Orthogonal Mixture-of-Expert Low-Rank Adapter for Continual Learning

Anonymous Authors[1]

Abstract

Continual Learning (CL) aims to prevent catastrophic forgetting during downstream finetuning. While Parameter-Efficient Fine-Tuning (PEFT) methods mitigate this by shielding pre-trained weights, they still suffer from severe cross-task interference. Existing solutions either use independent routers, causing structural misalignment, or rigid orthogonal constraints, severely limiting model plasticity. We propose the Orthogonal Mixture-of-Expert Low-Rank Adapter (OMoE-LoRA), which integrates an end-to-end contrastive soft router within the down-projection matrix to avoid misalignment, and an orthogonal constraint exclusively on the up-projection matrix to suppress cross-talk without sacrificing plasticity. Experiments on the MTIL benchmark demonstrate OMoE-LoRA achieves comparable accuracy with state-of-the-art method while effectively reducing trainable parameters.

1. Introduction

Continual Learning (CL) aims to enable artificial intelligence systems to sequentially acquire new knowledge without catastrophically forgetting previously learned information. This capability is indispensable for deploying models in dynamic, real-world environments such as personalized assistants, autonomous driving, and continuously evolving medical diagnosis. Within this field, Parameter-Efficient Fine-Tuning (PEFT) methods like Low-Rank Adapters (LoRA) freeze the pre-trained backbone and introduce lightweight bypass modules, effectively shielding the foundational representations.

However, when sequentially updated across multiple tasks, these adapters still overwrite previously learned concepts, leading to severe cross-task interference. To address this, MoE-Adapters (Yu et al., 2024) utilizes an Out-of-Distribution (OOD) detector as a router to direct inputs to specific expert modules. However, their detectors are trained independently from the downstream task, which increases the possibility of incorrect routing, severely degrading the final accuracy. Alternatively, orthogonal methods like InfLoRA (Liang & Li, 2024) attempt to enforce zero-interference by restricting the optimization direction of each task to a calculated null space fixed before training. While preventing cross-talk, this rigid constraint severely limits the experts' plasticity and accuracy when fitting complex new distributions.

To resolve these issues, we propose the Orthogonal Mixture-of-Expert Low-Rank Adapter (OMoE-LoRA), a framework that aligns the routing mechanism and suppresses cross-talk without sacrificing adaptability, achieving comparable accuracy with the state-of-the-art method on MTIL benchmark, with effective reduction on parameters. Our main contributions are:

- We utilize contrastive learning to train the LoRA down-projection matrix as an end-to-end, task-focused soft router, eliminating the structural misalignment caused by independent OOD detectors.
- We enforce an orthogonal subspace constraint on the up-projection matrix, providing a geometric safeguard that isolates task-specific capabilities while remaining plasticity.

2. Related Works

Parameter-Efficient Continual Learning. To balance catastrophic forgetting and efficiency, parameter-efficient fine-tuning (PEFT) methods such as Adapters (Houlsby et al., 2019) and LoRA (Hu et al., 2022) freeze pre-trained weights and add lightweight modules. LAE (Gao et al., 2023) unifies several PEFT variants into an ensemble framework, while ZSCL (Zheng et al., 2023) maintains zero-shot ability through feature-space distillation and parameter-space averaging, and establishes the MTIL benchmark used in this work.

Mixture-of-Experts Task Routing. MoE-based methods

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML 2026 Workshop "Continual Adaptation at Scale: Towards Sustainable AI". Do not distribute.

