CONCORD: Concept-Informed Diffusion for Dataset Distillation

Supplementary Material

The appendix is organized into the following sections: In Sec. A we provide additional justification from the literature as far as the utility of using LLM-based concept informed learning. In Sec. B we introduce the implementation details of our method. In Sec. C we present more experiment results and analysis on the proposed CONCORD method. In Sec. D we show example generated images to better illustrate the effect of the proposed concept informing method.

A. Further Analytical Grounding

Perspective from XAI Explainable Artificial Intelligence (XAI) is an emerging area within machine learning that aims to provide users with greater insight into the functioning and mechanisms of black-box models, such as neural networks. Standard practice often involves knowledge extraction techniques, where broader, less precise, but simpler and thus more intuitive models are presented to explain the behavior of complex machine learning models. However, it has been noticed that this paradigm can be amended with the success and proliferation of LLMs [13]. In particular, LLMs enable a more iterative and active learning procedure, where user prompts can directly inform the learning process to accommodate user needs. Simultaneously, LLMs can periodically generate language-based explanations, offering updates on model progress and adjustments.

One of the key advantages of LLMs is their potential for personalization [6]. Given the rich variety of concepts in the extensive training data and the depth of the developed models, LLMs can foster a detailed and more human-centric understanding. This allows the models to tune the learning process towards specific application-driven concerns. The effectiveness of LLMs as explainers [25] provides the clear potential to address important use case concerns that are difficult to represent through standard analytical loss functions. This adaptability allows LLMs to bridge gaps between the learning objectives and real-world applications.

Perspective from Instrumental DD In the recent work [26], it has been argued that an important analytical consideration often overlooked in most DD optimization formulations is its instrumentality. Specifically, synthetic data is typically not just used to solve the same learning problem in the same setting, but rather the dataset is expected to be used in some broader applications of interest to the user. These applications may have information needs that are not inherently condensed by standard off-the-shelf DD algorithms. By including additional custom criteria into the DD optimization formulation, while still incorporating existing powerful tools, DD can be more effectively steered towards performance on desired use cases. In this work, concepts are employed to facilitate natural human taxonomy with respect to object identification and recognition. This consideration substantially improves the desired property in an explainable way by aligning the synthetic dataset with the textual concept attributes.

B. More Implementation Details

Baselines We adopt Minimax [15] and Stable Diffusion unCLIP Img2Img [41] as the baselines to illustrate the efficacy of our proposed concept informing method. These two baselines represent two different application scenarios, as outlined below.

For Minimax, a fine-tuning process is conducted on DiT with the ImageNet-1K data. While it is expected that fine-tuning on target datasets yields superior performance, it also demands more resource consumption. Additionally, class labels are utilized for conditioning the denoising process, which might be inconvenient when extending the model to broader datasets. We adopt the default parameter setting in the original paper. The entire ImageNet-1K is partitioned into 50 subsets, each containing data of 20 classes. For each subset, a DiT model [39] is fine-tuned for 8 epochs. The mini-batch size, representative weight, and diversity weight are set as 8, 0.002, and 0.008, respectively. During inference, the corresponding fine-tuned model is loaded to generate data for specific classes.

For unCLIP Img2Img, we utilize the pre-trained model without any fine-tuning adjustments[1]. Random real images are fed into the model simultaneously with text prompts to generate high-quality samples without losing the information of the original data distribution. We adopt 28 prompt templates for generating images, *e.g.*, "a photo of a nice {$class_name}" [40]. The utilization of text prompts for conditioning provides significant flexibility, enabling data generation for custom datasets without extra training efforts. While the absence of fine-tuning may lead to a slight reduction in generation quality, it allows for direct application of the proposed CONCORD method to any custom data given relevant text descriptions. During inference, the same pre-trained model is adopted for generating images for all target categories.

Concept Acquirement We use GPT-4o to retrieve descriptive concepts for different categories. The full adopted prompt is as follows:

[1] https://huggingface.co/radames/stable-diffusion-2-1-unclip-img2img

You are an expert in computer vision and image analysis. Here is the task: <task>I want to use some visual descriptions to identify different categories in ImageNet dataset. Please first consider whether there exist categories with similar appearance to {$class_name}. Then please give 10 short descriptions describing the appearance features that the {$class_name} has and can be used to distinguish it from other classes. The phrases should only focus on visual appearance of body parts or components instead of functioning. Each phrase should be detailed but also shorter than 128 characters. Each phrase starts with non-capitalized characters.</task> Give the answer in the form of <answer>["$class_name", ["phrase1", "phrase2", "phrase3", "phrase4", "phrase5", "phrase6", "phrase7", "phrase8", "phrase9", "phase10"]]</answer>.

After retrieving the original concepts, we perform a similarity calculation between the textual concepts and real images of the corresponding category. The top 5 most similar concepts are selected for the subsequent informed diffusion process, as described in Sec. 3.2. This approach helps ensure that the selected concepts align closely with the real images, thereby enhancing the validity of the concepts used in the diffusion process to a certain extent.

