SG-LoRA: Semantic-guided LoRA Parameters Generation

Supplementary Material

Appendix Overview

The supplementary material is organized into the following sections:

A. Experimental Details

A.1. Image-text retrieval dataset

In this work, we generate fine-grained captions for the image-text retrieval task using Qwen2-VL. As illustrated in Figure. 6, our in-context learning approach provides the Multimodal Large Language Model (MLLM) with a small number of demonstration examples, enabling it to generate detailed captions for target images. Using this method, we have produced four diverse and highly relevant captions for each image in the entire OxfordPets dataset and the Flowers102 dataset, a subset of the MS-COCO dataset. These fine-grained descriptions serve as a valuable resource for downstream tasks such as fine-grained image-text retrieval and facilitate further research in ZSOA.

As shown in Figure. 7, we present examples of image-text pairs from three datasets. Compared to the original captions in the COCO dataset (labeled as *Original Caption* in the top), our generated captions more accurately reflect the content of the corresponding images. Thus, these datasets enable us to assess the model's robustness both in in-domain retrieval and in generalizing to novel domains and compositional scenarios.

A.2. LoRA Expert Repository Details

We present the expert tasks included in each dataset's expert repository in Table. 6, as set in the main manuscript.

Table 6. Selected Expert Tasks for Each Datasets

MS-COCO	OxfordPets	Flowers102
Airplane	*Abyssinian*	*Sweet pea*
Truck	*American bulldog*	*Tiger lily*
Traffic Light	*American Pit Bull Terrier*	*Monkshood*
Cat	*Birman*	*King Protea*
Horse	*Bombay*	*Corn Poppy*
Giraffe	*British Shorthair*	*Daffodil*
Handbag	*German Shorthaired*	*Sunflower*
Snowboard	*Havanese*	*Osteospermum*
Wine Glass	*Keeshond*	*Anthurium*
Banana	*Leonberger*	*Hibiscus*
Hot Dog	*Newfoundland*	*Desert-Rose*
Laptop	*Scottish Terrier*	*Mallow*

B. Aggregation function for Sematic Prior

We begin with the classical identity in probability theory known as the Law of Total Variance: for a random vector Θ and a discrete conditioning variable C , the following identity holds:

$$\mathrm{Var}(\Theta) = \mathbb{E}_C\big[\mathrm{Var}(\Theta \mid C)\big] + \mathrm{Var}_C\big(\mathbb{E}[\Theta \mid C]\big), \quad (10)$$

Proof Sketch: Using the identity $\mathrm{Var}(\Theta) = \mathbb{E}[\Theta^2] - (\mathbb{E}[\Theta])^2$ and the Law of Iterated Expectations, we write:

$$\mathbb{E}[\Theta^2] = \mathbb{E}_C\big[\mathbb{E}[\Theta^2 \mid C]\big], \quad \mathbb{E}[\Theta] = \mathbb{E}_C\big[\mathbb{E}[\Theta \mid C]\big], \quad (11)$$

From the definition of conditional variance, $\mathrm{Var}(\Theta \mid C) = \mathbb{E}[\Theta^2 \mid C] - (\mathbb{E}[\Theta \mid C])^2$, we substitute and obtain:

$$\mathrm{Var}(\Theta) = \mathbb{E}_C\big[\mathrm{Var}(\Theta \mid C)\big] + \mathbb{E}_C[(\mathbb{E}[\Theta \mid C])^2] - (\mathbb{E}_C[\mathbb{E}[\Theta \mid C]])^2, \quad (12)$$

where the last two terms equal $\mathrm{Var}_C(\mathbb{E}[\Theta \mid C])$.

Task Semantic via Aggregation. Suppose we have K expert tasks, each providing a LoRA parameter set with mean $\boldsymbol{\mu}_i$ and variance $\boldsymbol{\sigma}_i^2$. For an unseen task $\mathcal{T}^*$ with expert weight vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k]^T$, where $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$, we construct the prior as a weighted combination. The prior mean is:

$$\boldsymbol{\mu}^* = \sum_{i=1}^{k} \alpha_i \boldsymbol{\mu}_i, \quad (13)$$

Applying the Law of Total Variance, the element-wise prior variance for $\mathcal{T}^*$ is:

$$\boldsymbol{\sigma}^{2^*} = \sum_{k=1}^{K} \alpha_k \boldsymbol{\sigma}_k^2 + \sum_{k=1}^{K} \alpha_k (\boldsymbol{\mu}_k - \boldsymbol{\mu}^*) \odot (\boldsymbol{\mu}_k - \boldsymbol{\mu}^*), \quad (14)$$

[Instruction]
You are an image description assistant. Please analyze the given image carefully and provide a detailed description of its contents.