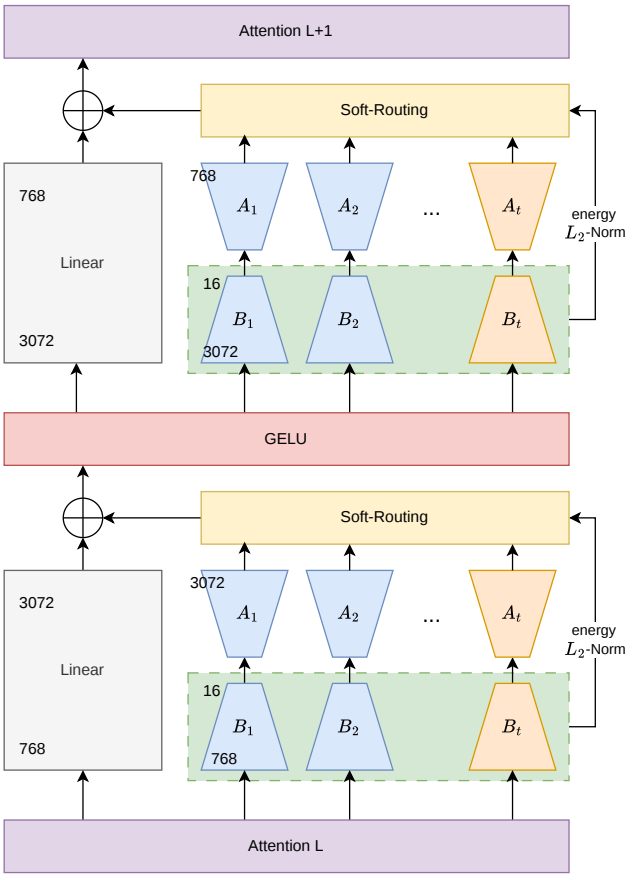


Figure 1. The architecture of our OMoE-LoRA in a certain MLP of a Transformer. During the learning of t-th task, the pretrained weight and the formerly learnt expert modules are frozen, and a new LoRA module is attached.

route inputs to task-specific adapters to reduce cross-task interference. MoE-Adapters (Yu et al., 2024), SLIM (Han et al., 2025), and C-LoRA (Zhang et al., 2025) all adopt independently trained routers to dynamically assign tasks. However, these decoupled routers create structural misalignment with the downstream objective. PASs-MoE (Hou et al., 2026) explicitly identifies this router–expert misalignment and mitigates it with pathway activation subspaces.

Orthogonal Subspace Projection. Orthogonal constraints eliminate interference by keeping task-specific updates in disjoint subspaces. InfLoRA (Liang & Li, 2024), O-LoRA (Wang et al., 2023), and OPLoRA (Xiong & Xie, 2026)enforce orthogonality through orthogonal projection on the gradient or LoRA weights. DualLoRA (Chen et al., 2024) and Rank-1 Expert Pool (Fa et al., 2026) further combine orthogonal losses with dynamic memory or sparse rank-1 composition. These methods face a dilemma between the plasticity and protectivity, either freezing the orthogonal subspace before training or only applying a soft loss function on orthogonality.

3. Methodology

3.1. Framework Overview

In this paper, we present a novel Low-Rank Adapter (LoRA) finetuning framework for vision-language models, designed to precisely train a bypass for each task in continuous learning.

During the training of the t-th task, a new LoRA module, which will be trained as the expert of the t-th task, is attached to each Linear layer of every MLP inside the Transformer. The pretrained weights and formerly attached experts are completely frozen. The overview of our framework is shown in Figure 1.

Let $x \in \mathbb{R}^d$ be the input to the module. A conventional LoRA computes the feature increment as $y = ABx$. In our continuous learning paradigm, the forward propagation of our augmented layer is formulated as:

$$y = Wx + \sum_{i=1}^{t} w_i y_i = Wx + \sum_{i=1}^{t} w_i (A_i B_i x) \quad (1)$$

where W denotes the pretrained weight, and A_i, B_i denote the up-project and down-project matrices of the i-th task expert LoRA module respectively. The routing coefficient w_i is a dynamic scalar determining the activation strength of the i-th expert, which is computed by the softmax of the L_2-norm of $B_i x$:

$$w_i = \frac{\exp(||B_i x||_2/\tau)}{\sum_{j=1}^{t} \exp(||B_j x||_2/\tau)} \quad (2)$$

3.2. Energy-Based Soft Routing

A LoRA module can be decomposed into a down-project matrix (B_t) and an up-project matrix (A_t). We explicitly formulate the activation intensity of the t-th expert as the L_2-norm of its down-projected representation: $E_t = \|B_t x\|_2$. Given that A_t is later constrained within an orthogonal subspace (as described in Section 3.3), a lower norm $\|B_t x\|_2$ mathematically bounds the expert's final contribution, ensuring irrelevant experts produce near-zero activations.

