic characteristics. Figure 7 visualizes the negative energy distributions of 500 samples per configuration on ImageNet. As $\bar{\rho}$ and $\bar{\mu}$ increase (from left to right and top to bottom), the generated samples shift toward lower energy regions, indicative of stronger outlier characteristics. Conversely, low guidance strengths produce samples closer to the in-distribution manifold. This tradeoff between semantic plausibility and anomaly severity is crucial: overly strong guidance may yield unrealistic or trivial outliers, while weak guidance may fail to leave the data manifold meaningfully.

By covering the full grid, our approach generates a continuum of near-OOD samples spanning subtle to extreme shifts. This diversity enhances outlier exposure training by reducing overfitting to a narrow anomaly profile and promoting robustness to real-world distributional shifts.

A.3 Implementation Details and Datasets

A.3.1 Details of Balanced Sampling

To support diverse and robust outlier exposure, we generate a wide spectrum of OOD samples using a balanced grid-based strategy described earlier. Specifically, we sample a large number of images across different guidance strengths, covering both image-level and feature-level scores. The sampling setup is consistent across both ImageNet-100 and CIFAR-100.

Image-level guidance (GOOD$_{img}$). For image-based guidance, we adopt a full Cartesian product of guidance strengths with $\bar{\rho}, \bar{\mu} \in \{0.1, 0.5, 1.0, 2.0, 5.0\}$, resulting in 25 unique parameter pairs. For each parameter configuration, we sample 5 images per class across 100 classes, yielding: $25 \times 5 \times 100 = \mathbf{12{,}500}$ samples.

Feature-level guidance (GOOD$_{feat}$). For feature-based guidance, we follow the kNN-based sparsity scoring method proposed in [50]. Unlike energy scores, kNN distances are not normalized to $[0, 1]$ due to their small dynamic range; instead, we apply a fixed scaling factor of 5 before computing gradients. For simplicity, we set $\bar{\rho} = \bar{\mu}$ and choose 7 representative values:

Table 4: Comparison of computational stages required by existing outlier exposure methods. Traditional approaches involve either expensive manual collection or multi-stage generation pipelines, while GOOD eliminates most stages by directly guiding diffusion sampling.

Method	Manual Data Collection	ID Embedding Alignment	OOD Embedding Sampling	OOD Sample Generation	Training and Detection
WOODS [28]	✓				✓
SAL [12]	✓				✓
Dream-OOD [13]		✓	✓	✓	✓
FodFoM [5]		✓	✓	✓	✓
NCIS [11]		✓	✓	✓	✓
BOOD [34]		✓	✓	✓	✓
GOOD (Ours)				✓	✓

$\{0.2, 0.4, 1.0, 1.5, 2.0, 3.0, 4.0\}$. For each configuration, we generate 15 images per class, resulting in: $7 \times 15 \times 100 = \mathbf{10{,}500}$ samples.

This balanced sampling strategy ensures that our generated OOD dataset covers a rich diversity of anomaly levels—ranging from near-distribution hard negatives to semantically abstract outliers—thereby improving generalization during training and evaluation.

A.3.2 Computational Cost Comparison

Outlier exposure methods generally fall into two categories: data-collection-based and generation-based. The former, such as WOODS [28] and SAL [12], rely on manually curated external datasets, which are costly to collect and often domain-dependent. The latter—including Dream-OOD [13], FodFoM [5], NCIS [11], and BOOD [34]—adopt multi-stage generation pipelines that incur substantial computational overhead. These generation-based approaches typically involve: (1) constructing or aligning latent feature spaces from ID embeddings, (2) synthesizing OOD features within these spaces, (3) decoding them into images using pretrained or fine-tuned generative models (e.g., diffusion or GANs), and (4) training a detection model on the resulting samples. As summarized in Table 4, all stages incur non-trivial computational and engineering costs.

In particular, Dream-OOD reports a total compute time exceeding 16 hours on ImageNet-100, including 8.2h for latent space construction, 10.1h for diffusion-based image generation, and 8.5h for training. BOOD reports similar costs, with 0.62h for building the latent space, 0.1h for OOD feature synthesis, 7.5h for image generation, and 8.5h for regularized training. NCIS requires up to 13h for ID embedding extraction alone (on 8×A100 GPUs), plus additional time for training the cVPN projection network.

In contrast, our method GOOD simplifies the pipeline by directly guiding the diffusion sampling process with gradients from a pretrained classifier. As shown in the shaded row of Table 4, GOOD bypasses most stages: it does not require latent space construction, embedding alignment, or feature synthesis. The only computational components are the forward/backward passes of the classifier during sampling, and the final outlier exposure training. In practice, each 512×512 image generated with OOD guidance (based on a 224×224 classifier input) takes approximately $5.7\times$ longer than unconditional sampling. However, unlike prior methods that typically require generating over 100,000 OOD samples, our method only needs about one-fifth as many—making the total computational cost comparable or even lower. As a result, the overall computational cost remains manageable. This efficient design supports scalable, high-quality OOD sample generation and enables seamless integration into large-scale training pipelines without retraining or architectural changes.

A.3.3 Distribution Analysis across OOD Datasets

For ImageNet-100, we follow [27] and evaluate OOD detection performance on four standard benchmarks: Textures [7], iNaturalist [52], Places [62], and SUN [55]. For CIFAR-100, we use five common OOD datasets: SVHN [38], Places365 [62], LSUN-R [60], ISUN [56], and Textures [7].

