Supplementary Materials for "Robustness Under Data Scarcity: Few-Shot Continual Adversarial Training for Evolving Threats"

In this supplementary material, we conduct additional experiments to further evaluate the effectiveness of our method and discuss its practical applicability.

- We show that the performance gains are not solely due to ADM: under the same pre-trained model, our method still outperforms SSEAT and LBGAT, thanks to GMM replay and MDB mitigating catastrophic forgetting.
- We find that features sampled from the replay GMM match adversarial features better than those from a single Gaussian.
- We investigate the impact of attack order by constructing different permutations of attack sequences on the ImageNet-1K dataset.
- We visualize clean-sample representations with t-SNE and observe that our method yields more compact and better-separated clusters than SSEAT.
- We conduct more challenging few-shot training experiments on ImageNet-1K, where each training stage randomly selects 10 classes and a random number of samples per class in the range of 0–10.
- To evaluate model performance under more demanding scenarios, we construct a long attack sequence consisting of 10 different adversarial attacks encountered during training on the ImageNet-1K dataset.
- We track the clean and adversarial accuracy at each training stage and compare our method with SSEAT to analyze stage-wise robustness.
- We evaluate the defense performance of our method on both short and long attack sequences using CIFAR-10 and CIFAR-100 datasets. Our method consistently achieves the best performance across all settings.
- We further assess robustness under varying attack intensities, and our method consistently attains the highest accuracy across all PGD configurations.
- We simulate realistic data-limited settings by training with adversarial samples from only 10 ImageNet-1K classes per stage, under which our method still outperforms all baselines.
- Our method achieves the lowest computational cost. It introduces only minimal overhead while delivering strong continual few-shot robustness.

1. More experimental results.

Gains Not Attributable to ADM Alone Even when baselines are strengthened with ADM pre-training and Mixup (Tab. 1), our method still outperforms SSEAT and LBGAT in both clean accuracy and robustness, indicating that the gains cannot be attributed to ADM alone. Instead, the additional improvements come from the proposed GMM replay and MDB mechanisms, which help alleviate catastrophic forgetting across attack stages.

Table 1. Comparison with ADM-pretrained backbone and Mixup.

Method	FGSM	PGD	CW	AA	Df	Clean
PGD-AT	32.05	25.02	27.48	21.36	24.11	48.72
LBGAT	34.12	26.84	28.91	22.95	25.38	49.85
SSEAT	38.96	31.62	33.71	27.24	29.85	53.12
Ours	**42.71**	**33.53**	**36.02**	**28.91**	**32.11**	**64.51**

Table 2. Feature-space distance between adversarial and replayed.

Method	FGSM	PGD	CW	AutoAttack
Single Gaussian	1.42	1.55	1.68	1.74
GMM (Ours)	**0.87**	**0.92**	**1.01**	**1.05**

GMM Empirical Validation. Each Gaussian component uses a diagonal covariance, ensuring memory efficiency while capturing main feature variations. To validate this, we generated 10 adversarial samples per class for each attack type (including AutoAttack and CW) and compared them to samples from the fitted GMMs. As in Tab. 2, GMMs achieve smaller feature-space distances than a single Gaussian, quantitatively supporting their ability to approximate adversarial feature distributions and justifying the diagonal-covariance design.

Impact of attack sequence order on model performance. To assess the stability and adaptability of our method under varying adversarial conditions, we evaluate its performance across different attack sequence orders, as reported in Tab. 3 and Tab.2 in the main submission. Regardless of the order in which attacks are presented, our method consistently outperforms baseline approaches in both clean and adversarial accuracy. This stable advantage across permutations underscores the robustness of our framework and the effectiveness of its core components. *In particular, the Multi-Domain Balanced Loss plays a crucial role by promoting stable optimization across heterogeneous adversarial distributions, reducing the bias introduced by sequence ordering.* These findings highlight that our method is resilient not only to the diversity of attack types but also to their presentation order—an essential characteristic for practical deployment where adversarial threats may emerge unpredictably and without a fixed structure.

