DeIL: Direct-and-Inverse CLIP for Open-World Few-Shot Learning

Supplementary Material

7. Related Work

7.1. Foundation Models

Currently, extensive research is being conducted on foundation models. In this context, we would like to present three models used in our paper: CLIP [35], CLIPN [40], and DALL-E [36].

CLIP (Contrastive Language-Image Pretraining) stands as a formidable vision and language model developed by OpenAI, representing a groundbreaking achievement in the field of artificial intelligence. Its strength lies in its seamless integration of both vision and language understanding, leading to outstanding performance across a wide range of tasks. The architecture of CLIP is built upon two central components: an image encoder and a text encoder. The image encoder meticulously processes input images, extracting high-level visual features that captures vital information about the image's content and style. Similarly, the text encoder takes on the responsibility of encoding natural language descriptions or prompts into a latent space representation. This encoding process is carefully designed to capture the semantic meaning and nuanced context embedded within the text. At the heart of CLIP's innovation lies the alignment of image and textual representations within a shared embedding space. This alignment is achieved through a contrastive learning framework, where the model is trained to maximize the similarity between corresponding image-text pairs while simultaneously minimizing the similarity between unrelated pairs. By training on extensive datasets containing diverse images and their associated textual descriptions, CLIP exhibits the capability to grasp a broad spectrum of concepts and relationships across different data modalities. Once trained, the CLIP model demonstrates its versatility by excelling in various tasks, including zero-shot image classification, caption generation, and image-to-text retrieval—all driven by the alignment of image and text embeddings. In essence, the CLIP model seamlessly harmonizes image and text encoders to create a unified representation space that enables comprehensive vision and language understanding. This exceptional capability not only underscores its impressive performance across multiple tasks but also positions it as a transformative tool for bridging the gap between visual and textual information.

CLIPN (CLIP Saying "No") [40] is a derivative model stemming from CLIP, mirrors its predecessor in terms of structure and training approaches. Its distinctive feature, however, lies in a specialized functionality. By modifying the prompt template, CLIPN alters the dynamic between image and text, offering insights into the likelihood that a certain sample falls outside a specific class. Moreover, CLIPN's capacity to alter the dynamic between image and text through prompt modifications has far-reaching implications. It allows users to gain insights into the likelihood that a given sample falls outside a specific predefined class. This is immensely valuable in scenarios where anomaly detection or outlier identification is essential. CLIPN can provide valuable indications of when an image-text pair deviates from the norm, aiding in tasks such as fraud detection, quality control, or security monitoring.

DALL-E (Diverse All-scale and Latent Length Encoder) represents a cutting-edge generative model developed by OpenAI, symbolizing a significant milestone in the field of artificial intelligence. This innovative model seamlessly merges the domains of natural language comprehension and image generation, enabling it to produce highly intricate images directly from textual descriptions. The architecture of the DALL-E model can be deconstructed into two fundamental components: an encoder and a decoder. The encoder assumes the role of skillfully processing natural language descriptions, transforming them into latent vectors. During this process, it captures the intricate semantic nuances embedded within the input text. This encoding process entails converting textual input into a numerical representation comprehensible to the model. To execute this task effectively, the encoder frequently employs advanced techniques such as transformers or recurrent neural networks, facilitating the extraction of meaningful and context-rich features from the text. Conversely, the decoder component shoulders the responsibility of utilizing these encoded latent vectors to generate corresponding images that faithfully mirror the essence of the input textual description. The decoder harnesses the capabilities of influential generative models like autoregressive models and transformers to decode the latent vectors into visually coherent, diverse, and expressive images. By adeptly conditioning the image generation process on the encoded textual context, DALL-E exhibits the remarkable ability to generate images that seamlessly align with the provided textual descriptions. DALL-E has not only demonstrated remarkable prowess but has also emerged as a crucial bridge connecting the domains of natural language processing and computer vision. This bridge empowers users to effortlessly generate a wide spectrum of diverse and lifelike images solely based on textual input. Consequently, DALL-E finds applications across a multitude of domains, encompassing content creation, virtual reality, creative design, and even in assisting artists and designers in translating their abstract concepts into visual reality. In summary, the DALL-E model can be likened to a symphony, composed of an encoder and a decoder, orchestrated to craft images from textual descriptions. It leverages state-of-the-art techniques in language comprehension and image synthesis to conjure visually coherent and diverse images that faithfully encapsulate the essence of the provided text. DALL-E's ability to transmute textual prompts into stunning visual outputs holds immense potential across various creative and practical domains.