[Demonstration 1]
Image: <image_path1>
Input: A photo of pug.
Output: The pug has a fawn coat, black mask, and wrinkled face with a short-muzzled expression. It wears a green collar with a tag, sits on a rock, and has a solemn gaze.

[Demonstration 2]
Image: <image_path2>
Input: A photo of birman.
Output: The Birman cat has a sleek cream and brown coat, striking blue eyes, and a dark brown nose. It wears a blue collar and is perched on a scratching post.

[Target Query]
Image: <image_path>
Input: A photo of {class name}.
*Output: **fine-grained caption for the target image***

Figure 6. Examples of caption generation using Qwen2-VL.

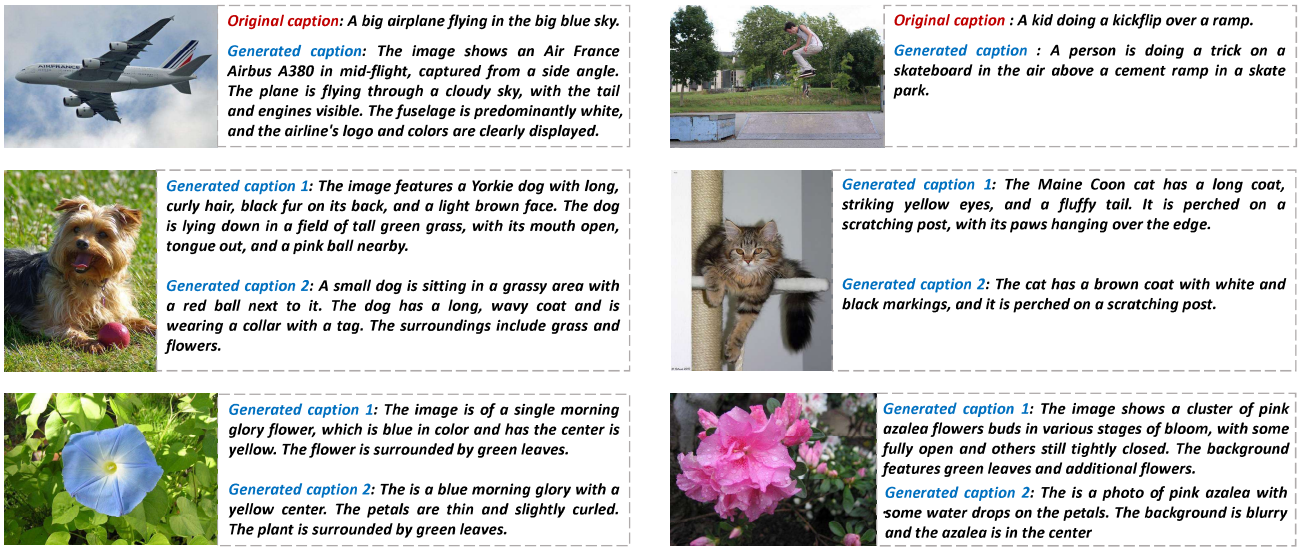

Figure 7. Illustration of generated captions: results on the MS-COCO (top), OxfordPets (middle), and Flowers102 (bottom) datasets.

where $\odot$ denotes element-wise multiplication. In Eq.14, the first term, which represents the expectation of the conditional variances (computed as a weighted average of the per-task variances), quantifies the uncertainty within each expert task. The second term, which is the variance of the conditional means, quantifies the dispersion or uncertainty among different expert tasks. Eq.13 and Eq.14 together define the task semantic, providing a more informative prior than modeling the parameters as a standard Gaussian, as it ensures that both intra-task variation and inter-task uncertainty are captured when modeling the latent prior for each unseen task.