Informing The informing process involves similarity calculation between embeddings of images and textual concepts. We use a CLIP model with ViT-L as the visual encoder, pre-trained on LAION-2B data [45] to encode these embeddings. The model weights can be downloaded from Hugging Face[2]. The generation process involves 50 denoising steps for each sample. Prior to denoising, 5 descriptive concepts from the same class as well as 10 negative concepts each from a different class are retrieved for the sample. Before extracting text embeddings, the concepts are grouped with the corresponding class name using the following format:

{$class_name} with {$concept}.

During each denoising step, the similarity between the generated sample and corresponding concepts is calculated for the informing objective in Eq. 11. The informing weight λ is set as 1 for optimal performance. The concept informing guides the denoising process to obtain completeness on essential details, and thereby enhances the instance-level quality of the generated images.

[2]https://huggingface.co/laion/CLIP-ViT-L-14-laion2B-s32B-b82K

Table 7. Comparison with different optimization objectives and their combination on ImageWoof.

Method	IPC		
	1	10	50
None	$16.7_{\pm 0.7}$	$37.9_{\pm 1.1}$	$63.6_{\pm 0.6}$
Classifier	$16.9_{\pm 0.7}$	$38.5_{\pm 1.0}$	$65.2_{\pm 0.8}$
Contrastive	$\mathbf{17.4}_{\pm 1.1}$	$\mathbf{40.7}_{\pm 0.4}$	$\mathbf{66.1}_{\pm 1.1}$
Combination	$16.7_{\pm 0.3}$	$38.8_{\pm 1.9}$	$65.7_{\pm 0.7}$

Validation We adopt the validation protocol in RDED [50] to evaluate the performance of distilled data. We mainly employ a ResNet-18 [17] architecture for experiments, with additional ones run on ResNet-101 and ConvNets as shown in Tab. 1. Specifically, for ImageWoof, ImageNet-100, ImageNet-1K, we adopt 5-layer, 6-layer, and 4-layer ConvNets, respectively, consistent with the settings in RDED. For ImageWoof, the images are resized to 128×128 on ConvNet-5, while for all other cases, the images are resized to 224×224 for evaluation.

For ImageNet and its subsets, we employ pre-trained models[3] to generate soft labels and apply Fast Knowledge Distillation [47]. The models are trained for 300 epochs using the AdamW optimizer, with an initial learning rate of 0.001 and a weight decay of 0.01. A cosine annealing scheduler is used to adjust the learning rate. The mini-batch size for evaluation is set the same as IPC, *e.g.*, a mini-batch size of 10 is adopted for evaluating 10-IPC sets. The applied data augmentation techniques include patch shuffling [50], random crop resize [58], random flipping and CutMix [63]. After training on the distilled dataset, the model is then evaluated with the original validation set, and Top-1 accuracy is used as the validation performance. Each experiment is performed for three times, and the mean accuracy and standard variance are reported in the results.

For the Food-101 dataset, since no pre-trained models are provided by RDED, we train a ResNet-18 model on the original training set for 300 epochs, and use it for soft-labeling. It is important to note that the utilization of pre-trained models is independent from the sample generation process, and is only for fair comparison with state-of-the-art methods, which can be omitted in actual applications.

C. Extended Experiments and Analysis

Ablation on Denoising Steps In the main experiments, we adopt 50 denoising steps for sample generation. The effect of varying the number of denoising steps is evaluated and presented in Fig. 5, with the unCLIP Img2Img model as the baseline. As the number of denoising steps increases, the accuracy of the baseline distilled data shows an upward trend under the IPC setting of 10, while it is relatively

[3]https://github.com/LINs-lab/RDED

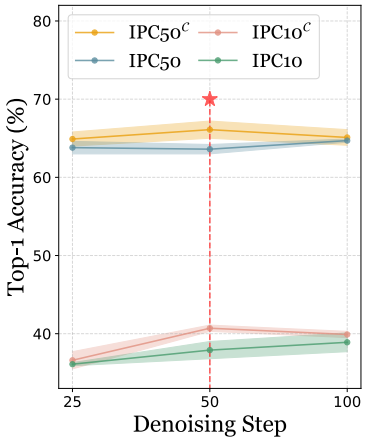


Figure 5. Parameter analysis on the denoising step number.

Table 8. Validation performance on ImageNet-1K with ResNet-101. Minimax^C indicates the application of CONCORD.

IPC	10	50
RDED	48.3$_{\pm 1.0}$	61.2$_{\pm 0.4}$
Minimax^C	**50.1**$_{\pm 0.9}$	**66.0**$_{\pm 0.7}$

Table 9. Comparison with adding concepts to text inputs. Minimax^C indicates the application of CONCORD.