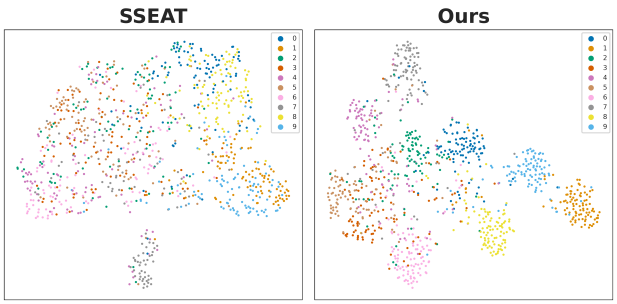

Figure 1. t-SNE visualizations of feature representations learned by SSEAT and our method on clean samples from 10 random ImageNet-1K classes on short attack sequences.

Table 3. Comparison results under a short attack sequence [CW, Df, FGSM, AA, PGD] on the ImageNet-1K.

METHODS	CW	Df	FGSM	AA	PGD	Clean
PGD-AT	26.03	22.67	29.19	19.42	23.75	46.87
AWP	27.29	24.05	30.79	21.06	25.12	45.98
LBGAT	28.94	25.66	32.63	23.05	27.02	49.61
RIFT	29.82	26.24	33.46	23.88	27.87	47.44
AFD	30.57	27.38	34.85	24.75	29.07	50.03
SSEAT	32.15	28.34	36.28	25.94	30.23	51.14
Ours	**36.45**	**32.72**	**41.08**	**30.54**	**34.82**	**62.18**

Feature Representation Visualization. We use t-SNE to visualize clean sample features from 10 random ImageNet-1K classes. As shown in Figure 1, our method yields more compact and well-separated clusters than SSEAT, indicating more discriminative and structured feature representations that contribute to improved robustness and clean accuracy.

Robust generalization under extreme few-shot settings. Unlike conventional fixed-shot experiments, we design a more practical setting by allowing the number of training samples per class to fluctuate randomly within the range of 0–10. This simulates real-world constraints where labeled data is scarce and unevenly distributed. As shown in Tab. 4, such randomness significantly increases the difficulty of learning robust representations, especially under adversarial perturbations. Despite this challenge, our method consistently achieves the best performance across all attacks. Notably, our model improves FGSM robustness by 9.11% over SSEAT and retains high clean accuracy, demonstrating its ability to generalize from unstable and minimal supervision. *These results highlight the effectiveness of our method in learning transferable and stable features, making it particularly suitable for real-world low-resource adversarial settings.*

Table 4. Defense results on ImageNet-1K against [FGSM, PGD, CW, AA, Df], where each attack is trained with 10 randomly selected classes and 0–10 randomly sampled training examples per class.

METHODS	FGSM	PGD	CW	AA	Df	Clean
PGD-AT	17.42	12.63	13.05	10.92	12.18	34.16
AWP	18.79	13.84	14.72	11.47	13.01	33.05
LBGAT	20.32	15.03	16.54	12.86	14.02	35.29
RIFT	21.67	16.12	17.28	13.45	14.71	36.84
AFD	22.83	17.34	18.41	14.23	15.68	37.65
SSEAT	24.16	18.29	19.87	15.06	16.31	38.92
Ours	**33.27**	**26.58**	**25.13**	**22.36**	**22.41**	**53.18**

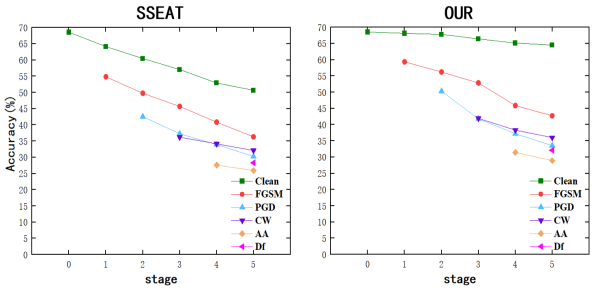

Figure 2. The accuracy of our method and SSEAT on both current and past attack data at each stage across the attack sequence [FGSM, PGD, CW, AA, Df].