To understand how different scoring functions separate ID and OOD samples, Figure 8 presents histograms of normalized detection scores across the four ImageNet-100 OOD datasets. Each column corresponds to a different OOD dataset, and each row shows a different score type: energy-based score (top), k-nearest neighbor (KNN) distance in feature space (middle), and our proposed unified score (bottom). We observe that the energy score distributions (top row), which reflect low image likelihood, tend to have broader spread and more overlap with ID validation scores—especially for

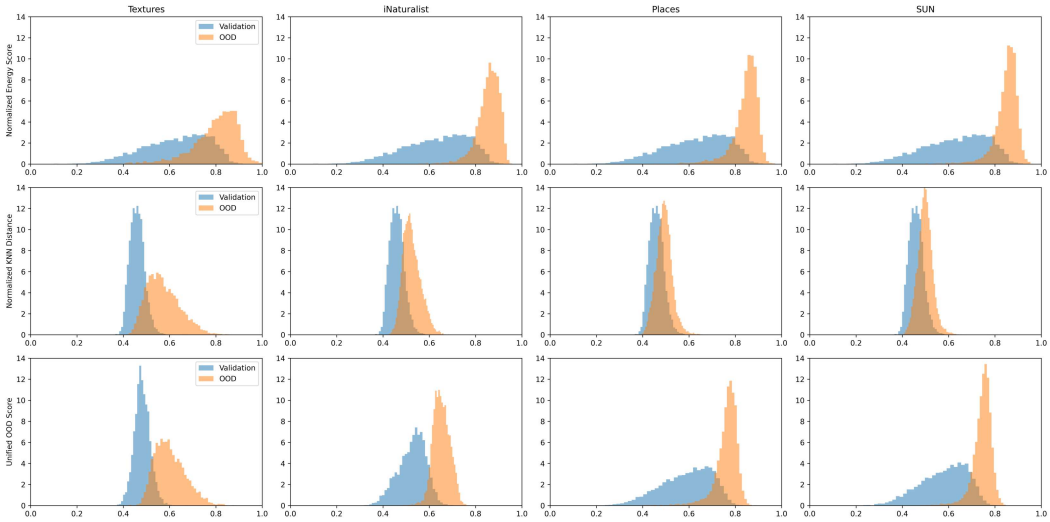

Figure 8: Distributions of OOD detection scores across four OOD datasets (Textures, iNaturalist, Places, and SUN) and three score types: energy-based score (top row), k-nearest neighbor (KNN) distance (middle row), and a unified OOD score (bottom row) that adaptively fuses the two based on KL divergence. Each subplot compares the score distributions for the ID validation set (Imagenet-100) and the corresponding OOD dataset. All scores are normalized to the range [0, 1].

iNaturalist, Places, and SUN. In contrast, the KNN-based feature distance (middle row) produces sharply peaked and better-separated distributions for certain datasets, particularly Textures, indicating that these OOD samples are more distinguishable in feature space than in pixel space.

Interestingly, the separability patterns differ by dataset: Textures is clearly more separable via feature distance, while the other three datasets (iNaturalist, Places, and SUN) are more distinguishable by energy score. To leverage the strengths of both signals, we compute a KL divergence between ID and OOD score distributions for each type and use it to adaptively weight the two, producing a unified OOD score (bottom row). As shown in the last row of Figure 8, this fused score consistently improves separation across all four datasets by emphasizing the more discriminative signal for each case.

This analysis supports the use of our unified score as a reliable, adaptive OOD detection signal that balances image-space and feature-space cues.

A.4 Additional Results and Case Studies

A.4.1 Additional Results on CIFAR

To verify that our guidance directions reflect semantically meaningful trajectories relative to the classifier, we conduct two controlled experiments on CIFAR datasets, each designed to probe the behavior of GOOD in feature space.

(a) OOD sample diversity via Stable Diffusion. We first generate OOD samples conditioned on the CIFAR-100 class "mouse" using a high-resolution Stable Diffusion model (512/128). Figure 9(a) showcases a diverse range of mouse-related outputs. The samples vary from realistic mice to abstract renderings, cartoons, and even symbolic representations—demonstrating how GOOD naturally induces both semantic preservation and controlled abstraction. This supports its ability to generate meaningful OOD variants within the scope of a class concept.

(b) Semantic alignment in latent space via t-SNE. Next, we study the semantic effect of guidance using a 32×32 DDPM model trained on CIFAR-10. We partition the 10 classes into two halves: five seen classes are used to train a classifier, while the other five are held out. Using this setup, we perform three types of sampling: (1) unconditional generation, (2) ID-guided sampling (i.e., maximizing

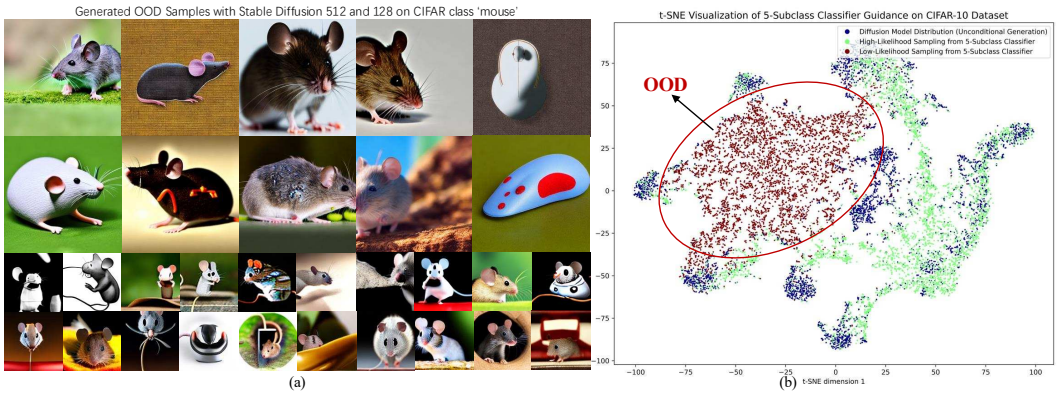

Figure 9: (a) OOD samples from the CIFAR-100 mouse class using Stable Diffusion (512/128). (b) t-SNE plot of samples from a 5-class CIFAR-10 classifier using DDPM: OOD guidance pushes samples toward unseen classes; ID guidance pulls them into known class regions.

