Supplementary Material of Hierarchical Divide-and-Conquer Grouping for Classification Adaptation of Pre-Trained Models

A. Detailed Information of Hierarchical Grouping

As mentioned in Section 3, we progressively divide the unified test space into hierarchical ones. The whole grouping process is given in Algorithm. 1. In addition, we also explore the optimal depth of the proposed hierarchical grouping structure. Specifically, we not only change the number of separation with balanced way (*i.e.,* keep the same depth for base and novel branches) as shown in Table. 1 of main paper, we also evaluate the unbalanced separation in Table. b. Intuitively, when we adopt two step *i.e.,* grouping for novel branch of AwA2, we observe a slight degradation. Therefore, we speculate that the optimal depth for each datasets or different branches is inconsistent. In fact, we can search the optimal best plan by greedy search. As we experimentally explore the performance of popular benchmarks, we set the depth of each dataset by 2 to maintain the simplicity of the method.

In Fig. a, we report the distributions of $\mathrm{score}^{\mathrm{d}}$, $\mathrm{score}^{\mathrm{b}}$ and $\mathrm{score}^{\mathrm{n}}$, respectively. Notably, the first column highlights a strikingly low fractional coincidence across different domains, underscoring the remarkable ability of our method to effectively segregate the base and novel domains within a straightforward and training-free framework.

B. Contribution of Each Strategy

In order to better explore what strategies can bring significant gains, we provide Table. a. It can be seen that after the information of VLM was introduced into f-VAEGAN and TF-VAEGAN, there was a significant improvement in the base classes, but it is accompanied by a serious degradation of novel performance. This indicates that directly applying VLM to the original zero-shot structure is not appropriate, as it will exacerbate prediction bias.When we compare the proposed HDG-PSVMA method, we observe significant improvements across all datasets, which confirms the effectiveness of our proposed grouping strategy even under the strict zero-shot setting. Furthermore, compared with CLIP, our method performs particularly outstanding in terms of improving the performance on novel classes. This indicates that the hierarchical divide-and-conquer grouping indeed broke the prior data dependency problem in the VLMs adaptation process and improve the model's generalization ability.

C. LLMs-guided descriptions

Examples for LLMs-guided Descriptions. To enhance the diversity of the generation samples, we propose a LLMs-guided method to generate various descriptions for novel classes. Here, we provide some examples which are obtained by GPT-4 on AwA2:

bat:

The bat hung upside down from the cave ceiling, its wings folded neatly around its body.

The bat flew gracefully through the night sky, silhouetted against the full moon.

The bat clung to the tree trunk, its sharp claws digging into the bark.

The bat's eyes gleamed in the darkness, reflecting the light from the distant streetlamp.

The bat swooped low over the field, its keen ears detecting the flutter of insect wings.

The bat nestled among the leaves, hidden from predators by its mottled brown fur.

The bat spread its wings wide, revealing the delicate webbing between its fingers.

The bat darted through the air, executing sharp turns to avoid obstacles.

The bat emitted a series of high-pitched squeaks, using echolocation to navigate.

The bat landed softly on the window ledge, its nose twitching as it sniffed the air.

blue whale:

The blue whale surfaced slowly, sending a towering spout of mist into the air.

The blue whale glided gracefully through the ocean, its massive body cutting through the water with ease.

The blue whale breached, its enormous body rising out of the water before crashing back down with a thunderous