IPC	10	50
unCLIP	37.9$_{\pm 1.1}$	63.6$_{\pm 0.6}$
prompt	37.1$_{\pm 0.7}$	63.6$_{\pm 0.9}$
Minimax^C	**40.7**$_{\pm 0.4}$	**66.1**$_{\pm 1.1}$

consistent for the 50-IPC setting. For concept informing, fewer denoising steps result in insufficient informing, leading to performance similar to that of the original images. Conversely, when too many denoising steps are used, the informing starts to disrupt the standard denoising process, leading to a drop in performance. While more fine-grained tuning of the informing weight could potentially mitigate the negative effect, more denoising steps also lead to extra computational costs. Therefore, we adopt 50 denoising steps as the standard setting in our experiments.

Combining Concept Informing and Classifier Guidance The proposed CONCORD method informs the diffusion process to contain more discriminative details for enhancing instance-level image quality. While concept informing sharing similarity to classifier guidance, the key difference is that CONCORD utilizes the similarity between generated samples and concepts as optimization targets, without relying on pre-trained classifiers. We also experiment with combining these two kinds of constraints to simulate scenarios where pre-trained classifiers are available. As shown in Tab. 7, when functioning independently, our proposed contrastive concept informing outperforms classifier guidance. It supports our hypothesis that detailed descriptions provide richer information compared with the category-level labels, and are more helpful in refining the instance-level sample quality. However, when both types of guidance are combined, classifier guidance does not provide additional information and disrupts the concept informing process. As a result, the combined approach shows less effective performance improvement on the generated images. Therefore, in the main experiments, we exclusively use the proposed concept informing, as it delivers better overall results and saves extra computational consumption.

Table 10. Inference time cost comparison of generating one sample under the mini-batch size of 1 between the baselines and the proposed CONCORD method.

Method	Minimax	Minimax^C	unCLIP	unCLIP^C
Time (s)	2.1	4.3	9.7	20.8

Results with ResNet-101 We additionally evaluate the validation performance on ImageNet-1K with ResNet-101 on the generated images and present the results in Tab. 8. Consistent with Tab. 1, the advantage of CONCORD is more significant with ResNet-101 than ResNet-18.

Adding Concepts to the Prompt We test directly adding the retrieved concepts to the prompt as conditioning. As shown in Tab. 9, the enriched prompt (denoted as "prompt" in the table) cannot achieve better performance compared with the baseline. This might partly be attributed to limited diversity of fixed prompts. Comparatively, the proposed CONCORD method is also informed by negative concepts, which enhance the stability and diversity of the sample generation process. Moreover, for DiT-based approaches, where the conditioning comes from class labels, the concept informing provides better flexibility.

Extra Computational Cost We report the inference time cost for generating an image on both Minimax and unCLIP Img2Img in Tab. 10. Comparatively, introducing CONCORD increases the original inference cost by approximately 1-1.5 times. unCLIP Img2Img involves Stable Diffusion v2-1 model [43], which demands more computational resources compared with Minimax, which uses a DiT model [39] as the denoising backbone. During inference, Minimax with CONCORD only requires about half the time of the unCLIP baseline. Although Minimax performs better as a baseline, the advantage is based on extra fine-tuning processes on the target dataset. Therefore, the model choice

Table 11. Performance comparison with state-of-the-art methods on ImageWoof. The superscript $^{\mathcal{C}}$ indicates the application of our proposed CONCORD method. **Bold** entries indicate best results, and underlined ones illustrate improvement over baseline.

IPC (Ratio)	Test Model	Random	K-Center	Herding	IDM	Minimax	Minimax$^{\mathcal{C}}$	Full
1 (0.08%)	ConvNet	$16.3_{\pm 0.5}$	$15.8_{\pm 0.6}$	$16.8_{\pm 1.1}$	$17.1_{\pm 0.2}$	$16.7_{\pm 0.2}$	**17.8**$_{\pm 0.8}$	$69.0_{\pm 0.2}$
	ResNet-18	$15.1_{\pm 0.2}$	$15.7_{\pm 0.8}$	$16.1_{\pm 0.4}$	$16.7_{\pm 0.5}$	$15.3_{\pm 1.1}$	**16.9**$_{\pm 1.0}$	$76.9_{\pm 0.1}$
	ResNet-101	$14.0_{\pm 0.6}$	$13.7_{\pm 1.2}$	$14.1_{\pm 0.6}$	$16.3_{\pm 0.6}$	$14.2_{\pm 1.1}$	**14.9**$_{\pm 1.3}$	$77.6_{\pm 0.2}$
10 (0.8%)	ConvNet	$40.5_{\pm 1.5}$	$37.1_{\pm 0.9}$	$41.2_{\pm 0.4}$	$38.5_{\pm 0.6}$	$41.2_{\pm 0.8}$	**43.1**$_{\pm 0.5}$	$69.0_{\pm 0.2}$
	ResNet-18	$34.3_{\pm 1.6}$	$33.1_{\pm 0.5}$	$36.8_{\pm 0.6}$	$36.5_{\pm 1.2}$	$42.8_{\pm 1.1}$	**44.4**$_{\pm 0.9}$	$76.9_{\pm 0.1}$
	ResNet-101	$32.1_{\pm 1.0}$	$31.6_{\pm 0.3}$	$33.8_{\pm 0.4}$	$30.8_{\pm 1.2}$	$35.7_{\pm 0.9}$	**36.5**$_{\pm 0.9}$	$77.6_{\pm 0.2}$
50 (3.8%)	ConvNet	$60.9_{\pm 0.9}$	$57.7_{\pm 1.2}$	$60.4_{\pm 0.8}$	$61.0_{\pm 0.6}$	$61.1_{\pm 0.8}$	**62.5**$_{\pm 0.9}$	$69.0_{\pm 0.2}$
	ResNet-18	$67.1_{\pm 1.0}$	$64.3_{\pm 0.9}$	$67.6_{\pm 0.5}$	$64.9_{\pm 0.6}$	$67.8_{\pm 0.5}$	**69.2**$_{\pm 1.0}$	$76.9_{\pm 0.1}$
	ResNet-101	$61.4_{\pm 0.7}$	$58.8_{\pm 0.4}$	$60.8_{\pm 0.3}$	$57.2_{\pm 0.6}$	$62.2_{\pm 0.6}$	**63.6**$_{\pm 0.2}$	$77.6_{\pm 0.2}$