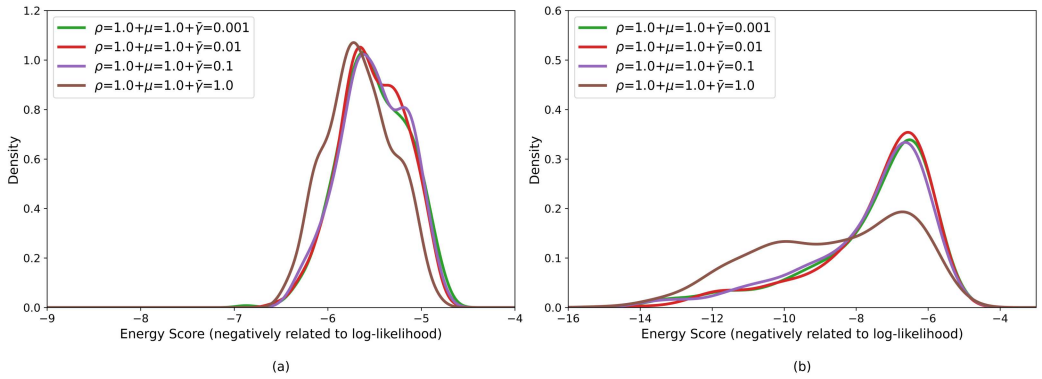

Figure 10: Energy score distributions of OOD samples generated with different $\bar{\gamma}$ values, under fixed $\bar{\rho} = \bar{\mu} = 1.0$. Each curve is computed from 500 samples. (a) ImageNet-100, (b) CIFAR-100.

likelihood under the seen-class classifier), and (3) OOD-guided sampling (i.e., minimizing likelihood under the same classifier).

Figure 9(b) visualizes the resulting samples using t-SNE. Unconditional samples (green) spread across the generative manifold. ID-guided samples (blue) are pulled toward the classifier's known class regions, forming tight clusters. In contrast, OOD-guided samples (red) are pushed away from the seen-class subspace and drift toward unfamiliar, in-between regions. This contrast highlights that GOOD's guidance directions are indeed semantically aligned: positive guidance avoids known categories, while negative guidance reinforces them.

A.4.2 Additional Ablation Study

Effect of $\bar{\gamma}$ in TFG. The hyperparameter $\bar{\gamma}$ controls the strength of Gaussian smoothing in the implicit dynamics of TFG. It determines how much noise is added when computing the smoothed target function $\tilde{f}(x)$, which in turn affects the stability and directionality of the guidance signal.

To evaluate its effect, we fix $\bar{\rho} = \bar{\mu} = 1.0$ and vary $\bar{\gamma} \in \{0.001, 0.01, 0.1, 1.0\}$, generating 500 OOD samples for each setting. Figure 10 shows the resulting energy score distributions on ImageNet-100 (left) and CIFAR-100 (right). These scores are inversely related to image likelihood, and are commonly used to characterize the degree of distributional shift.

We observe that small values of $\bar{\gamma}$ (e.g., 0.001, 0.01) produce sharp, unimodal distributions centered around a typical energy range, indicating stable and consistent guidance. As $\bar{\gamma}$ increases, the distributions become flatter or multimodal—especially on CIFAR—suggesting that excessive smoothing can

inject noisy or less directional gradients, resulting in more diverse but potentially less semantically aligned samples.

Overall, GOOD exhibits robustness to a range of $\bar{\gamma}$ values. However, we find that moderate values (e.g., 0.1 for ImageNet, 0.001 for CIFAR) provide the best trade-off between stability and anomaly diversity. These values are therefore used as default in all main experiments.

A.5 Additional Visualizations

To further illustrate the generative behavior of our approach, we present additional OOD samples produced by GOOD_{img} and $\text{GOOD}_{\text{feat}}$ on ImageNet-100. For each of the first 72 classes, we generate one sample per class under each guidance strategy, using fixed parameters.

Figures 11, 13, and 15 show results from GOOD_{img} with $\bar{\rho} = \bar{\mu} = 5$, where guidance is based on low image-level likelihood (free energy). These samples tend to exhibit a global shift toward abstraction—often distorting not just local texture but also the overall scene composition. In many cases, the image foreground, background, and object boundaries are all rendered with atypical patterns, surreal colors, or unnatural materials. Nonetheless, class semantics such as shape or context remain loosely preserved, resulting in globally outlying yet semantically grounded samples.

In contrast, Figures 12, 14, and 16 show results from $\text{GOOD}_{\text{feat}}$ with $\bar{\rho} = \bar{\mu} = 4$, where guidance targets sparsity in the deep feature space. Unlike GOOD_{img}, this variant preserves the overall image realism and structure but introduces rare or atypical class-specific features—such as uncommon textures, poses, or environmental contexts. For example, cats with distorted fur patterns or birds in unfamiliar postures. These samples lie closer to the data manifold but remain feature-wise distinctive, resembling "hard negatives" rather than globally abstract outliers.