To optimize B_t as an intrinsic energy filter, we apply a contrastive-margin objective over a target batch X_t and a reference set X_{ref}:

$$\mathcal{L}_{route} = \mathbb{E}_{x \in X_t}[\max(0, m - \|B_t x\|_2)] + \mathbb{E}_{x \in X_{ref}}[\|B_t x\|_2] \quad (3)$$

We choose ImageNet as the reference dataset for the visual side of CLIP, and Conceptual Captions for the language

side, as these datasets encompass highly diverse features representing the universal distribution the model might encounter. By employing this contrastive learning strategy, the energy filter learns to actively suppress ubiquitous background noise. Furthermore, because the down-project matrix B_t simultaneously participates in the forward pass to optimize the target task, it is intrinsically coupled with the actual feature extraction process. This powerful synergy between the contrastive routing loss and the downstream task objective ensures that the filter achieves exceptional target-matching precision.

By leveraging this intrinsic energy mechanism, we establish a cohesive, end-to-end routing paradigm. We eliminate external routing classifiers, effectively bypassing the cascaded error propagation and objective misalignment typically associated with disjointed router architectures. Instead, we dynamically blend expert outputs via a Softmax function. We explicitly favor this soft routing over Top-1 hard gating to maintain structural consistency between training and inference, gracefully preserve representations of historical experts, and strictly bound the residual stream's L_2-norm ($\sum_i w_i = 1$) to prevent variance explosion in the subsequent `LayerNorm`.

3.3. Orthogonal Projection

While soft routing regulates expert magnitudes, concurrently activating multiple experts can lead to severe feature interference. To guarantee zero-interference among continuous tasks, we introduce an orthogonal projection strategy.

For a new task t, its up-project matrix A_t is first initialized as an arbitrary orthogonal matrix. During training, A_t is continuously regularized via Newton-Schulz iteration to strictly maintain its intrinsic orthogonality, thereby establishing the orthonormal basis for the expert's feature subspace.

Let $P_{\text{past}} = U_{\text{past}} U_{\text{past}}^{\top}$ denote the projection matrix onto the subspace spanned by all historical tasks. To prevent interference with historical tasks, A_t is projected into the null space of U_{past} before participating in the forward pass. The projected up-weight is computed as:

$$\tilde{A}_t = (I - P_{\text{past}}) A_t \tag{4}$$

The actual feature increment injected into the residual stream is thereby calculated as $y_t = \tilde{A}_t B_t x$. This geometric constraint guarantees that the newly injected feature increments are strictly orthogonal to all previous representations ($\langle y_{\text{past}}, y_t \rangle = 0$). Upon completing the training for task t, its optimized up-project matrix $\tilde{A}_t$ is concatenated with U_{past} to update the historical subspace (e.g., updating an initially empty U_{past} to $\tilde{A}_1$ after the first task), forming an orthogonal direct sum in the high-dimensional space.

4. Experiments

4.1. Experiment Settings

Datasets. We evaluate our method using Multi-domain Task Incremental Learning (MTIL) Benchmark following ZSCL. The model is continually trained on Aircraft, Caltech101, DTD, EuroSAT, Flowers, Food, MNIST, OxfordPet, and StandfordCars, and evaluated on all datasets every time the training on one dataset is done.

Metrics. We adopt the same metric proposed by ZSCL: Transfer, Average, and Last.

Implementation Details. We use the CLIP model with ViT-B/16 variant as our backbone. The rank of LoRA experts is set as $r = 16$. For all datasets, we train 1k iterations. For soft routing, the temperature τ is set to 1.0. The batchsize for training is set to 32.

4.2. Results

Accuracy. As summarized in Table 1, our method achieves an overall performance broadly comparable to the state-of-the-art ZSCL. The primary limitation we observed is a noticeable accuracy drop on the initial *Aircraft* dataset, which we attribute to insufficient regularization constraints when training the very first task. However, across the remaining datasets, our framework surpasses ZSCL on multiple datasets, demonstrating comparable results in the key Transfer, Average, and Last metrics.