7.2. Contrastive Learning

Since its inception, contrastive learning has garnered significant attention within the fields of machine learning and computer vision [7–9, 22]. At its core, this approach centers on the concept of identifying compatible positive and negative samples and calculating their loss functions. The primary objective is to minimize the distance between anchor samples and their corresponding positives while simultaneously maximizing the distance between anchor samples and negatives. The selection of positive and negative samples can be achieved through a variety of methods, offering adaptability to suit specific requirements. For example, one can opt for label-based selection [24], where samples with identical labels are designated as positives, while the remainder serve as negatives. Alternatively, positive samples can be generated through rotations and translations of anchor samples, with all other samples designated as negatives [7]. Researchers also have the flexibility to devise custom selection criteria tailored to the demands of the task or, in some cases, forgo the use of negative samples entirely [18]. Contrastive learning strategies have found applications across a broad spectrum of downstream tasks, spanning from image classification to object detection and retrieval. Its versatility and adaptability render it a valuable tool in the realm of machine learning research. Particularly noteworthy is contrastive learning's significant contributions to the field of few-shot learning [13, 20, 28, 30, 42]. In our paper, we leverage the Direct-and-Inverse concept, harnessing models such as CLIP and CLIPN to precisely extract positive and negative samples. This innovative approach has propelled few-shot learning to new heights, addressing the challenges associated with limited data. In summary, contrastive learning has emerged as a potent paradigm in machine learning and computer vision, providing a flexible and effective methodology for sample selection and loss calculation. Its impact extends across a wide range of applications, with its remarkable contributions to few-shot learning standing out as a testament to its significance in addressing complex and challenging problems in the field.

8. Experiments

8.1. Additional Performance Comparison

Fig. 7 presents a comprehensive analysis of the experimental outcomes for other five datasets, including EuroSAT [23], FGVC [31], Flower102 [32], StanfordCars [26], and UCF101 [39]. The results can be categorized into two components: the upper part of the figure illustrates performances under 1-shot conditions, with variations in noisy label proportions. Conversely, the lower part focuses on outcomes where a consistent noisy label proportion of 0.3 is maintained across various few-shot scenarios. These experimental findings across datasets are in alignment with results from six other datasets, further solidifying the claim that our method outperforms others in terms of both performance and stability.

Furthermore, as depicted in Fig. 8, the learning progress over 50 epochs is meticulously visualized. It captures both the training loss and test accuracy within the context of the 1-shot ImageNet scenario. These graphical representations vividly illustrate the rapid convergence of our method.

Additionally, in contrast to our previous experiments where ResNet50 served as the backbone architecture for CLIP, this study employs DeIL with a variety of visual encoders for comparisons with other methods. As depicted in Table 5, DeIL consistently showcases superior performance across different visual backbones. Particularly, when utilizing the ViT-B/16 backbone, this phenomenon becomes particularly pronounced. These observations underscore the versatility and generalizability of our approach across various network architectures.