Methods	Backbone	AwA2			CUB			SUN		
		B	N	H	B	N	H	B	N	H
f-VAEGAN [7]	Res-101	70.6	57.6	63.5	60.1	48.4	53.6	38.0	45.1	41.3
F-VAE+SHIP [6]	VLMs	95.9$_{+25.3}$	61.2$_{+3.6}$	74.7$_{+11.2}$	82.2$_{+22.1}$	22.5$_{-25.9}$	35.3$_{-18.3}$	-	-	-
TF-VAEGAN [3]	Res-101	75.1	59.8	66.6	64.7	52.8	58.1	40.7	45.6	43.0
TF-VAE+SHIP [6]	VLMs	96.3$_{+21.2}$	43.7$_{-16.1}$	60.1$_{-6.5}$	84.4$_{+19.7}$	21.1$_{-31.7}$	34.0$_{-24.1}$	-	-	-
PSVMA [2]	ViT-Base	77.3	73.6	75.4	77.8	70.1	73.8	45.3	61.7	52.3
HDG-PSVMA (Depth=1, Frozen)	Vit-Base	78.9	75.3	77.1	78.0	72.5	75.1	49.2	63.9	55.6
HDG-PSVMA (Depth=1, Tuning)	Vit-Base	79.1$_{+1.8}$	77.2$_{+3.6}$	78.1$_{+2.7}$	78.3$_{+0.5}$	74.4$_{+4.3}$	76.3$_{+2.5}$	55.6$_{+10.3}$	64.4$_{+2.7}$	59.7$_{+7.4}$
CLIP [5]	VLMs	92.9	86.6	89.6	55.1	54.9	55.0	40.2	49.4	44.3
HDG-CLIP (Depth=1, Frozen)	VLMs	93.0	90.2	91.6	57.1	71.5	63.5	45.7	66.1	54.0
HDG-CLIP (Depth=1, Tuning)	VLMs	93.9	94.2	94.0	73.8	75.2	74.5	79.8	71.2	75.3
HDG-CLIP (Depth=2, Tuning)	VLMs	94.5$_{+1.4}$	94.1$_{+7.5}$	94.3$_{+4.7}$	78.4$_{+23.3}$	78.0$_{+23.1}$	78.2$_{+23.2}$	81.4$_{+41.2}$	73.3$_{+23.9}$	77.1$_{+32.8}$

Table a. GZSC performance (%) comparisons on three benchmarks. “Frozen” indicates that no fine-tuning is performed on the subspace. Depth=1 refers to inter-class grouping. Depth=2,3 refers to recursive intra-class grouping. Gain or degradation refers to the performance relative to same variant. Our method is marked with a blue background color

Algorithm 1 Hierarchical Grouping Strategy

```
Input: $x_i \in \mathcal{X}_{base}$: Base Image Sets; $\widetilde{x}_i \in \mathcal{G}$: Novel Generation Image Sets; $x_j \in \mathcal{X}_{test}$: Test Image Sets; $\mathcal{T}$ : $Threshold$
Output: Fine-grained class labels: $Base1, Base2$ and $Novel1, Novel2$
for each $x_j$ in $\mathcal{X}_{test}$:
repeat
    compute multi-modal distance in Eq.2 and then compute the scores of $score^d$, $score^b$ in Eq.3 and Eq.6, respectively;
    while $j < max\ number\ of\ test\ images$ and $x_j$ has not been grouped do
        if $score^d \geq \mathcal{T}$ then
            assign coarse-grained stage label Base to $x_j$
            if $score^b \geq 0$ then
                assign fine-grained stage label Base1 to $x_j$
            else
                assign fine-grained stage label Base2 to $x_j$
            end if
        else
            assign coarse-grained stage label Novel to $x_j$
            if $score^n \geq 0$ then
                assign fine-grained stage label Novel1 to $x_j$
            else
                assign fine-grained stage label Novel2 to $x_j$
            end if
            stop
```

splash.

The blue whale’s tail fluke emerged from the sea as it prepared to dive deep into the abyss.

The blue whale swam alongside a pod of dolphins, dwarfing them with its immense size.

The blue whale’s mouth opened wide, filtering vast amounts of krill through its baleen plates.

The blue whale’s eye, small in comparison to its body, scanned the ocean depths.

The blue whale floated near the surface, its smooth blue-gray skin glistening in the sunlight.

The blue whale’s call echoed through the water, a deep, resonant sound that could travel for miles.

The blue whale moved slowly, conserving energy as it navigated the cold, nutrient-rich waters.

bobcat

The bobcat crouched low in the tall grass, its eyes fixed on its prey.

The bobcat leapt gracefully onto a rock, scanning the area for any signs of movement.

The bobcat’s ears twitched, picking up the faint rustle of leaves in the wind.

The bobcat padded silently through the forest, its fur blending seamlessly with the underbrush.

The bobcat snarled, revealing sharp teeth as a warning to intruders.

Table b. Effects of the proposed grouping strategy depth. Depth=2 for Base Only denotes that we conduct two stage grouping for base branch and one stage for novel classes.