should consider multiple factors, including the balance between training and inference consumption.

Comparison to More Baselines In addition to the results in Tab. 1, we also conduct experimental comparison with random sampling, K-Center [46], Herding [57] and IDM [66] in Tab. 11. For the methods based on original samples, we first resize the images to 128×128 for ConvNet and 224×224 for ResNet before running validation.

K-Center and Herding are two methods for selecting coresets from the original data, with unstable performance improvement compared with random sampling. IDM is a dataset distillation method based on distribution matching, which is effective under small IPC settings. However, as the required sample number increases, the generated images often perform worse than randomly selected original samples. The baseline Minimax comparatively provides more stable information condensation across different IPC settings. When combined with the proposed CONCORD method, the overall dataset quality is significantly enhanced, surpassing all other methods in terms of accuracy. Especially for ConvNet and ResNet-18 architectures, training with 50 images per class achieves less than 10% performance gap from training with the entire original set. As larger models (*e.g.*, ResNet-101) require more data and training iterations to get good performance, there still remains a certain performance margin between distilled data and the original full set.

Validation Results on ImageNet-Sketch In addition to standard test sets, we also validate the performance of distilled data on ImageNet-Sketch, which is out of the standard image distribution. The experiment is conducted on the ImageWoof subset. The results are presented in Tab. 12. CONCORD helps both Minimax and unCLIP generate more robust images, except for the IPC-50 setting of unCLIP. It validates that more complete and correct visual details are also helpful for OOD test sets.

Table 12. Validation results on the Out-Of-Distribution ImageNet-Sketch testing set. The training set is ImageWoof. $^{\mathcal{C}}$ indicates the application of CONCORD.

IPC	10	50
Minimax	$11.8_{\pm 1.6}$	$21.2_{\pm 1.6}$
Minimax$^{\mathcal{C}}$	$13.0_{\pm 0.8}$	$23.0_{\pm 1.2}$
unCLIP	$16.8_{\pm 0.7}$	$20.1_{\pm 0.9}$
unCLIP$^{\mathcal{C}}$	$17.4_{\pm 0.4}$	$19.1_{\pm 1.1}$

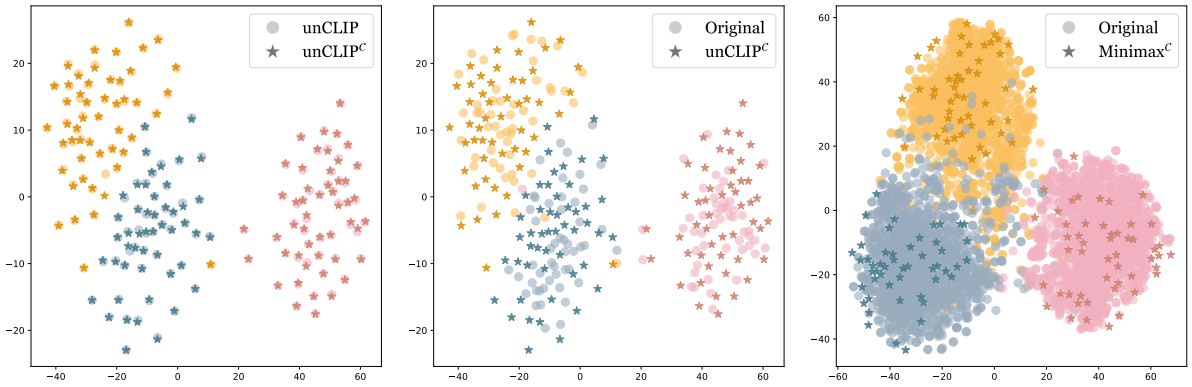
Table 13. Validation results on ImageNet-Nette. $^{\mathcal{C}}$ indicates the application of CONCORD.