Computational Cost. As detailed in Table 2, our method shows better efficiency among existing methods. Compared to ZSCL, the integration of LoRA drastically reduces the number of trainable parameters. Compared to MoE-Adapters, the proposed soft routing mechanism eliminates the need of standalone DDAS domain discriminator. This reduction highlights the parameter efficiency and structural advantage of our method.

Energy Visualization. We visualize the transition of the proposed energy on all datasets at each training stage at Figure 2. The distribution shows two things:

- As training progresses, the energy gap between the current task and non-target tasks continuously expands, eventually forming a sharp and well-defined decision boundary, demonstrating the effectiveness of the proposed energy routing mechanism.
- The discriminability of the energy distribution enhances from shallow to deep Transformer layers, with deeper layers establishing the task boundary at a significantly earlier epoch. This indicates that the routing mechanism aligns with the hierarchical architecture of CLIP, effectively capturing higher-level semantic information in deeper layers.

Table 1. Accuracy Comparison across multiple datasets on the MTIL Benchmark. The row of "Full FT" and "LoRA" in category "CLIP" means adopting the corresponding finetuning method on that specific dataset only, rather than continual finetuning.

	Method	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	Average
CLIP	Zero-shot	24.8	88.4	68.4	42.9	54.9	71.0	88.5	59.4	89.1	64.0	65.2
	Full FT	57.3	96.1	89.7	78.9	98.9	95.0	92.6	99.7	94.1	87.8	89.0
	LoRA	55.2	97.0	88.8	79.2	98.9	95.7	92.8	99.6	94.9	86.7	88.9
Transfer	Continual FT	-	78.2	59.6	37.2	32.9	41.8	63.0	52.6	67.6	30.7	51.5
	WiSE-FT	-	79.2	56.6	34.3	36.0	23.1	48.9	46.7	49.8	14.6	43.3
	SECA	-	77.3	66.3	42.4	51.2	63.7	85.8	45.7	83.1	60.8	64.0
	ZSCL	-	**87.2**	**68.4**	**45.6**	49.3	69.5	**88.3**	58.2	89.6	61.5	**68.6**
	MoE-Adapters	-	80.4	68.3	42.7	**51.7**	**73.0**	85.1	61.0	84.4	60.2	67.4
	Ours	-	84.6	66.5	43.1	49.4	69.5	88.0	**62.3**	**90.6**	**62.7**	68.5
Average	Continual FT	22.1	89.6	66.6	59.7	69.5	64.0	72.9	66.7	72.7	36.2	62.0
	WiSE-FT	9.1	79.1	60.8	41.3	63.4	31.7	57.8	62.6	57.6	19.5	48.3
	SECA	35.5	86.3	68.6	49.9	65.5	70.5	86.5	57.7	84.4	62.2	66.7
	ZSCL	**45.6**	93.0	81.4	**66.8**	78.3	81.4	**90.0**	70.5	90.8	64.1	76.2
	MoE-Adapters	40.7	92.4	**84.3**	54.3	**79.8**	76.5	87.8	72.6	85.3	61.4	73.5
	Ours	40.2	**94.6**	81.5	65.5	78.8	**82.2**	89.7	**73.5**	**91.4**	**64.8**	**76.2**
Last	Continual FT	20.7	89.6	65.4	70.9	93.2	85.8	86.2	99.5	93.1	86.0	79.0
	WiSE-FT	7.5	76.4	50.1	35.2	76.4	36.2	81.8	**99.6**	89.6	63.3	61.6
	SECA	34.2	85.7	69.0	53.1	74.9	73.6	86.9	82.1	86.8	74.5	72.1
	ZSCL	40.3	93.7	83.0	**74.6**	96.9	92.0	**92.4**	99.2	**95.4**	**87.6**	**85.5**
	MoE-Adapters	**40.7**	92.9	**88.3**	56.0	**98.6**	79.2	91.7	**99.6**	88.9	72.6	80.9
	Ours	34.8	**95.7**	84.1	71.1	98.2	**94.0**	92.2	99.5	94.5	84.2	84.8

Table 2. Comparison of computational costs during training. "DDAS" refers to the task discriminator module in MoE-Adapters. Δ indicates the relative reduction of our method compared to the adapter-based method MoE-Adapters. The trainable parameter count and gpu memory consumption data are from MoE-Adapters.