8.2. Additional Ablation Study

DeIL-Pretrainer In the original text, Equation (4) introduces a critical parameter denoted as ϵ, which plays a pivotal role in determining the performance of the initial step of the DeIL-Pretrainer. To assess the impact of this parameter on the results, we conducted an ablation experiment. In this experiment, we consider a scenario with 1,000 images and their corresponding labels. Among these, 500 images have correct labels, while the remaining 500 are associated with incorrect labels. The objective is to examine how accuracy changes concerning whether a sample does not belong to a specific class under different values of ϵ. It is imperative to note that the setting of ϵ should be greater than 0.5 for the following reasons: Let's assume the predicted value is p representing the probability of a sample not belonging to class A, while $(1 - p)$ represents the probability of a sample belonging to class A. When the predicted value p falls below 0.5, it suggests that the likelihood of the sample belonging to class A outweighs the likelihood of it not belonging to class A. Therefore, to achieve the purpose of the original text, we set the threshold ϵ to a value greater than

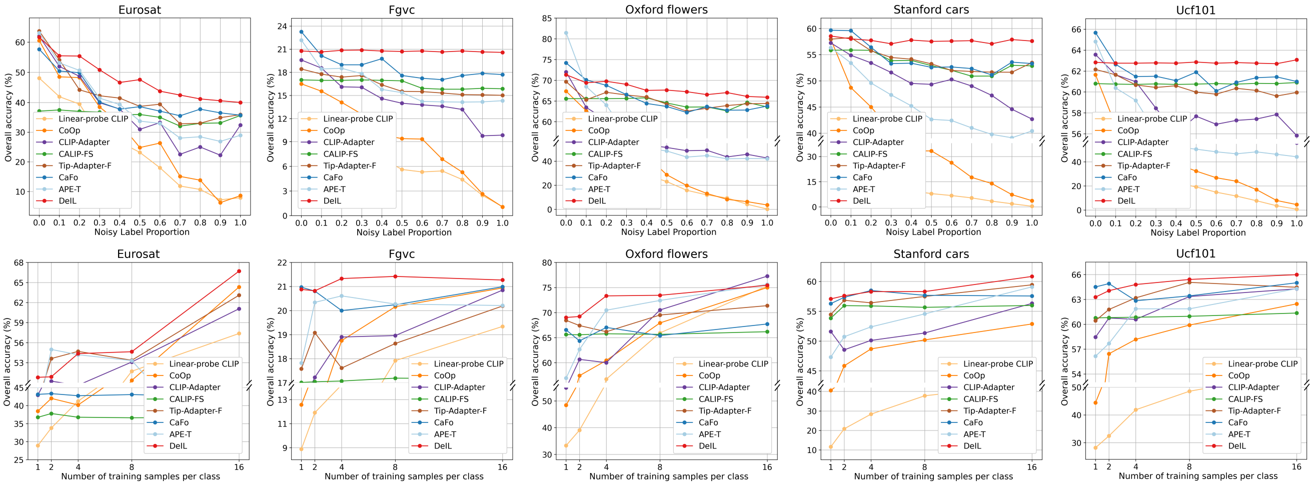

Figure 7. Performance (%) comparison on other datasets. The upper column presents the results under 1-shot conditions with different noisy label proportions, while the lower column presents the results with a fixed noisy label proportion of 0.3 on varying few-shot settings.