IPC	10	50
Minimax	$53.0_{\pm 1.4}$	$71.2_{\pm 0.8}$
Minimax$^{\mathcal{C}}$	$54.2_{\pm 0.7}$	$73.3_{\pm 0.8}$
unCLIP	$46.9_{\pm 1.9}$	$67.3_{\pm 0.8}$
unCLIP$^{\mathcal{C}}$	$49.1_{\pm 1.5}$	$68.4_{\pm 0.7}$

Validation Results on ImageNette We further distill images for ImageNette, where the difference between classes is larger, indicating that it is an easier subset for classifiers. Compared with both the Minimax and unCLIP baselines, CONCORD brings a stable performance boost across all IPC settings. The results suggest that CONCORD is also useful for circumstances where the classes are not so fine-grained. The supplemented visual details still help distilled images obtain better validation performances.

Feature Distribution Visualization We provide the feature distribution comparison in Fig. 6 to illustrate the effects of our proposed CONCORD method.

First, the left figure shows the t-SNE features of samples generated with and without CONCORD. CONCORD works as a training-free guidance at the inference stage, without changing the main object in the images. By refining essen-



(a) (b) (c)

Figure 6. Feature distribution visualization of (a) samples generated by unCLIP with and without CONCORD; (b) samples generated by unCLIP with CONCORD and original samples used for conditioning; (c) samples generated by Minimax with CONCORD and original samples. Different colors indicate different categories.

tial details in the generated samples, CONCORD enhances instance-level concept completeness and improves the overall quality of the distilled datasets. However, these detail refinements have a mild effect on the feature distribution, indicating that with an already well-structured distribution, CONCORD can further improve performance without disrupting the underlying data distribution.

Second, the middle figure compares the generated images with the original ones used as conditioning in unCLIP Img2Img. The generated images closely align with the original distribution, validating their efficacy in capturing the properties of the original dataset. It demonstrates the suitability of using these generated images for training models.

Lastly, the right figure shows the distribution of samples generated by Minimax with CONCORD and the entire original set. The generated samples demonstrate comprehensive coverage over the original distribution, ensuring that they represent a wide range of instances. While with sufficient diversity brought by samples distributed near decision boundaries, the generated samples also reduces noise in the overlapping regions between categories. It makes the generated dataset stable and effective for training models, when computational resources are limited.

Analysis on the Retrieved Concepts The effects of different prompts and LLMs have been quantitatively investigated in Tab. 4. We further conduct qualitative comparison for the retrieved descriptions to explicitly analyze the informing effects of different concepts. As shown in Fig 7, descriptions of indigo bunting and streetcar are retrieved based on three settings: prompt for zero-shot classification on GPT-4o (denoted as "Cls"), our adopted prompt on GPT-3.5, and our prompt on GPT-4o. The "Cls" prompt retrieves general descriptions about the object. However, in many cases the retrieved descriptions are still too coarse for fine-grained informing. The descriptions retrieved by GPT-3.5 are more detailed, but contain a large number of non-visual attributes, which cannot provide valid signal during concept informing. Comparatively, our adopted prompt on GPT-4 successfully emphasizes the detailed visual features of corresponding categories. These fine-grained descriptions enables the proposed CONCORD method to effectively enhance instance-level concept completeness and further improves the overall quality of the distilled datasets.

D. Sample Comparison

We further present more example generated images in the following sections.

Comparison with Baselines Firstly, we show comparison on Minimax and unCLIP Img2Img with and without applying the proposed CONCORD method in Fig. 8 and Fig. 9, respectively. When baseline methods fail to present essential features and often lead to image defects, CONCORD significantly enhances the concept completeness in samples.

Failure Cases We also present failure cases where the proposed CONCORD method fails to correct or supplement essential features in the images in Fig. 10. It can be seen that the informing tries to modify some defects in the original image, but the eventual refinement is limited. There are also some cases where the informing fails to find the missing or incorrect details. Especially the informing fails to refine the details when the number of body parts is incorrect or the body part is completely missing in the original generation results. There is still much space for further improving the instance-level sample quality for dataset distillation.