Together, these visualizations highlight the complementary behavior of the two guidance modes: GOOD_{img} encourages holistic abstraction, while $\text{GOOD}_{\text{feat}}$ promotes subtle novelty. This duality enables us to generate a diverse spectrum of OOD examples—ranging from boundary-adjacent to semantically twisted—without retraining, simply by switching the guidance target.

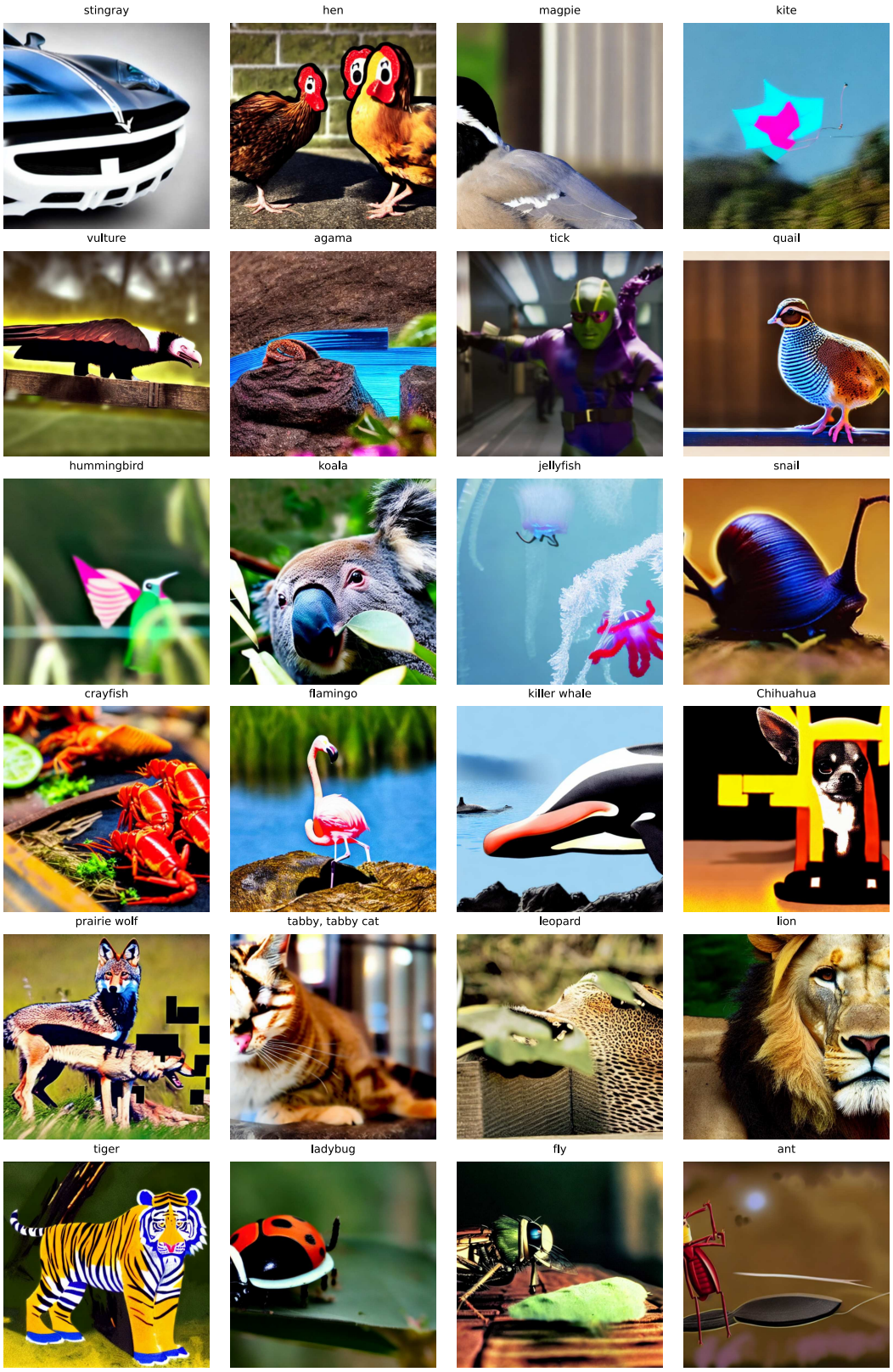

Figure 11: OOD samples generated by GOOD_{img} on ImageNet, guided by low image likelihood (free energy). One sample per class is shown.