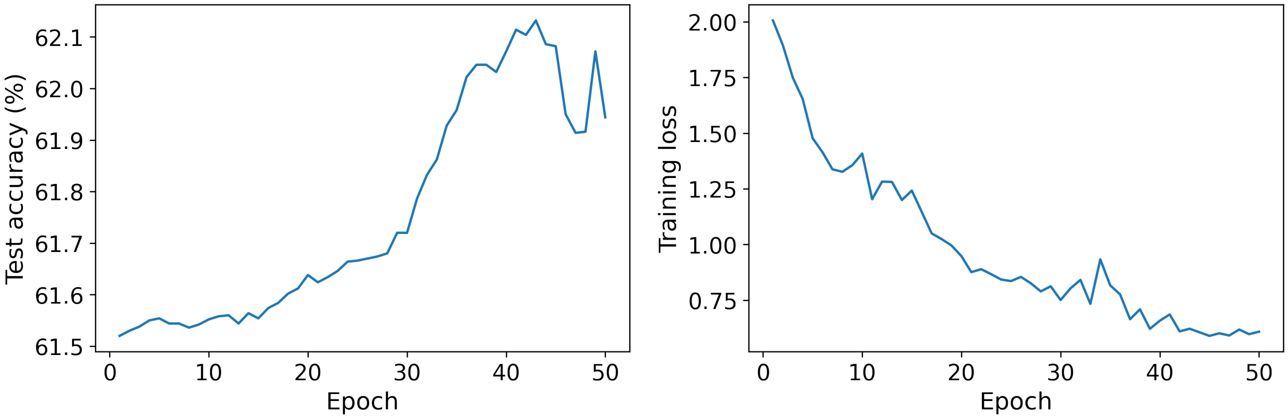

Figure 8. The learning curve of training loss and test accuracy (%) of our DeIL on 1-shot ImageNet. The noisy label proportion is fixed at 0.3.

0.5. The results of our experimentation are summarized in Tab. 6. Upon careful examination, it becomes evident that when ϵ is set to 0.5, optimal results are achieved.

Additionally, we present the predicted values obtained through the inverse-concept method in Fig. 9. These values represent the probability distribution of 1,000 samples not belonging to 1,000 categories. The x-axis corresponds to $1,000 \times 1,000$ values, while the y-axis represents the predicted probability values. The figure illustrates the results with one data point selected for every 200 points. It is noteworthy that a significant majority of values tend to cluster around either 0 or 1, indicating a high degree of confidence in most predictions made by the method.

DeIL-Adapter We first discuss the $loss_{cls}$. In the original text, Equations (15) and (16) highlight the significant impact of the parameters α and β on the results. In this context, we conduct experiments to assess the consequences of varying these parameters, and the findings are summarized in Tab. 7. Our primary objective through this extensive examination and fine-tuning of hyperparameters is to unravel the intricate relationship between alpha, beta, and the resulting loss values. Ultimately, this endeavor aims to shed light on how these parameters exert their influence on the overall outcomes of our computations. This meticulous analysis yields valuable insights that can be instrumental in optimizing the performance and accuracy of our model. It is worth noting that these experiments were conducted using the 1-shot ImageNet with fixed values for the noisy label proportion (0.3) and training epochs (40).

Then, let's dive into a detailed discussion of $loss_{nce}$. As evident from Equation (25) in the original text, the param-

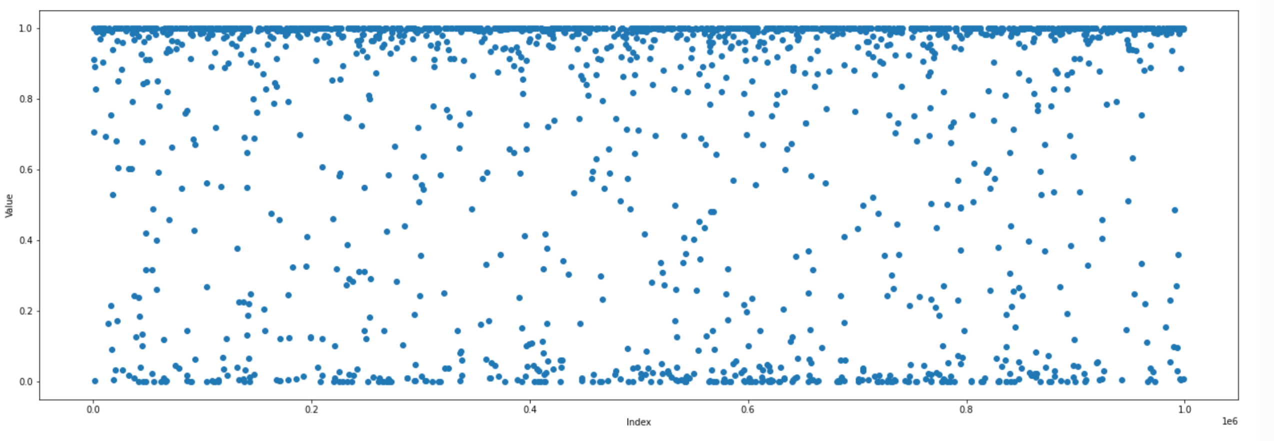

Figure 9. Probability distribution of 1000 samples not belonging to 1000 categories. The x-axis corresponds to 1000 × 1000 values, while the y-axis represents the predicted probability values. The figure illustrates the results with one data point sampled for every 200 points.