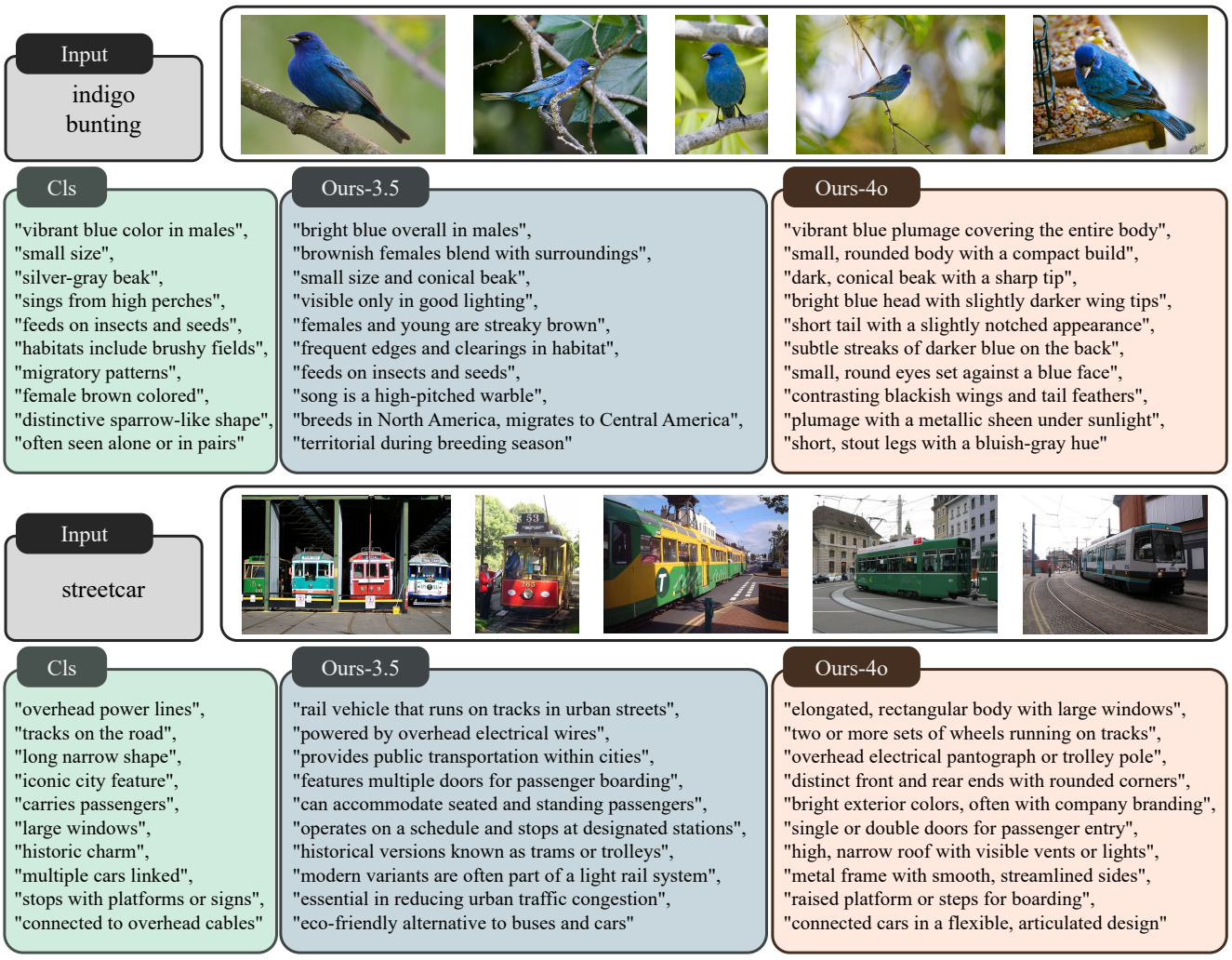


Figure 7. The comparison between concepts retrieved by different prompts and LLMs. "Cls" refers to prompts used for zero-shot classification attribute retrieval. Example images of corresponding classes are also presented to show their appearance features.

More Sample Visualization Additionally, we present more example samples across various categories to demonstrate the overall high quality of the dataset generated by the proposed CONCORD method. Specifically, in Fig. 11 we present samples generated for the Food-101 dataset. In Fig. 12 images of animal categories in the ImageNet-1K dataset are generated by Minimax with CONCORD applied. In Fig. 13 we show images of other categories in the ImageNet-1K dataset. The high-quality generated samples form an effective surrogate dataset, which achieves state-of-the-art performance.