Facing longer attack sequences, our proposed few-shot CAT method consistently demonstrates significant advantages. As shown in Tab. 5, our method achieves the highest adversarial robustness across all attack types, significantly outperforming existing baselines. Simultaneously, it maintains the highest accuracy on clean samples, effectively balancing the trade-off between robustness and accuracy. *In data-constrained few-shot fine-tuning scenarios, our approach greatly enhances model adaptability to diverse and evolving attacks, showcasing its strong potential for real-world deployment in dynamic adversarial environments.*

Accuracy of the model on current and past types of samples at each stage. Figure 2 shows a stage-wise accuracy comparison between our method and SSEAT under the attack sequence [FGSM, PGD, CW, AA, Df]. For early-stage attacks such as FGSM and PGD, *our method consistently achieves higher accuracy throughout all subsequent stages*, indicating stronger retention of adversarial robustness over time. These results demonstrate that our method not only enhances robustness against the current attack at each stage but also preserves defense capabilities against earlier attacks, showcasing superior generalization and long-term stability in continual adversarial training.

Table 5. Evaluation under a longer attack sequence with 10 various attacks on ImageNet-1K under a 10-shot setting.

METHODS	FGSM	BIM	PGD	SA	BS	MCG	DIM	CW	AA	Df	Clean
PGD-AT	22.43	21.96	23.15	19.88	18.64	18.35	20.02	19.47	21.80	22.12	42.57
AWP	18.17	17.28	21.66	16.25	15.82	17.44	15.63	18.91	18.03	20.70	39.83
LBGAT	21.08	22.16	24.53	20.34	19.76	21.85	19.47	23.61	22.74	25.23	43.06
RIFT	20.32	21.55	23.09	20.88	19.64	21.06	18.59	24.16	22.53	24.44	41.22
AFD	21.04	21.82	23.73	20.26	19.08	20.41	18.94	23.42	21.95	23.86	41.75
SSEAT	20.97	22.31	23.14	21.37	20.43	21.56	19.81	20.70	21.44	23.06	40.16
Ours	**27.64**	**26.97**	**26.44**	**25.83**	**25.35**	**26.98**	**26.13**	**26.72**	**24.91**	**28.33**	**50.26**

Table 6. Comparison under a short attack sequence on CIFAR-10 with 10-shot setting.

METHODS	FGSM	PGD	CW	AA	Df	Clean
PGD-AT	63.54	67.68	58.10	60.36	59.77	72.55
AWP	48.67	64.10	56.39	49.72	65.24	68.41
LBGAT	58.91	66.27	62.48	63.10	63.42	72.38
RIFT	54.12	63.81	60.20	57.76	63.23	69.55
AFD	55.41	65.19	61.35	59.89	62.07	70.02
SSEAT	60.03	63.67	65.89	64.87	65.50	70.36
Ours	**74.85**	**73.92**	**68.43**	**70.77**	**71.16**	**78.06**

Table 7. Comparison under a short attack sequence on CIFAR-100 with 10-shot setting.

Method	FGSM	PGD	CW	AA	Df	Clean
PGD-AT	43.02	41.75	41.10	37.45	36.41	54.89
AWP	38.54	40.87	36.40	38.23	36.52	50.15
LBGAT	41.12	42.96	41.68	40.28	37.12	54.77
RIFT	36.13	39.83	38.42	36.59	36.93	50.62
AFD	38.33	41.67	39.95	38.26	37.08	51.16
SSEAT	41.78	45.19	42.87	44.24	41.93	52.05
Ours	**57.62**	**55.68**	**54.33**	**54.26**	**54.92**	**61.03**