Methods	**Backbones**			
	RN50	RN101	ViT-B/32	ViT-B/16
Linear-probe CLIP (ICML'21) [35]	13.98	16.18	16.87	19.69
CoOp (IJCV'22) [49]	56.58	55.23	53.57	61.54
Tip-Adapter-F (ECCV'22) [47]	59.97	62.62	62.90	66.91
CLIP-Adapter (IJCV'23) [17]	55.19	55.03	52.69	58.24
CALIP-FS (AAAI'23) [19]	56.86	61.29	61.89	66.59
CaFo (CVPR'23) [48]	59.99	62.03	63.12	67.03
APE-T (ICCV'23) [50]	51.02	54.34	57.92	60.63
DeIL (Ours)	**62.28**	**63.09**	**65.62**	**70.90**

Table 5. Ablation Study (%) of CLIP's Visual Encoders. We conduct different visual backbones on the 1-shot ImageNet. The noisy label proportion is fixed at 0.3.

Method	ϵ				
	0.5	0.6	0.7	0.8	0.9
Noisy Label Identification	**87.0**	85.6	84.3	83.0	80.6

Table 6. Examining how accuracy (%) changes concerning *whether a sample does not belong to a specific class* under different values of ϵ on ImageNet. There are 1,000 images and their corresponding labels. Among these, 500 images have correct labels, while the remaining 500 are associated with incorrect labels. The reason for commencing from 0.5 is further elaborated in Sec. 8.2.

α	β				
	0.1	0.3	0.5	0.7	0.9
0.1	61.52	61.70	62.00	61.97	62.14
0.3	61.91	62.65	**62.75**	62.55	62.28
0.5	61.61	62.11	62.15	62.19	62.51
0.7	61.69	62.45	62.38	62.35	62.56
0.9	61.76	62.58	62.50	62.53	62.40

Table 7. Comparison results (%) of different hyperparameter (α and β) on ImageNet with 1-shot case. The noisy label proportion is fixed at 0.3. The training epoch is fixed at 40.

eter γ assumes a critical role in governing the behavior of $loss_{nce}$. To be more precise, the magnitude of γ dictates the extent to which the model relies on label information in the loss calculation. In light of this, we embark on an exploration of the impact of varying γ values through a series of experiments. The comprehensive results of these experiments are meticulously documented in Tab. 8.

DALL-E Next, we delve into the impact of the quantity of enhanced images on the results, with corresponding outcomes presented in Tab. 9. We observe that increasing the number of generated samples tends to result in enhanced performance. However, it's crucial to note that an excessive increase in the sample quantity can significantly prolong the training time. In the main text, we provide results based on an 8-dalle-shot configuration, striking a balance