C. Additional Results on Cross-dataset Image-text Retrieval

As shown in Table. 7, we present cross-dataset evaluation with the model trained on MS-COCO and tested on Flowers102. It can be observed that when there exists a noticeable gap between the expert task and the inference task, although the performance generally surpasses Zero-Shot CLIP, both AdapterSoup and Top-K LoRA Weighted fail to significantly outperform Model Soup. In contrast, the semantics-guided SG-LoRA demonstrates a strong ability in generating high-performance LoRA parameters.

Table 7. Cross-dataset evaluation with the model trained on MS-COCO and tested on Flowers102. The best results are highlighted in bold, and the second-best results are underlined.

Method	I2T Metrics			T2I Metrics		
	R@1	R@5	R@10	R@1	R@5	R@10
Zero-Shot CLIP	22.30	53.57	69.41	16.33	46.79	68.49
Oracle	26.23	59.17	77.55	21.57	57.56	80.13
Model Soups	23.50	<u>54.84</u>	<u>73.85</u>	17.23	<u>50.96</u>	73.48
AdapterSoup	<u>24.21</u>	54.26	72.83	<u>17.76</u>	50.93	<u>73.83</u>
Top-*K* LoRA Weighted	23.98	52.83	71.91	17.63	50.69	73.74
SG-LoRA	**26.83**	**56.63**	**74.16**	**20.52**	**53.71**	**76.69**

D. Additional Results on Classification Task

In addition to the image-text retrieval task, we also explored the performance of SG-LoRA on classification tasks. Specifically, we selected 20 superclasses from CIFAR-100 [18] as 20 distinct tasks, with each task corresponding to a 5-class classification. The superclasses were chosen as defined in the official CIFAR-100 hierarchy. We selected 8 of these tasks as expert LoRAs and performed inference on 6 unseen tasks, with the results presented in Table. 8. We observed that, compared to Zero-Shot CLIP, both Model Soups and the selection of expert parameters to construct LoRA improved classification performance on unseen tasks. We also evaluated MOLE [1] on the unseen tasks and observed that its performance remained suboptimal. This may be attributed to two factors: (1) deterministic LoRA experts merging lacks semantic guidance and therefore struggles to generalize without access to raw image features, and (2) the domain shift between seen and unseen tasks constrains MOLE's ability to adapt beyond its training domains. Furthermore, SG-LoRA achieved the best performance, indicating that our method is also applicable to classification tasks.

Table 8. Model Performance of image classification on CIFAR-100 superclass.

Method	Accuracy
Zero-Shot CLIP	72.30
MOLE	56.56
Model Soups	75.63
AdapterSoup	72.60
Top-*K* LoRA Weighted	72.70
SG-LoRA	77.50

E. More Ablations and Qualitative Analysis

E.1. Ablation on modalities of semantic priors.

The construction of semantic priors serves as the foundation for our SG-LoRA. In Table. 9, we compare the performance of semantic conditions across different modalities, where the conditional task description from each modality directly influences the selection of experts for unseen tasks by affecting α_k in Eq.6. The visual condition is obtained by averaging the visual embeddings of training set images within each task dataset using a frozen CLIP visual encoder. Experimental results show that the textual condition better captures the semantic relationships between tasks. This could be attributed to two factors. Firstly, the high degree of condensation in textual semantics might play a role. Secondly, the disparities between training and test set images (or the presence of noisy images) within a task could result in inaccuracies in the visual prior condition.

Table 9. Ablation study on modalities of semantic prior condition

Condition	Metrics		Dataset
	I2T R@1	T2I R@1	
Visual	73.16	52.70	MS-COCO
Textual	74.31	54.42	
Visual	86.30	70.12	Flickr30K
Textual	86.90	70.66	
Condition	**Metrics**		**Dataset**
	Accuracy		
Visual	73.83		CIFAR-100
Textual	77.50		

E.2. Sensitivity analysis on initialization of task description.

In Table. 10 , we report the results of trying several initialization for task description for in-dataset image-text retrieval on MS-COCO dataset and for coss-dataset image-text retrieval on Oxfordpets dataset. The results confirm that our SG-LoRA might be marginally sensitive to the prefix initialization.

Additionally, we conducted an ablation study on textual description for the classification task, as shown in Table. 11. It can be observed that, for classification task, more de-

Table 10. Sensitivity analysis of textual description on imgae-text retrievals.