Methods	**AwA2**			**CUB**		
	B	**N**	**H**	**B**	**N**	**H**
Depth=1	93.9	94.2	94.0	73.8	75.2	74.5
Depth=2 for Base and Novel	94.5	94.1	94.3	78.4	78.0	78.2
Depth=2 for Base Only	94.5	94.2	94.3	78.4	75.2	76.8
Depth=2 for Novel Only	93.9	94.1	94.0	73.8	78.0	76.2

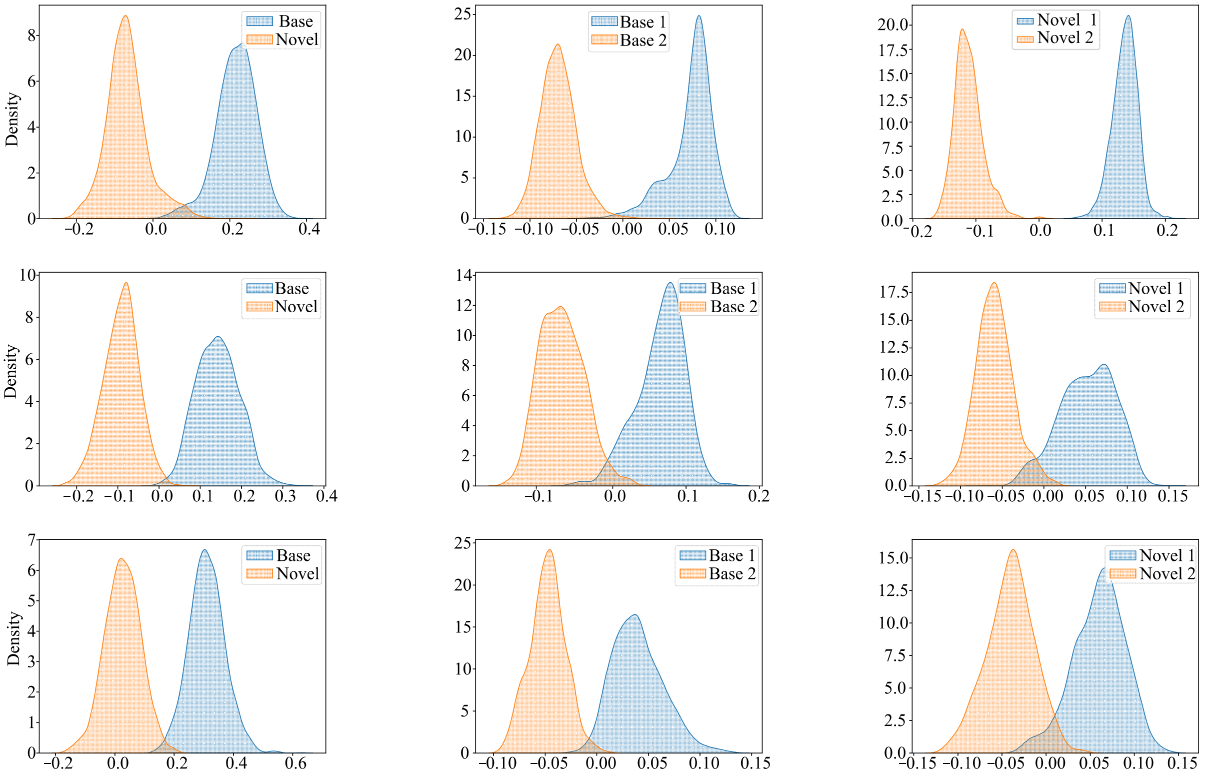

Figure a. Supplementary analysis of Figure 3 in the main paper. From top to bottom, we report the density maps of score^{d} (first column), score^{b} (second column), and score^{n} (third column) on AwA2, CUB and SUN, respectively.

The bobcat's tail flicked back and forth as it prepared to pounce.

The bobcat perched on a tree branch, surveying the ground below for any potential meals.

The bobcat stretched out in a patch of sunlight, enjoying the warmth on its fur.

The bobcat's keen eyes caught the glint of a rabbit's fur in the moonlight.

The bobcat slinked through the shadows, moving with a predator's stealth and precision.

dolphin:

The dolphin leapt out of the water, performing a graceful arc before splashing back down.

The dolphin swam alongside the boat, its dorsal fin cutting through the waves.

The dolphin's playful nature was evident as it chased after a school of fish.

The dolphin's sleek, gray body glistened in the sunlight as it rode the surf.

The dolphin communicated with its pod using a series of clicks and whistles.

The dolphin flipped its tail energetically, propelling itself through the water.