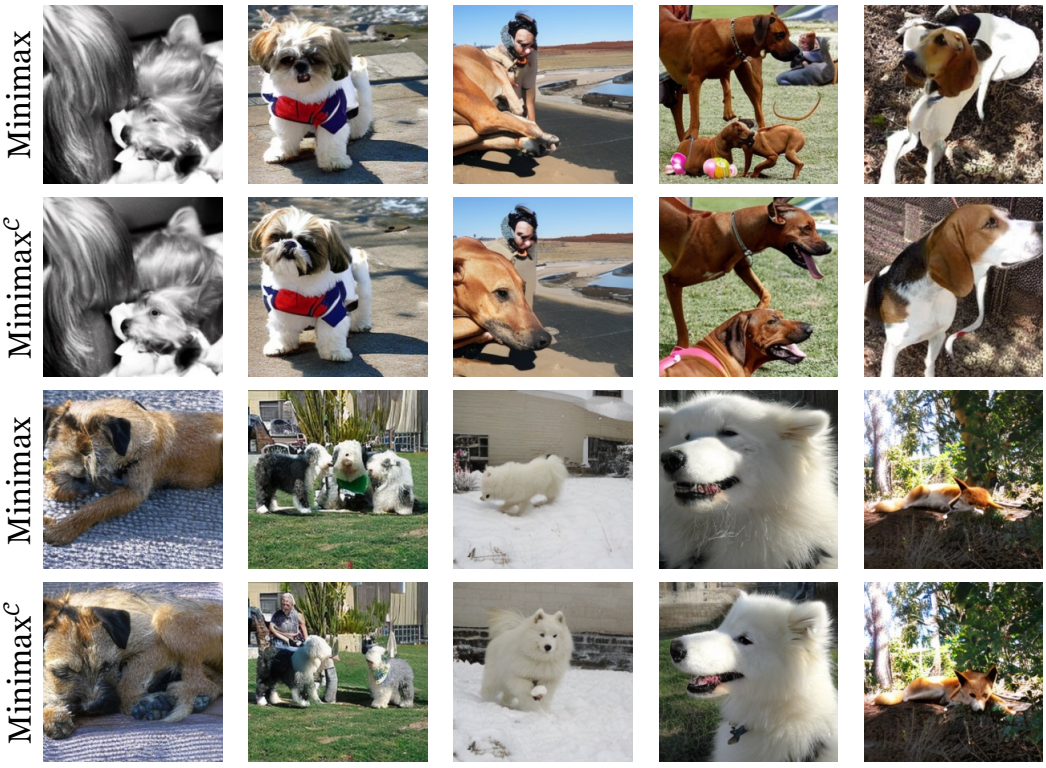


Figure 8. Example generated image comparison on Minimax with and without the proposed CONCORD method (denoted as $^{\mathcal{C}}$).

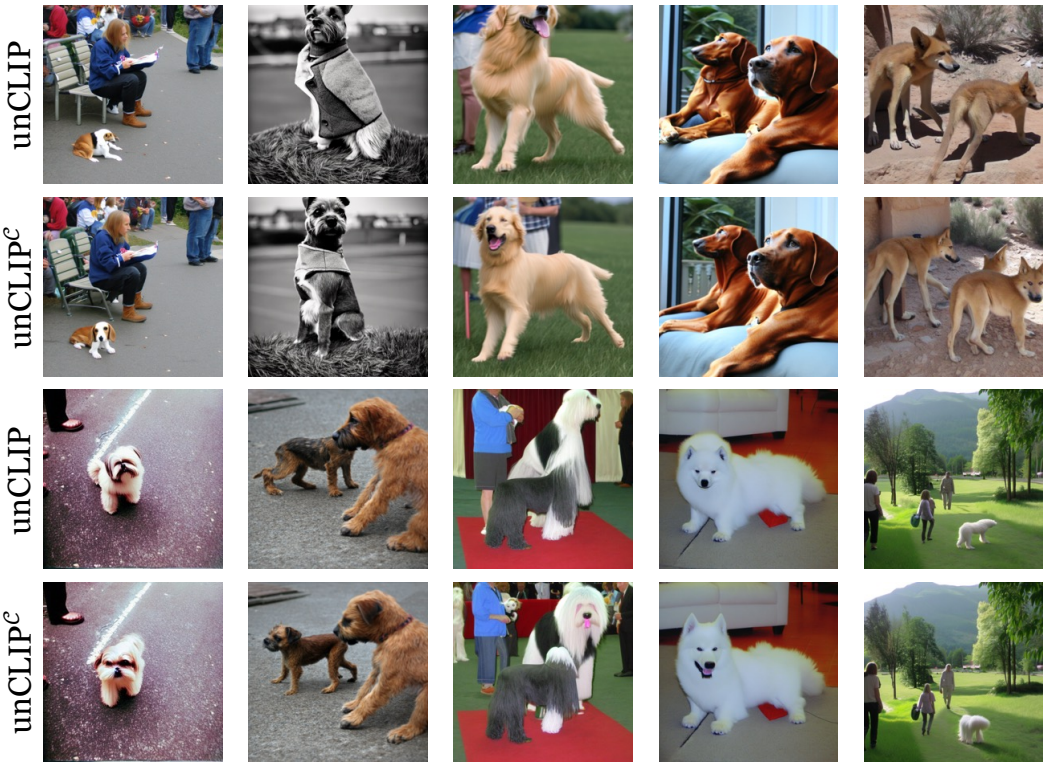


Figure 9. Example generated image comparison on unCLIP Img2Img with and without the proposed CONCORD method (denoted as $^{\mathcal{C}}$).