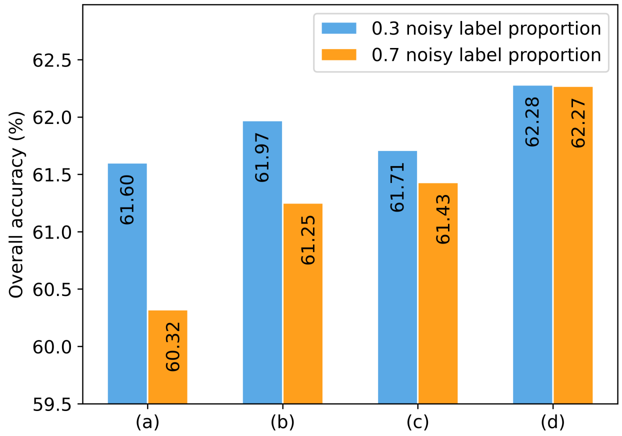

Figure 10. Efficiency of Direct-and-Inverse concept on DeIL-Pretrainer. (a) represents the selection of positive and negative samples based on the provided noisy labels. (b) indicates the selection of positive samples based on the Direct concept and negative samples based on the provided noisy labels. (c) signifies the selection of positive samples based on the provided noisy labels and negative samples based on the Inverse concept. (d) encompasses the selection of positive and negative samples based on the Direct-and-Inverse concept.

Method	γ				
	10^0	10^{-1}	10^{-2}	10^{-3}	0
DeIL	61.96	**62.28**	61.75	61.75	61.50

Table 8. Comparison results (%) of different hyperparameter (γ) on ImageNet with 1-shot case. The noisy label proportion is fixed at 0.3. The training epoch is fixed at 40.

between improved effectiveness and maintaining a manageable training duration. This equilibrium ensures a practical and efficient approach to achieving desirable outcomes. Additionally, we offer a visual representation of the generated images in Fig. 11.

8.3. Efficiency of Direct-and-Inverse Concept

In the original text, we discussed the significance of the Direct-and-Inverse Concept in the DeIL-pretrainer. Now, we delve deeper into its role in the DeIL-Adapter. The experimental results are depicted in Fig. 10. Upon observation, it becomes evident that our method has a positive impact, with its effectiveness becoming more pronounced as the ratio of label noise increases.

8.4. Visualization of Direct-and-Inverse Templates

Lastly, we present visualizations of the inverse and direct templates in Fig. 12.

DALL-E Shots	**Noisy Label**				
	0.1	0.3	0.5	0.7	0.9
1	61.36	61.18	61.15	61.11	61.06
2	61.56	61.23	61.14	61.10	61.12
4	62.20	61.80	61.93	61.87	61.94
8	62.37	62.28	62.42	62.27	62.11
16	**62.66**	**62.62**	**62.50**	**62.57**	**62.46**

Table 9. Ablation Study (%) of generated number via DALL-E. We compare different noisy label proportions on 1-shot ImageNet.