METHOD	TRAIN PARAMS ↓	GPU ↓
ZSCL [79]	149.6M	26290MIB
MOE-ADAPTERS	59.8M	22358MIB
DDAS	8.7M	2461MIB
OURS	22.1M	8872MIB
Δ (VS MOE-ADAPTERS)	**-63.0%**	**-60.3%**

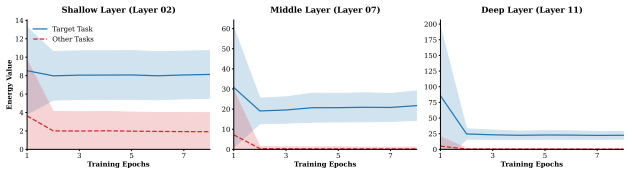


Figure 2. The transition of energy on the shallow (2nd), middle (7th), and deep (11th) layer of the Vision Transformer during the training of the dataset *StandfordCars*. Details in Appendix A.

5. Discussion

In this work, we presented an energy-based soft routing mechanism coupled with orthogonal subspace allocation to address catastrophic forgetting in continual learning. Our framework achieves parameter-efficient, interference-free adaptation by dynamically mapping distinct tasks into orthogonal sub-networks within a frozen pretrained backbone. Furthermore, our approach drastically reduces the computational overhead, requiring even fewer trainable parameters than existing adapter-based solutions.

Limitations and Future Work. Despite its strong performance across the MTIL benchmark, our framework exhibited a noticeable accuracy degradation on the initial *Aircraft* dataset. We attribute this to the fact that the first task's expert is initialized in an empty historical subspace ($U_{\text{past}} = \emptyset$), lacking the restrictive regularization that subsequent tasks benefit from. This highlights a potential area for future improvement: developing task-agnostic prior constraints or warmup strategies for the initial expert to better anchor the feature space. Future work will also explore extending this dual-stage suppression paradigm to large language models (LLMs) and investigating highly complex sequential tasks.

References

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Chen, H., Li, J., Gazagnadou, N., Zhuang, W., Chen, C., and Lyu, L. Dual low-rank adaptation for continual learning with pre-trained models. *arXiv preprint arXiv:2411.00623*, 2024.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Deng, L. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Fa, Z., Duan, Y., Zhang, J., Qi, L., Yang, W., and Shi, Y. Decomposing and composing: Towards efficient vision-language continual learning via rank-1 expert pool in a single lora. *arXiv preprint arXiv:2601.22828*, 2026.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Gao, Q., Zhao, C., Sun, Y., Xi, T., Zhang, G., Ghanem, B., and Zhang, J. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11483–11493, 2023.

Han, J., Du, L., Du, H., Zhou, X., Wu, Y., Zhang, Y., Zheng, W., and Han, D. Slim: Let llm learn more and forget less with soft lora and identity mixture. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4792–4804, 2025.

He, L., Cheng, D., Xu, D., Wang, H., and Wang, N. Harnessing textual semantic priors for knowledge transfer and refinement in clip-driven continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21645–21653, 2026.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Hou, Z., Guo, H., Ma, H., Sun, Y., Yang, Y., and Wang, J. Pass-moe: Mitigating misaligned co-drift among router and experts via pathway activation subspaces for continual learning. *arXiv preprint arXiv:2601.13020*, 2026.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Liang, Y.-S. and Li, W.-J. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647, 2024.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Wang, X., Chen, T., Ge, Q., Xia, H., Bao, R., Zheng, R., Zhang, Q., Gui, T., and Huang, X.-J. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, 2023.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7959–7971, 2022.