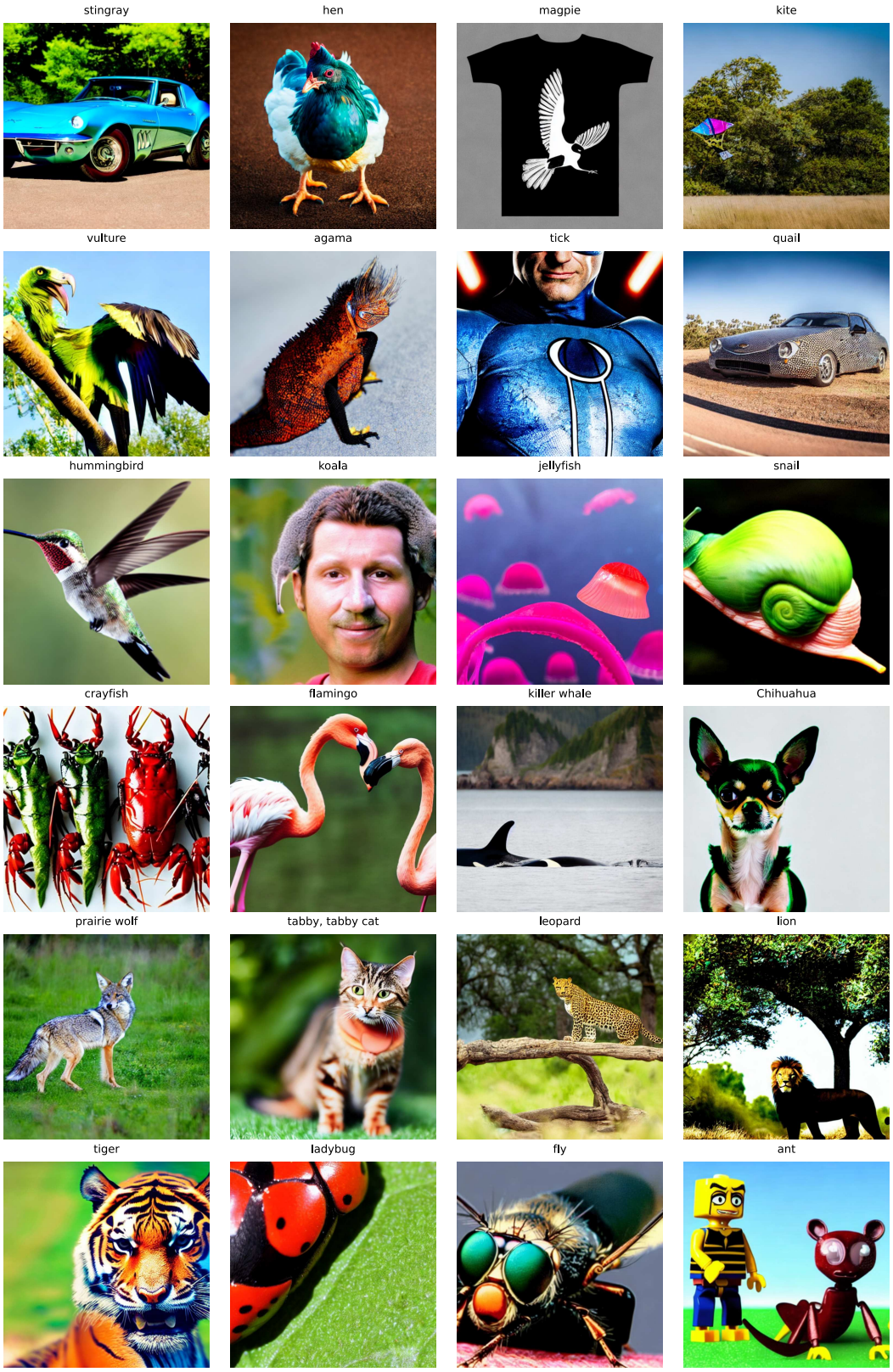

Figure 12: OOD samples generated by $\text{GOOD}_{\text{feat}}$ on ImageNet, guided by feature-space sparsity (kNN distance). One sample per class is shown.