Description	I2T Metrics			T2I Metrics		
	R@1	R@5	R@10	R@1	R@5	R@10
In-dataset results on MS-COCO						
A photo of <*class name*>	74.26	88.76	92.28	54.70	75.36	82.12
A photo of the <*class name*>	74.25	88.74	92.46	54.12	75.21	82.14
A photo of a <*class name*>	74.31	88.78	92.50	54.42	75.45	82.18
Cross-dataset results on Oxfordpets						
A photo of <*class name*>	55.95	80.57	87.61	38.74	66.36	76.13
A photo of the <*class name*>	55.27	80.42	86.97	38.54	66.71	76.42
A photo of a <*class name*>	55.41	80.73	87.33	38.84	66.77	76.69

Table 11. Sensitivity analysis of textual description on CIFAR-100 superclass Classification

Description	Accuracy (%)
A photo of a <*superclass name*>	75.77
This is a classification task for recognizing <*superclass name*>, which includes *class_1, ..., class_5*	77.50

tailed textual descriptions, which account for the specific categories within each task, can better capture the semantic relationships between tasks, thus leading to improved performance.

E.3. Impact of expert repository configuration.

Consistent with Table. 3, we further evaluated the impact of incorporating the *Dog* expert from the MS-COCO dataset on retrieval performance for two unseen dog tasks in the OxfordPets dataset. As demonstrated in Table 12, including the *Dog* expert in the expert repository consistently improves the performance.

E.4. Visulization of LoRA parameters.

As shown in Figure. 8 and Figure. 9, we visualize the average LoRA parameters for both the image–text retrieval tasks and the classification tasks. Notably, semantically related tasks lie closer to each other in the parameter space. For instance, in Figure. 8, "skateboard" is situated near "snowboard", and in Figure. 9, "household furniture" is positioned close to "household electrical devices." This observation further supports our motivation for using task descriptions as a semantic bridge to measure the proximity of unseen tasks to a set of known expert tasks. Leveraging this semantic guidance, our SG-LoRA models the target task's LoRA parameter distribution and consequently generates high-performing parameters for novel tasks.

To further investigate the parameter diversity of SG-LoRA, we conducted evaluations on the unseen '*Zebra*' task from the MS-COCO dataset at different training stages and visualized the generated LoRAs using t-SNE. As shown in Figure. 10, we observe that the distribution of LoRAs generated by SG-LoRA gradually aligns with that of Oracle LoRAs (directly trained in image-caption pairs), while still preserving diversity rather than extensively overlapping with the Oracle. This indicate that, by injected stochasticity, our method effectively explores the high-performance LoRAs in the parameter space. Additionally, by examining the bottom subfigure, we observe that in parameter space, both AdapterSoup (Tok-K LoRA Merging) and Tok-K LoRA Weighted lie closer to the mean of the Oracle LoRA compared to Model Soup. This is because the latter treats all experts equally, whereas the former two provide more informative semantic guidance, allowing the LoRA parameters to be better tailored to the current unseen task.

F. Comparison with generation-based methods

In our main manuscript, we primarily focus on zero-shot open-world adaptation to unseen tasks. Since generation-based approaches mainly aim at enhancing parameters on seen tasks, we additionally compare SG-LoRA with representative generation-based models in this section. As shown in Table.13, we use P-diff [45] and COND P-Diff [3] as baselines to evaluate the advantages of our method on two seen image–text retrieval tasks. We observe that both generation-based models and SG-LoRA can produce parameters comparable to those obtained from direct LoRA fine-tuning. However, SG-LoRA leverages expert knowledge more effectively and is therefore able to generate higher-quality LoRA parameters.

Table 12. Ablation on expert repository strategy for cross-dataset evaluation. We assess the impact of the *Dog* from MS-COCO dataset on the retrieval performance in two unseen dog tasks from OxfordPets datase (marked with gray text).

Expert strategy	*Pug* **I2T**			*Pug* **T2I**			**Expert strategy**	*Chihuahua* **I2T**			*Chihuahua* **T2I**		
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10
w/o *Dog* expert	45.00	66.00	78.00	32.00	51.00	61.25	w/o *Dog* expert	64.00	89.00	93.00	52.50	81.00	88.25
w/ *Dog* expert	46.00	67.00	78.00	32.00	51.75	62.75	w/ *Dog* expert	66.00	90.00	95.00	55.25	81.25	89.75

Table 13. Comparison with generation-based methods. We asses the the retrieval performance in two seen tasks from MS-COCO dataset (marked with gray text).