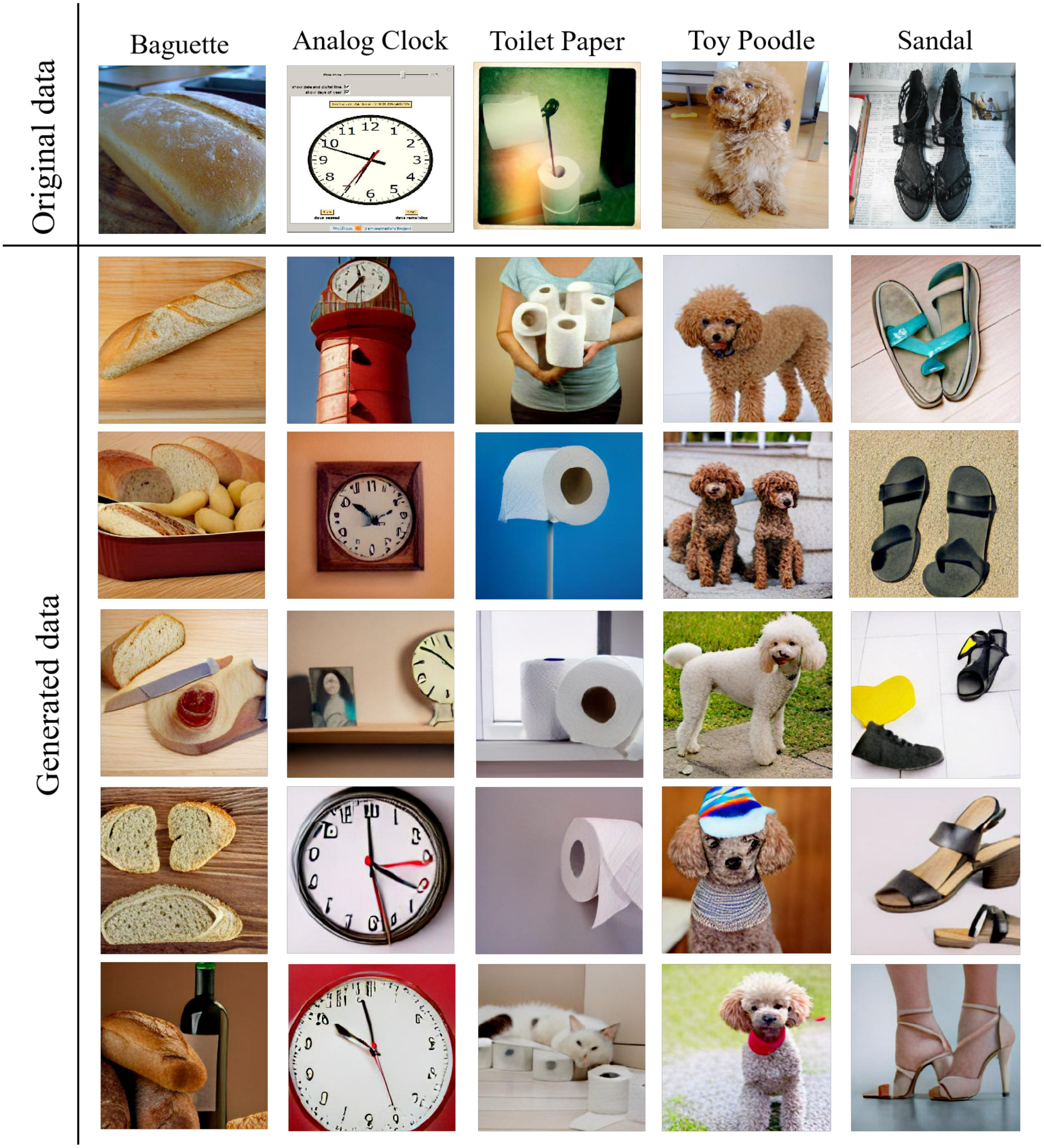

Figure 11. Visualizations of DALL-E's generated images. Examples are from ImageNet.

"a bad photo of no {tench}."
"a photo of no many {tench}."
"a sculpture of no {tench}."
"a low resolution photo of no {tench}."
"a rendering of no {tench}."
"graffiti of no {tench}."
"a cropped photo of no {tench}."
"a tattoo of no {tench}."
"a bright photo of no {tench}."
"a dark photo of no {tench}."
"a drawing of no {tench}."
"a photo of no {tench}."
"a close-up photo of no {tench}."
"a black and white photo of no {tench}."
"a painting of no {tench}."
"a sculpture of no {tench}."

(a) Inverse Template

“a {tench} is a freshwater fish of the carp family."
"a {tench} in a river."
“a {tench} is a freshwater fish found in Europe."
“a {tench} is a freshwater fish of the family Cyprinidae."
“a {tench} is a freshwater fish of the carp family."
“the image is of a {tench} fish swimming in a pond.",
"A {tench} in a fishpond."
“a {tench} is a small freshwater fish in the carp family."
“a tench is a freshwater fish in the carp family.",
“a {tench} is a freshwater fish that looks similar to a carp."
“a {tench} is a freshwater fish in the carp family."
“the {tench} is a fresh-water fish in the family Cyprinidae."
"{tench} are a freshwater fish found in Europe."
“the image is of a {tench} fish."
“the {tench} is a freshwater fish belonging to the carp family."
“a {tench} is a freshwater fish of the Cynoglossidae family."

(b) Direct Template

Figure 12. Visualization of Direct-and-Inverse templates.