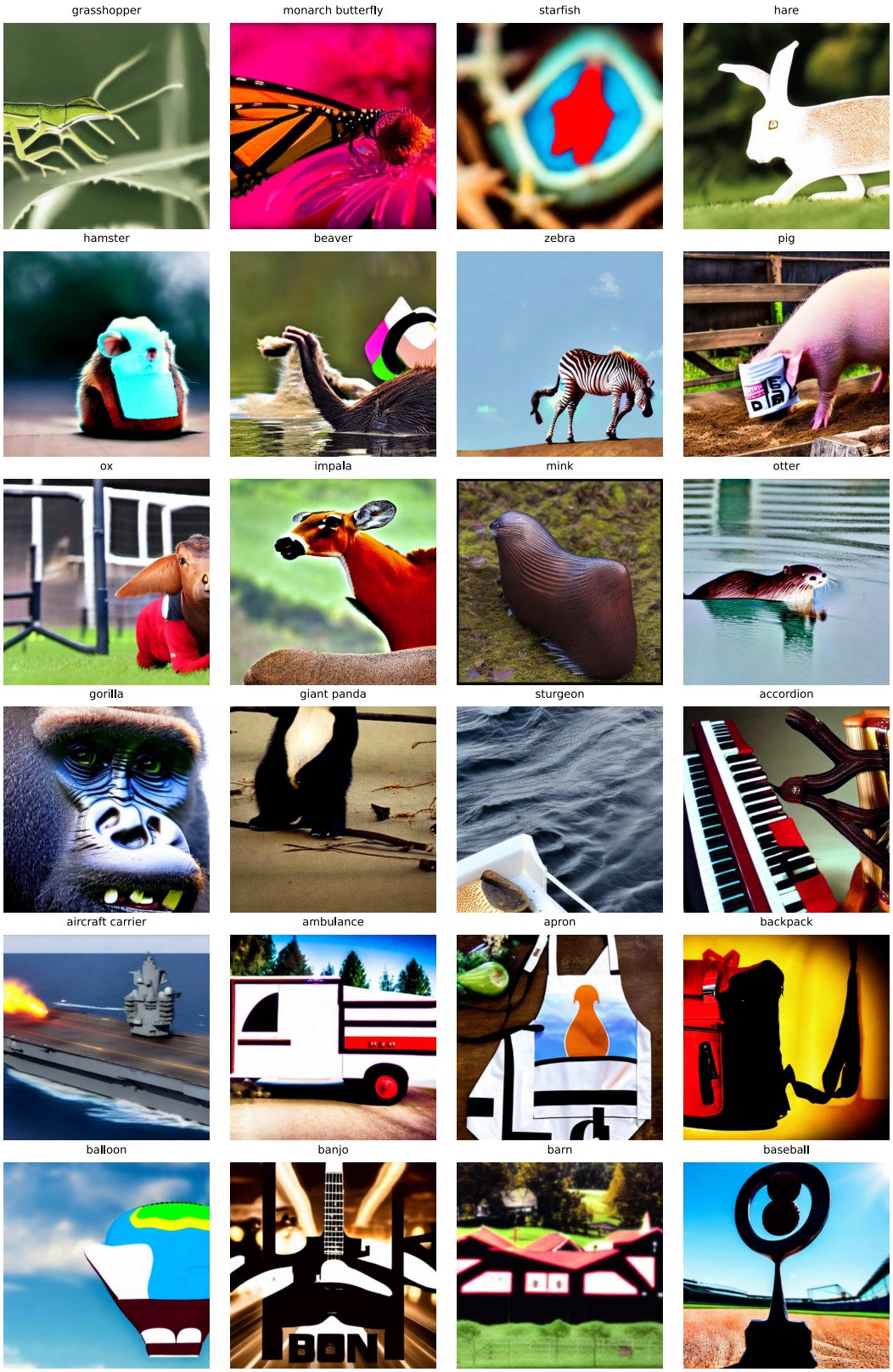

Figure 13: OOD samples generated by GOOD_{img} on ImageNet, guided by low image likelihood (free energy). One sample per class is shown.

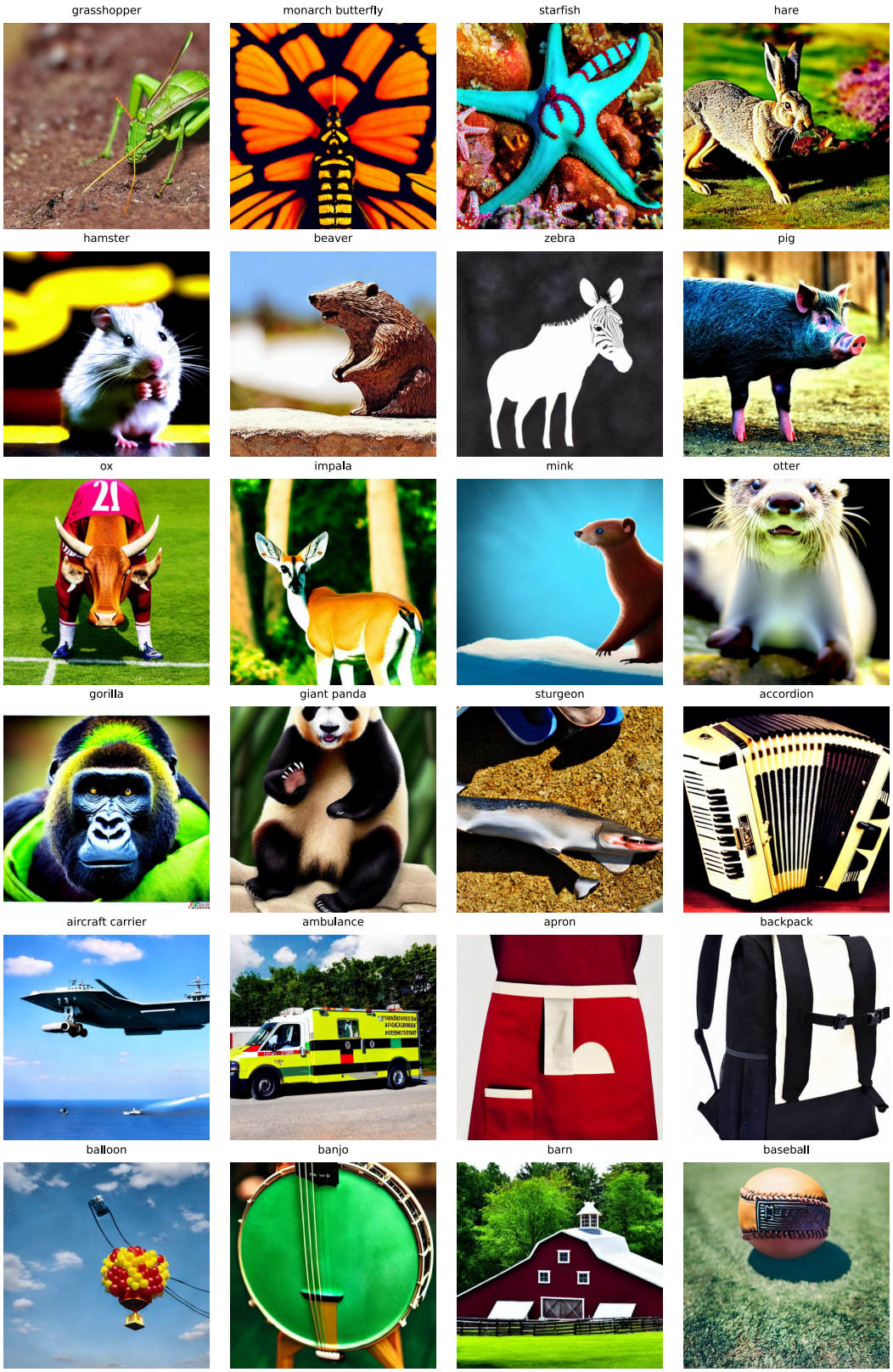

Figure 14: OOD samples generated by $\text{GOOD}_{\text{feat}}$ on ImageNet, guided by feature-space sparsity (kNN distance). One sample per class is shown.