Our method consistently outperforms existing approaches on CIFAR-10 and CIFAR-100 under both short and long attack sequences. As shown in Tab. 6 and Tab. 7, under short attack sequences, our method achieves substantially higher accuracy than all competing approaches on both CIFAR-10 and CIFAR-100. Notably, on CIFAR-100, it improves PGD robustness by over 5% compared to the second-best method, with even larger gains observed on CIFAR-10. These results validate the strength of our few-shot continual adversarial training (CAT) framework in extracting robust representations from limited data. Furthermore, as illustrated in Tab. 8 and Tab. 9, our method sustains high performance under long attack sequences, while other models experience significant degradation. This demonstrates our approach's effectiveness in mitigating catastrophic forgetting and maintaining resilience throughout prolonged adversarial exposure. *Together, these results confirm that our method delivers a powerful and generalizable defense solution for few-shot continual adversarial training challenges.*

Our method keeps the best robustness under varying attack intensities. Table 10 evaluates our method under different PGD settings (20/50 iterations; $\epsilon = 8/255$ and $16/255$). Our method consistently achieves the best performance across all configurations, demonstrating strong robustness against both weak and strong attacks. This indicates the effectiveness of our approach in adapting to varying attack intensities, and highlights its ability to generalize across adversarial scenarios of different severities.

Our method can realize the best defense performance with only 10 randomly selected training classes. To better reflect realistic, data-constrained scenarios, each training session uses attack samples from just 10 randomly chosen ImageNet-1K classes instead of the full set. As shown in Tab. 11, our method consistently outperforms all competitors on both clean and adversarial inputs, demonstrating superior robustness and adaptability under severe data limitations.

Our method achieves the lowest computational cost. As shown in Tab. 12, our method achieves the lowest total computational cost among all compared defenses. By efficiently replaying past-domain data using Gaussian mixture–based feature modeling and employing a lightweight MDB loss, our design introduces only minimal overhead while maintaining strong robustness, thereby achieving a superior efficiency–performance balance.

2. Discussion

Our study highlights the critical importance of addressing adversarial robustness in data-constrained continual learning environments. Existing adversarial defense strategies often assume access to abundant adversarial data and are

Table 8. Comparison under a long attack sequence on CIFAR-10 with 10-shot setting.

METHODS	FGSM	BIM	PGD	SA	BS	MCG	DIM	Clean
PGD-AT	61.35	60.74	64.20	53.18	50.47	48.31	54.62	72.44
AWP	47.22	41.35	56.38	44.09	42.85	44.12	46.88	67.11
LBGAT	58.26	61.17	62.91	54.72	58.19	59.86	56.40	72.11
RIFT	55.94	58.33	61.02	57.48	54.39	56.73	55.81	69.42
AFD	56.85	59.08	60.57	56.32	53.77	55.93	55.42	69.05
SSEAT	61.44	63.62	62.40	61.81	63.03	61.58	62.21	70.89
Ours	**72.68**	**71.20**	**70.74**	**70.93**	**72.01**	**71.54**	**72.17**	**74.65**

Table 9. Comparison under a long attack sequence on CIFAR-100 with 10-shot setting.

METHODS	FGSM	BIM	PGD	SA	BS	MCG	DIM	Clean
PGD-AT	43.26	40.43	42.18	37.35	35.28	34.16	36.90	55.29
AWP	38.03	36.42	40.52	33.44	32.18	32.12	36.45	51.02
LBGAT	41.22	43.38	42.85	38.17	39.24	40.31	38.60	54.78
RIFT	36.97	39.81	39.05	40.49	38.93	39.86	40.72	50.87
AFD	37.78	41.03	41.24	39.72	39.55	39.14	39.92	51.34
SSEAT	41.46	45.11	45.08	44.32	42.97	42.75	43.17	52.64
Ours	**54.35**	**52.90**	**53.71**	**54.06**	**55.10**	**54.08**	**55.43**	**57.26**

Table 10. Few-shot defense performance on various attack strengths on ImageNet-1K. The W in the lower right corner indicates a weak attack perturbation size of 8/255, and the S in the lower right corner indicates a strong attack perturbation size of 16/255.