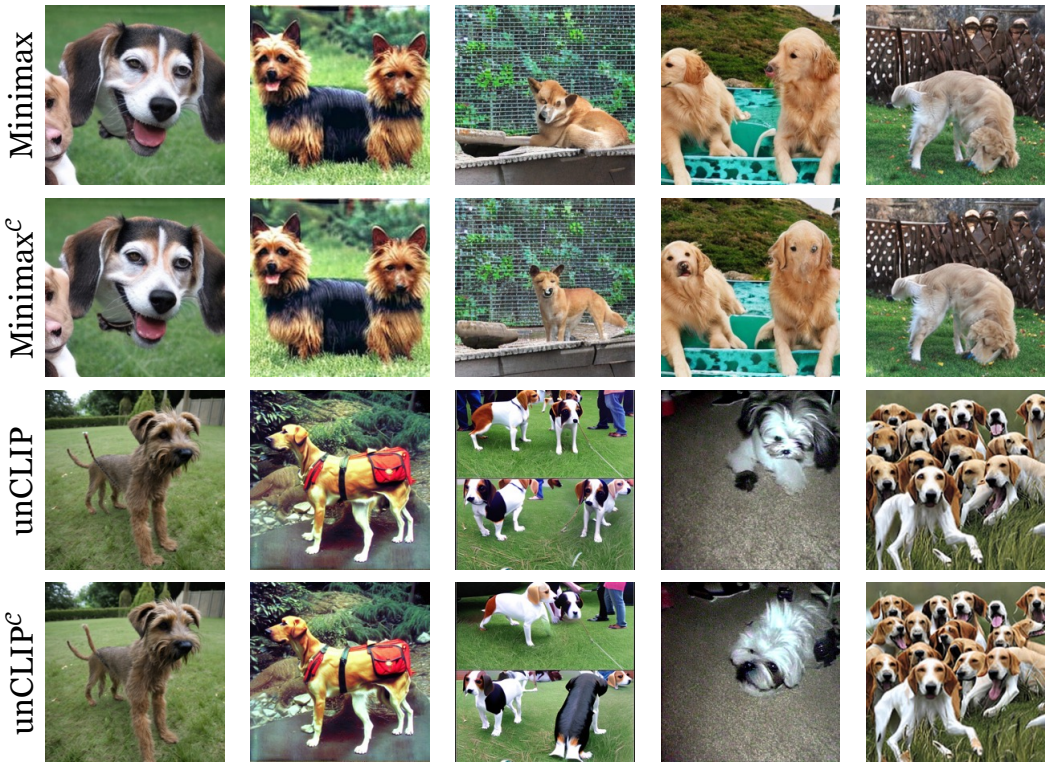


Figure 10. Example cases where CONCORD fails to supplement or modify incorrect concepts in the images ($^{\mathcal{C}}$ indicates the application of CONCORD).

Figure 11. Example images generated by the proposed CONCORD method on the Food-101 dataset. The class names are annotated below the images.

Figure 12. Example animal images generated by the proposed CONCORD method on the ImageNet-1K dataset. The employed diffusion pipeline is the fine-tuned Minimax model. The class names are annotated below the images.

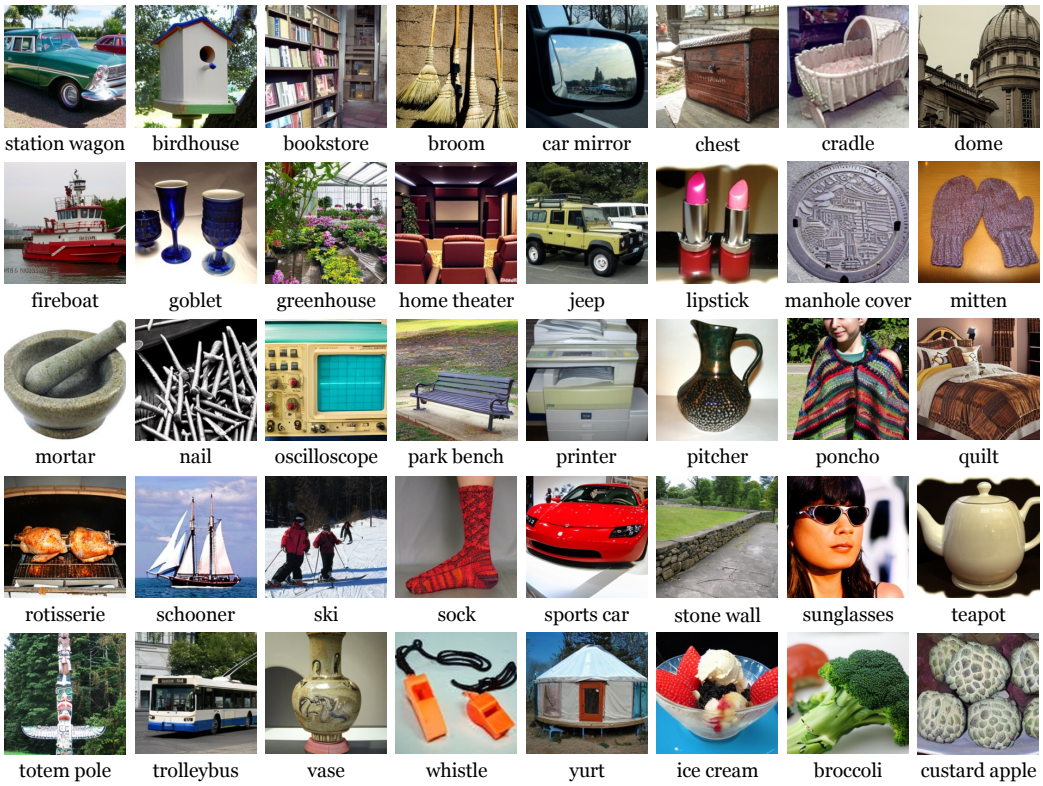


Figure 13. Example other images generated by the proposed CONCORD method on the ImageNet-1K dataset. The employed diffusion pipeline is the fine-tuned Minimax model. The class names are annotated below the images.