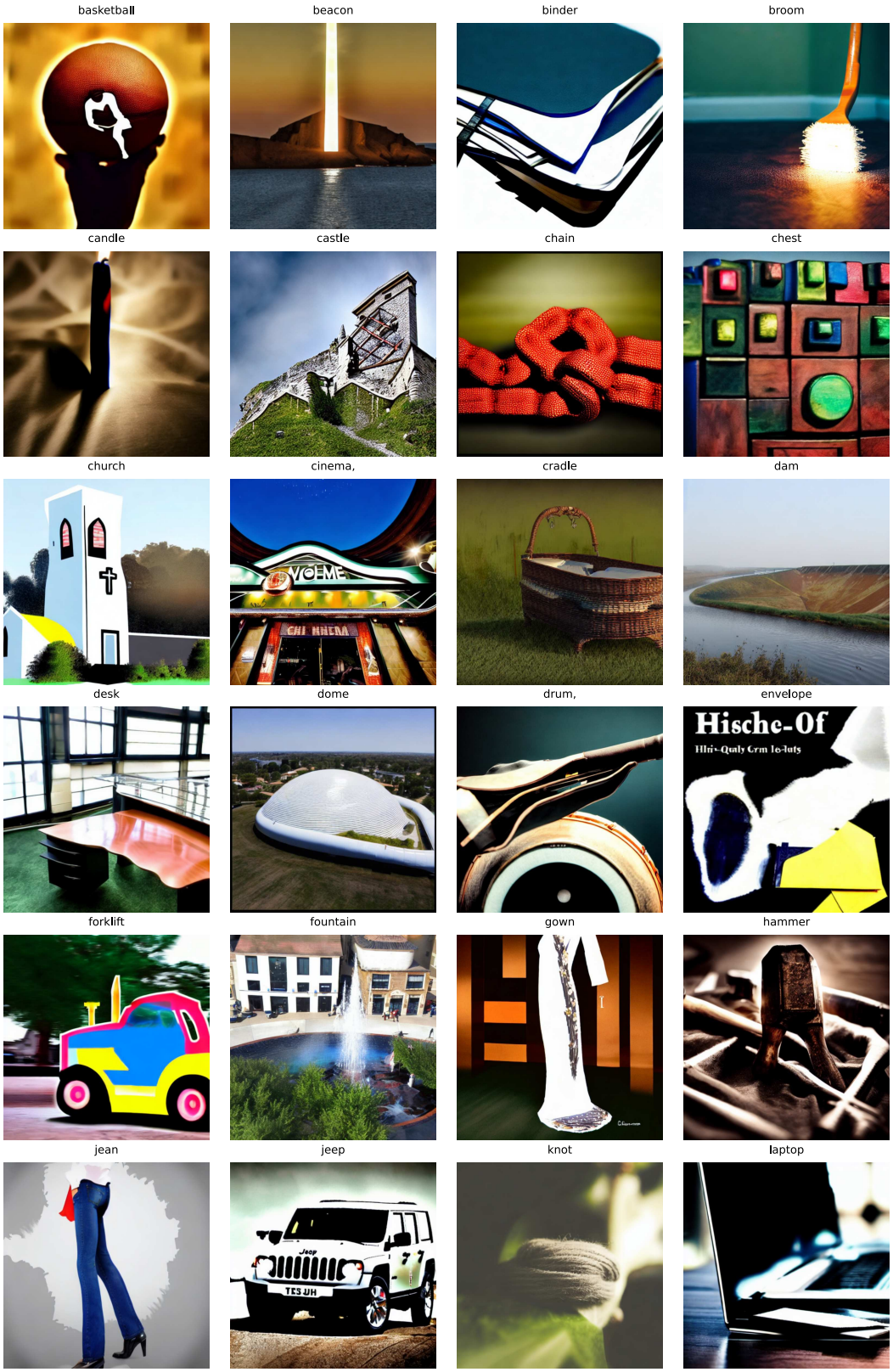

Figure 15: OOD samples generated by GOOD_{img} on ImageNet, guided by low image likelihood (free energy). One sample per class is shown.