Method	**I2T Metrics**			**T2I Metrics**		
	R@1	R@5	R@10	R@1	R@5	R@10
Motorcycle						
Oracle	85.56	97.82	99.46	68.94	91.55	95.50
P-diff	85.63	97.08	98.64	68.83	90.70	93.60
COND P-DIFF	86.30	97.17	98.70	69.03	91.56	94.29
SG-LoRA	87.74	97.28	98.91	71.05	91.69	95.10
Sheep						
Oracle	68.26	88.02	92.22	41.32	68.71	78.74
P-diff	66.07	85.43	89.22	39.28	65.88	75.76
COND P-DIFF	67.98	87.83	91.91	40.63	68.69	76.74
SG-LoRA	68.26	88.62	92.61	41.40	69.81	77.84

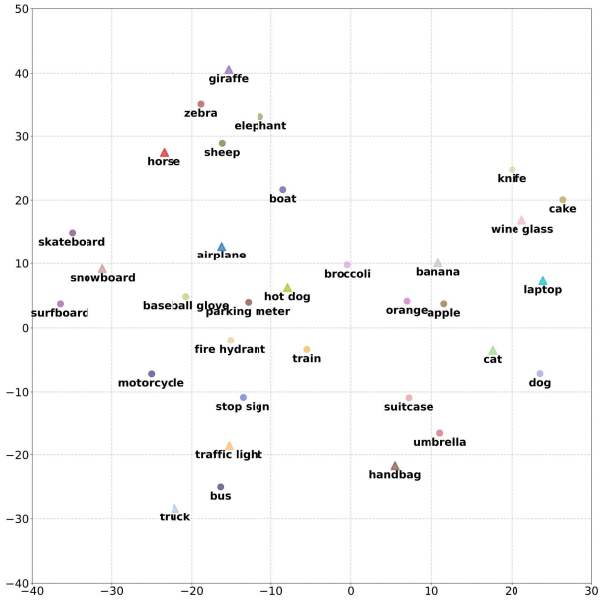

Figure 8. t-SNE visualization of the averaged LoRA parameters on MS-COCO dataset for image-text retrieval task. Triangular markers indicate expert LoRAs. Semantically similar LoRA parameters tend to cluster closely together.

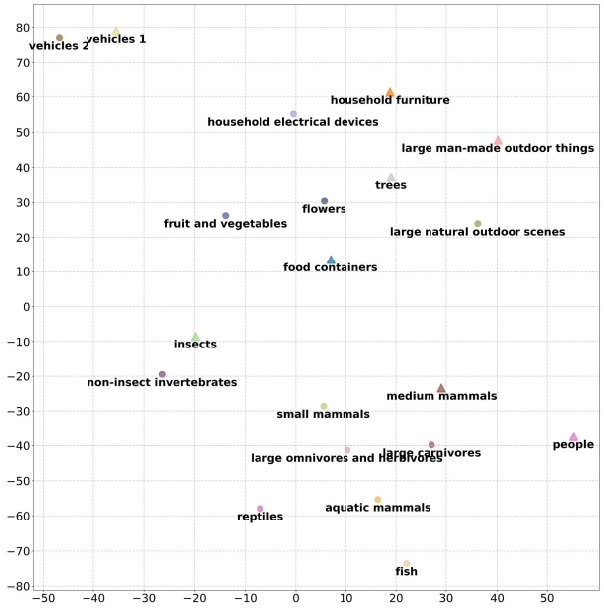

Figure 9. t-SNE visualization of the averaged LoRA parameters on CIFAR-100 for classification task. Triangular markers indicate expert LoRAs. Semantically similar LoRA parameters tend to cluster closely together.