The dolphin surfaced for air, its blowhole emitting a

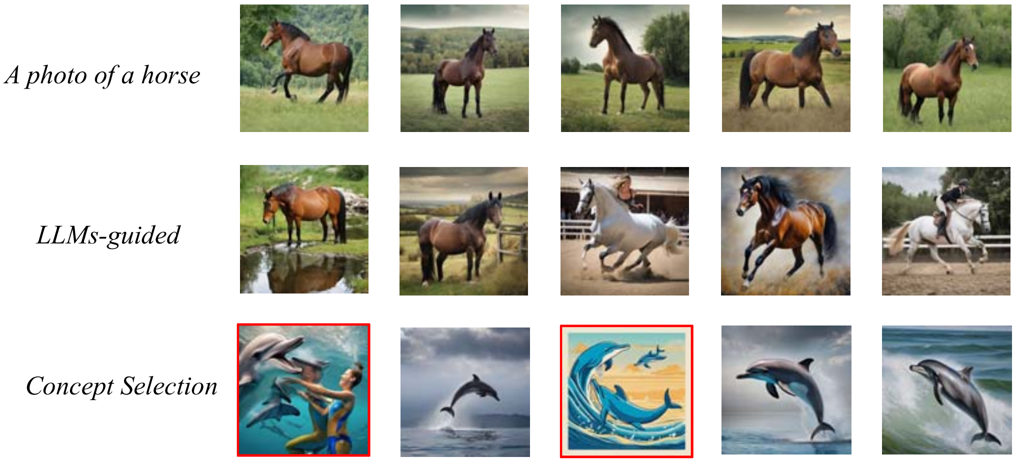

Figure b. Examples of different descriptions.Top: The generation samples which are generated by A photo of a horse. Middle: The generation samples which are generated by LLMs-guided descriptions. Bottom: The concept selection process. The red boxes represent samples that were discarded.

quick burst of mist.

The dolphin interacted with swimmers, gently nudging them with its snout.

The dolphin performed acrobatics, delighting the onlookers with its agility.

The dolphin's intelligent eyes observed the divers curiously as they explored the reef.

D. Details and Examples for Generation

In addition, we provide some examples in Fig. b. By comparing the first row and second row, we can observe that the generation samples equipped with LLMs-guided strategy have better diversity than the others. For instance, the horses in the second row have more colors, poses and backgrounds, and are closer to the state of the horse in nature. This indicates that the proposed LLMs-guided strategy ensures diversity while not sacrificing semantic relevance. Further, we visualize the process of concept selection at the bottom of Fig. b. We can see that the discarded samples have obviously different distribution than preserved ones. As for GPU and time utilization, we conduct all the generation processes with Stable-Diffusion-XL [4] on Nvidia RTX 4070Ti and the model generates an 1024*1024 image about every 15 seconds.

E. Details of Normalized Mutual Information

The Normalized Mutual Information (NMI) [1] is commonly used in clustering evaluation to measure the similarity between two clustering results. Formally, we have:

$$\mathrm{NMI}(X, Y) = \frac{2 \times I(X;Y)}{H(X) + H(Y)}, \tag{1}$$

where $I(X;Y)$ refers to Mutual information between X and Y, $H(X)$ and $H(Y)$ represent the entropy of X and Y, respectively. In this paper, we first assign labels of 0 to the base images (training set) $\mathcal{X}_{base}$ and 1 to the diffusion model generated images $\mathcal{G}$. Subsequently, we input and predict all the base images into hypothetical classes using Eq.3 (in main paper). By evaluating the NMI between the ground truth labels (i.e., $H(X)$) and the predicted labels (i.e., $H(Y)$) across various threshold values (T), we can quantify the extent of information about the true class labels captured by the specific clustering or grouping outcomes.

References

[1] Pablo A Estévez, Michel Tesmer, Claudio A Perez, and Jacek M Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2):189–201, 2009.

[2] Man Liu, Feng Li, Chunjie Zhang, Yunchao Wei, Huihui Bai, and Yao Zhao. Progressive semantic-visual mutual adaption for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15337–15346, 2023.

[3] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *European Conferenceon Computer Vision (ECCV)*, pages 479–495. Springer, 2020.

[4] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[6] Zhengbo Wang, Jian Liang, Ran He, Nan Xu, Zilei Wang, and Tieniu Tan. Improving zero-shot generalization for clip with synthesized prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3032–3042, 2023.

[7] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10275–10284, 2019.