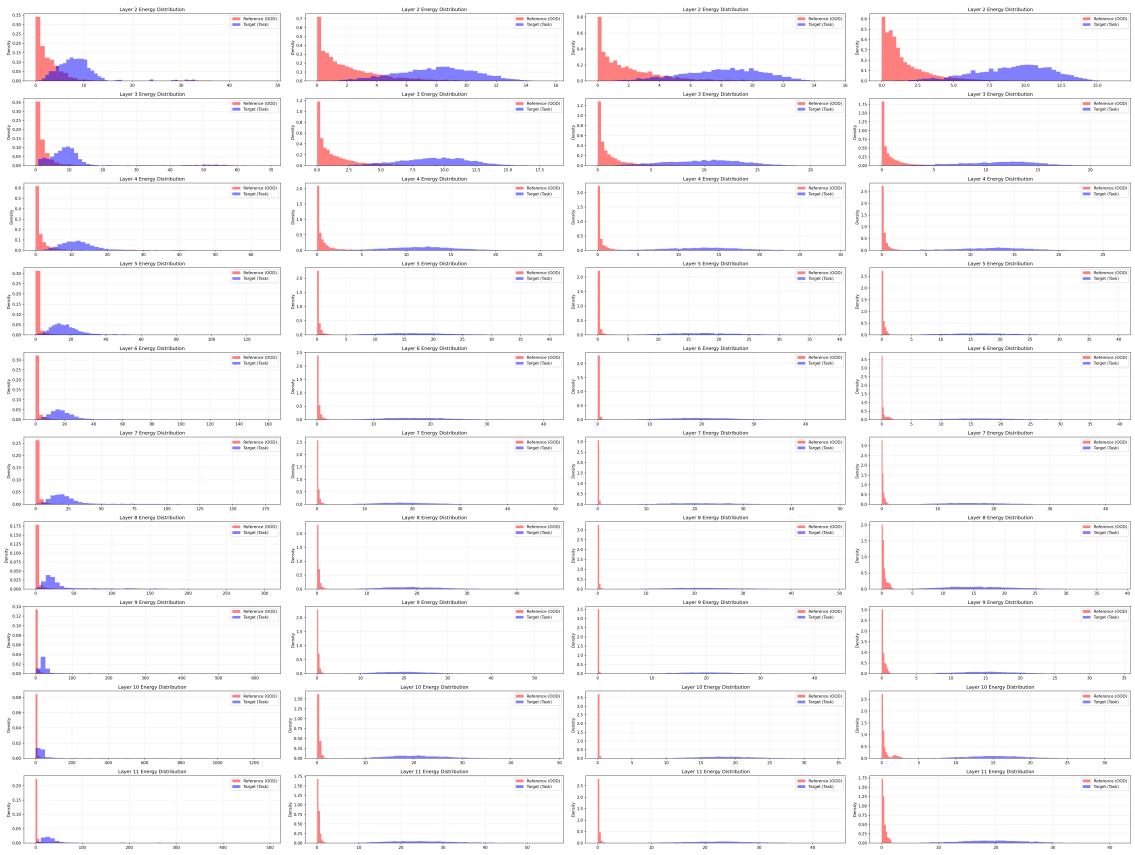
Xiong, Y. and Xie, X. Oplora: Orthogonal projection lora prevents catastrophic forgetting during parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34088–34096, 2026.

Yu, J., Zhuge, Y., Zhang, L., Hu, P., Wang, D., Lu, H., and He, Y. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23219–23230, 2024.

Zhang, X., Bai, L., Yang, X., and Liang, J. C-lora: Continual low-rank adaptation for pre-trained models. *arXiv preprint arXiv:2502.17920*, 2025.

Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., and You, Y. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19125–19136, 2023.

A. Energy Transition on Different Dataset and Different Epoch



(a) *(b)* *(c)* *(d)*

Figure 3. The energy of samples from target task domain and background domain while training on the dataset *StandfordCars*. Figure 3a, Figure 3b, Figure 3c compare the energy of the test dataset in target domain to the energy of the reference dataset, while Figure 3d compares to the test dataset in other domains.

Compare the result in Figure 3a, Figure 3b and Figure 3c with Figure 3d, we can conclude that the use of router loss $\mathcal{L}_{route}$ with reference dataset successfully distinguishes the target domain with others. It shows a sound transferring ability.