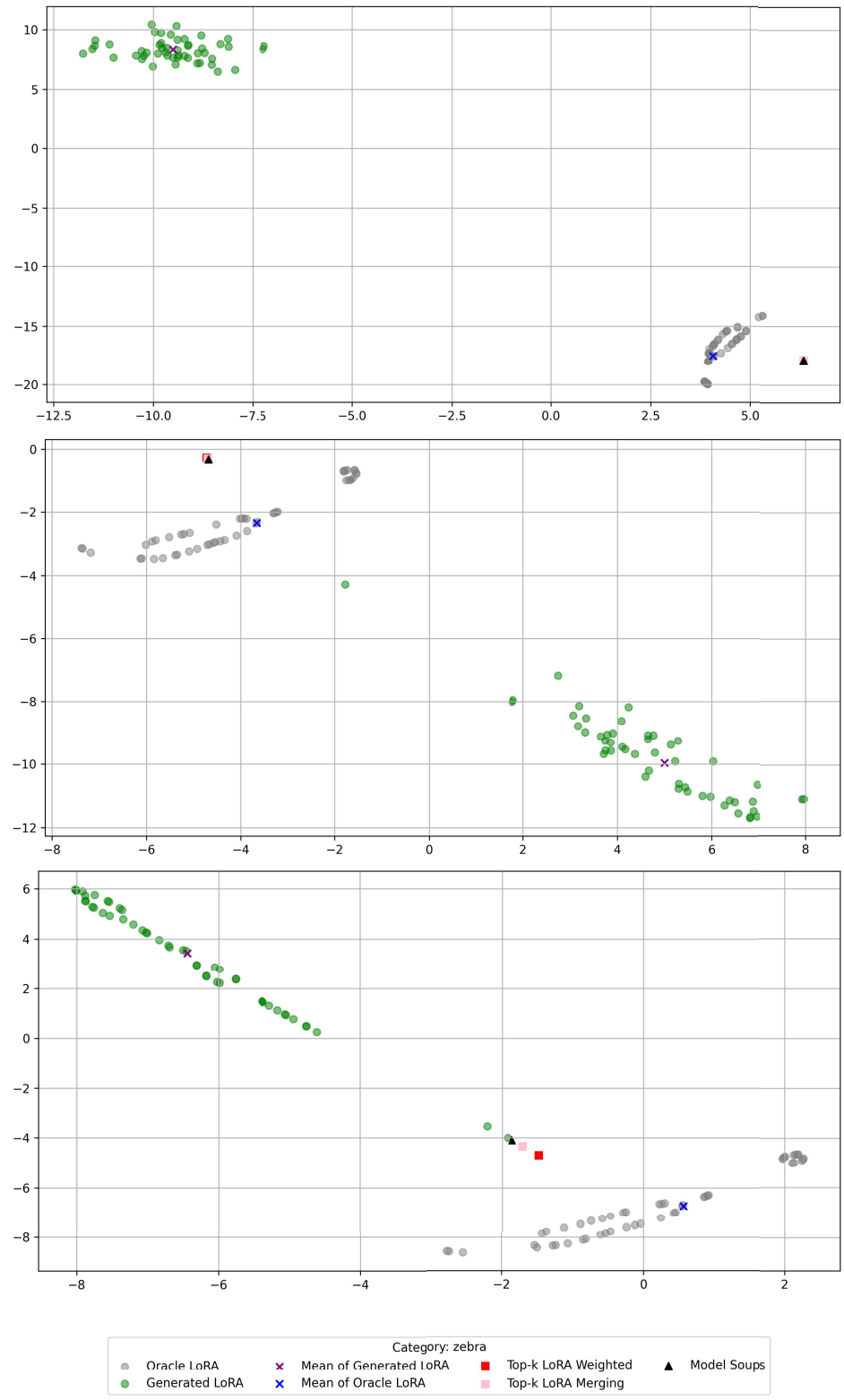

Figure 10. t-SNE visualization of LoRA parameters using different comparative methods, tested on the unseen '*Zebra*' retrieval task at different training stages. For SG-LoRA (in green) and Oracle LoRA (in gray), we randomly sampled 50 samples each. The subfigures from top to bottom represent increasing CVAE training epochs.

References

[1] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*, 2024.

[2] Kai Wang, Dongwen Tang, Boya Zeng, Yida Yin, Zhaopan Xu, Yukun Zhou, Zelin Zang, Trevor Darrell, Zhuang Liu, and Yang You. Neural network diffusion. *arXiv preprint arXiv:2402.13144*, 2024.

[3] Xiaolong Jin, Kai Wang, Dongwen Tang, Wangbo Zhao, Yukun Zhou, Junshu Tang, and Yang You. Conditional lora parameter generation. *arXiv preprint arXiv:2408.01415*, 2024.