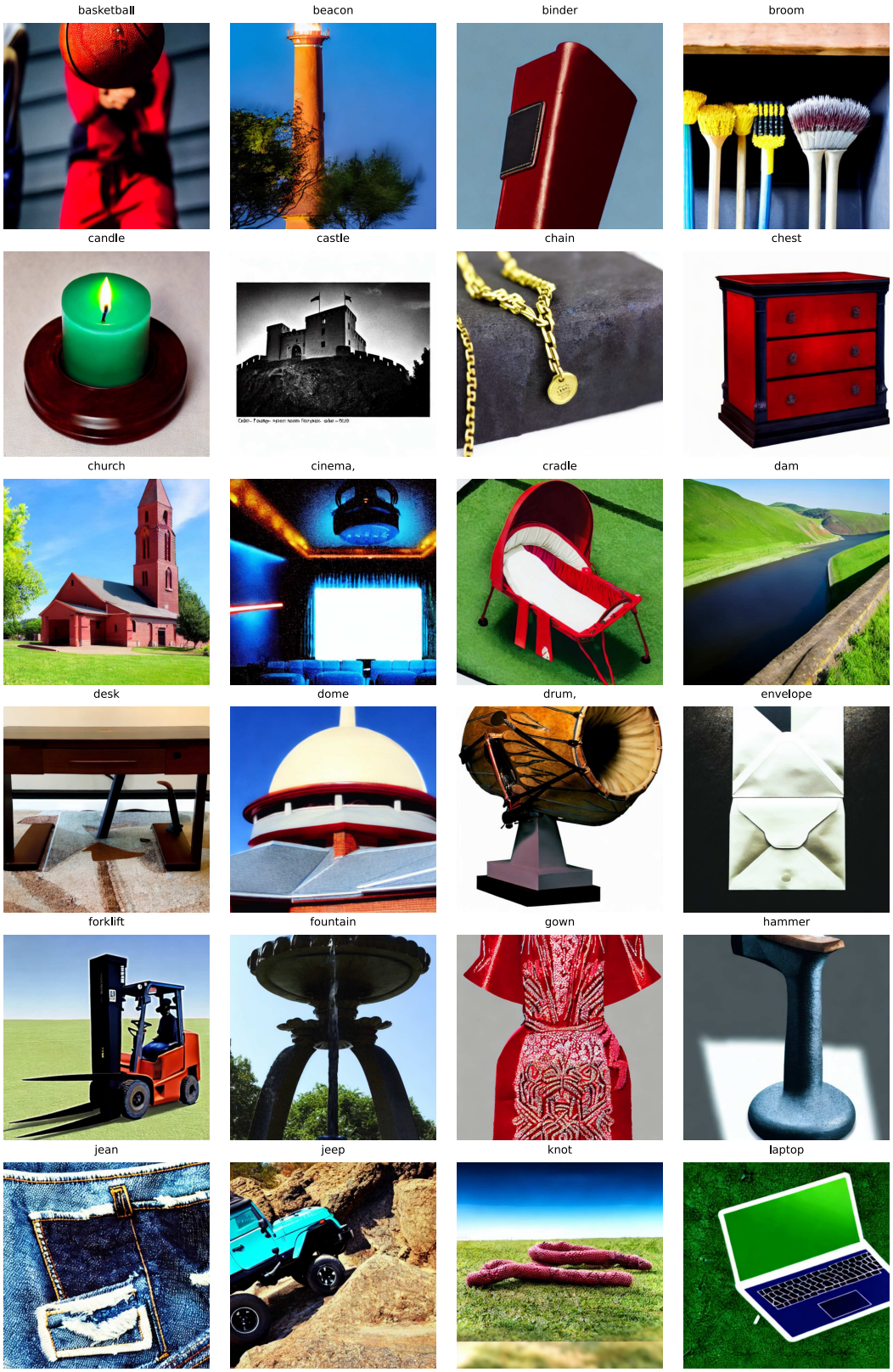

Figure 16: OOD samples generated by $\text{GOOD}_{\text{feat}}$ on ImageNet, guided by feature-space sparsity (kNN distance). One sample per class is shown.

References

[1] Faruk Ahmed and Aaron Courville. Detecting semantic anomalies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3154–3162, 2020.

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.

[3] Sima Behpour, Thang Long Doan, Xin Li, Wenbin He, Liang Gou, and Liu Ren. Gradorth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. *Advances in Neural Information Processing Systems*, 36:38206–38230, 2023.

[4] Francisco Caetano, Christiaan Viviers, Luis A Zavala-Mondragón, Peter HN de With, and Fons van der Sommen. Discopatch: Batch statistics are all you need for ood detection, but only if you can trust them. *arXiv preprint arXiv:2501.08005*, 2025.

[5] Jiankang Chen, Ling Deng, Zhiyong Gan, Wei-Shi Zheng, and Ruixuan Wang. Fodfom: Fake outlier data by foundation models creates stronger visual out-of-distribution detector. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1981–1990, 2024.

[6] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical image analysis*, 80:102479, 2022.

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[11] Lars Doorenbos, Raphael Sznitman, and Pablo Márquez-Neila. Non-linear outlier synthesis for out-of-distribution detection. *arXiv preprint arXiv:2411.13619*, 2024.

[12] Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? In *The Twelfth International Conference on Learning Representations*.

[13] Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *Advances in Neural Information Processing Systems*, 36:60878–60901, 2023.

[14] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*, 2022.

[15] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, et al. Manifold preserving guided diffusion. In *The Twelfth International Conference on Learning Representations*, 2024.

[18] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, pages 8759–8773. PMLR, 2022.

[19] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[20] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

[21] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

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

[25] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10951–10960, 2020.

[26] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

[27] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021.

[28] Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pages 10848–10865. PMLR, 2022.

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[30] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

[31] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[32] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[33] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

[34] Qilin Liao, Shuo Yang, Bo Zhao, Ping Luo, and Hengshuang Zhao. Bood: Boundary-based out-of-distribution data generation. 2025.

[35] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[36] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23946–23955, 2023.

[37] Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pages 15650–15665. PMLR, 2022.

[38] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

[39] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[40] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

[41] 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*, pages 10684–10695, 2022.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[43] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pages 8491–8501. PMLR, 2020.

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[45] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pages 32483–32498. PMLR, 2023.

[46] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[48] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in neural information processing systems*, 34:144–157, 2021.

[49] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European conference on computer vision*, pages 691–708. Springer, 2022.

[50] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

[51] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

[52] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[53] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022.

[54] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

[55] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[56] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[57] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

[58] Haiyun Yao, Zongbo Han, Huazhu Fu, Xi Peng, Qinghua Hu, and Changqing Zhang. Out-of-distribution detection with diversification (provably). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[59] Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Y Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *Advances in Neural Information Processing Systems*, 37:22370–22417, 2024.

[60] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[61] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184, 2023.

[62] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.