METHODS	PGD_S^{50}	PGD_S^{20}	PGD_W^{50}	PGD_W^{20}	Clean
PGD-AT	52.31	47.95	25.84	22.71	60.42
AWP	47.12	42.88	23.06	19.34	57.89
RIFT	50.84	46.33	24.57	21.86	60.11
LBGAT	49.71	45.32	25.12	21.42	59.24
AFD	48.92	44.75	22.39	19.22	58.64
SSEAT	49.53	45.88	26.10	21.76	56.71
Ours	**55.78**	**51.03**	**27.66**	**24.12**	**62.30**

Table 11. Defense results under [FGSM, PGD, CW, AA, Df] with only 10 random training classes for each attack over ImageNet-1K.

METHODS	FGSM	PGD	CW	AA	Df	Clean
PGD-AT	21.52	16.83	17.40	13.91	15.66	37.24
AWP	23.08	18.05	19.26	15.02	16.79	36.11
LBGAT	24.73	19.49	20.82	16.73	18.03	38.36
RIFT	25.84	20.31	21.76	17.21	18.92	39.77
AFD	26.97	21.18	22.62	18.07	19.63	40.86
SSEAT	28.34	22.15	24.04	18.92	20.25	42.02
Ours	**38.41**	**30.72**	**29.51**	**25.84**	**25.70**	**58.63**

prone to catastrophic forgetting when exposed to evolving threats. By introducing the FS-CAT setting and a customized framework that integrates margin-based generalization, distribution-aware replay, and multi-domain optimization, we demonstrate that it is possible to achieve both adaptability and stability even with limited adversarial samples. Through extensive evaluations on CIFAR-100, CIFAR-10, and ImageNet-1K, we show that our method significantly improves robustness under various attack sequences and permutations while maintaining high accuracy. This suggests strong potential for real-world deployment, where models must generalize to previously unseen threats without sacrificing performance on benign inputs. In principle, the FS-CAT framework is compatible with more advanced or adaptive attack generation strategies, such as learnable or policy-driven attacks, which could further create more challenging few-shot continual scenarios. However, such approaches typically involve heterogeneous parameter spaces and additional optimization complexity. In this work, we focus on strengthening the defense side—particularly the stability–plasticity trade-off

Table 12. Comparison of the total computational cost (Cost) on ImageNet-1K.

METHODS	Cost (PFLOPS)
AWP	1256.7
RIFT	1243.6
LBGAT	1334.7
AFD	1272.9
SSEAT	1232.4
Ours	**1231.2**

and the proposed GMM-based replay and MDB mechanisms—while integrating adaptive attack strategies remains a promising direction for future research. While our current framework focuses primarily on image classification, extending FS-CAT to other tasks such as object detection or multimodal learning represents a promising direction for future work. In addition, exploring self-supervised or unsupervised extensions could further reduce reliance on labeled data in few-shot scenarios.

Broader Impact The proposed FS-CAT framework contributes to the development of resilient and robust machine learning systems capable of defending against a wide range of adversarial attacks using limited training data. This is particularly valuable in domains such as autonomous driving, healthcare, and security-critical infrastructure, where robustness is essential but access to adversarial data is often scarce or prohibitively expensive.

Learnable and Adaptive Attack Strategies. We appreciate the reviewers' excellent suggestions. FS-CAT is compatible with such approaches: in principle, learnable or policy-driven attacks could replace the fixed attacks per stage to create more challenging few-shot continual scenarios. However, each attack type has a heterogeneous parameter space and optimization dynamics, requiring separate strategy networks and substantially increasing complexity. Our focus is on the defense side, *i.e.*, stability–plasticity trade-offs and GMM/MDB components, while integrating adaptive attacks is a promising direction for future research. We will cite and discuss the relevant works in the revision.