Unveiling CLIP Dynamics: Linear Mode Connectivity and Generalization

Alireza Abdollahpourrostam [1] **Amartya Sanyal** [2] **Seyed-Mohsen Moosavi-Dezfooli** [3]

Abstract

Zero-shot models like CLIP are often fine-tuned on a target dataset to improve its accuracy further, but this can compromise out-of-distribution (OOD) robustness. Robust Fine-Tuning (`RFT`) (Wortsman et al., 2022c), which interpolates between the *zero-shot* and fine-tuned models, has been proposed to address this issue. However, understanding when `RFT` actually improves OOD error remains limited. In this work, we empirically investigate the robustness of `RFT` in CLIP models, focusing on two key factors: 1) the *presence* or *absence* of barriers in the interpolation path between the zero-shot and fine-tuned models, and 2) fine-tuning choices such as data augmentation and learning rate magnitude. Our extensive experiments reveal that the *absence* of barriers correlates with larger gains in OOD accuracy for `RFT`. Additionally, we show that fine-tuning without data augmentation and using smaller learning rates consistently results in lower OOD errors. While similar findings have been reported for CNN models, this is the first work, to the best of our knowledge, to study these properties for CLIP models [1].

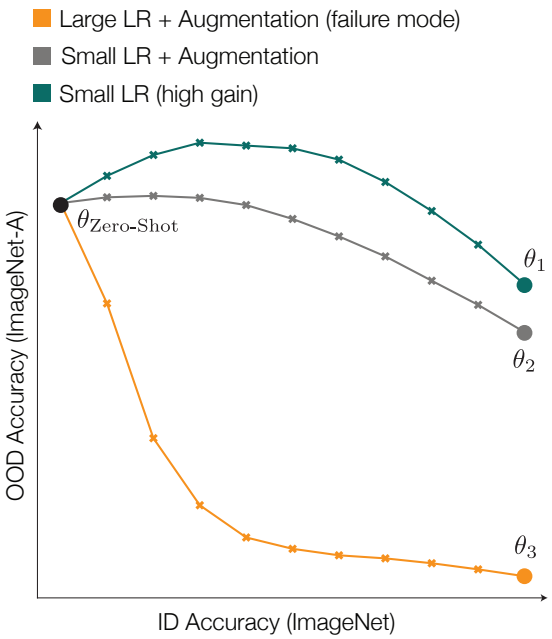

Figure 1: This research highlights the impact of learning rate magnitude and the use of augmentation on the fine-tuning of CLIP. These factors play a crucial role in determining the ***success*** or ***failure*** of interpolation within the weight space.

1. Introduction

Understanding the behavior of large machine learning models like CLIP (Radford et al., 2021) on OOD tasks is important for their safe deployment. Analyzing their behavior on a linear path between the initial and the final parameters has been proposed as a simple yet insightful approach this. However, prior works (Vlaar & Frankle, 2022; Lucas et al., 2021; Neyshabur et al., 2020; Draxler et al., 2018; Entezari et al., 2022; Chatterji et al., 2020) has primarily focused on CNN models for this analysis and whether such analysis extends to other kinds of architecture has not been thoroughly explored. On the other hand, several works have shown that while foundation models like CLIP exhibit outstanding zero-shot OOD performance, this can be further improved if they are fine-tuned on the relevant target domain. However, this improvement comes at the cost of reduced performance on domains that it is not trained on. To solve this problem, inspired by the above-mentioned works on linear interpolation in CNNs, Wortsman et al. (2022b) showed that on the linear path connecting the zero-shot model and the final fine-tuned model, there exists a model with better OOD performance and proposed an algorithm called *Robust Fine Tuning* (`RFT`) to find this parameter. However, `RFT` does not always succeed in achieving large improvement in OOD accuracy compared to the zero-shot model, and very little understanding exists of when the improvement is large and

[1]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. Work done during an internship at LTS4. [2]Max Planck Institute for Intelligent Systems, Tuebingen. [3]Independent Researcher. Correspondence to: Alireza Abdollahpourrostam <alirezaabdollahpour1380@gmail.com>.

Published at ICML 2024 Workshop on Foundation Models in the Wild.

[1]Code for the experiments is published `https://github.com/alirezaabdollahpour/CLIP_Mode_Connectivity`

when it isn't. In this work, we aim to address this lack of knowledge. Inspired by earlier work on the linear interpolation between two CNN models, we first procvide extensive experimental results to examine the correlation between the linear path's geometry and CLIP's capability to generalize on OOD tasks. We aim to address the following question:

> *How does loss monotonicity on OOD samples relate to CLIP generalization?*

Second, we investigate the role of the complexity fine-tuning algorithm on CLIP's OOD generalization. In particular, we ask the following question

> *How does data augmentation and choice of learning rate affect OOD generalization for CLIP?*

Robust Fine-Tuning (`RFT`) method has two steps: first, they fine-tune the *zero-shot* model on the target distribution. Second, they combine the original *zero-shot* and fine-tuned models by linearly interpolating between their weights, coined as weight-space ensembling. Nevertheless, the connection between linear interpolation and OOD generalization for CLIP has not been thoroughly investigated. The question of why the linear interpolation between *zero-shot* and fine-tuned CLIP models succeeds in OOD tasks, and the conditions under which the linear path between two CLIP models indicates robust generalization performance on OOD tasks, remains an unresolved problem. The latest advancements in the loss landscape of CNNs and the connection between linear paths in CNNs and generalization have motivated us to reconsider these discoveries within the context of a foundation model such as CLIP. Our objective is to bridge the gap between assumptions made about linear interpolation and loss landscape geometry in the context of CNN models and the generalization capabilities of CLIP. We intend to determine the conditions under which linear interpolation *may* be ***successfully*** done between two CLIP models.

2. Exploring the Monotonicity of Loss Landscapes and the Existence of Barriers

We begin the definition of loss barrier in the context of OOD loss landscape geometry. Then, we explore the relationship between barriers and the generalization of CLIP for OOD tasks.

Loss barrier. For loss landscapes, *barriers* refer to regions of increased loss encountered along the interpolation path between two sets of model parameters.

We examine a CLIP architecture that is parametrized by $\boldsymbol{\theta}$ and is fine-tuned on a task represented by a training set S_{train} and a test set S_{test}. In the following, as we are interested in the generalization of CLIP on OOD tasks, we consider OOD loss and accuracy and write $\mathcal{L}(\boldsymbol{\theta}), \mathcal{A}(\boldsymbol{\theta})$ for $\mathcal{L}(\boldsymbol{\theta}, S_{\text{OOD}}), \mathcal{A}(\boldsymbol{\theta}, S_{\text{OOD}})$. Assume that we have fixed two different different sets of weights $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. Let $\mathcal{L}_\alpha(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \mathcal{L}(\alpha\boldsymbol{\theta}_0 + (1-\alpha)\boldsymbol{\theta}_1)$ and $\mathcal{A}_\alpha(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \mathcal{A}(\alpha\boldsymbol{\theta}_0 + (1-\alpha)\boldsymbol{\theta}_1)$ for $\alpha \in [0, 1]$ be the loss and accuracy, respectively, of the CLIP network created by linearly interpolating between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. Then, building upon the (Frankle et al., 2020) definition for linear interpolation instability, we define it for CLIP on OOD as the following notion.

Definition 1. *The difference between the supremum of the loss for any interpolation* $\sup_\alpha \mathcal{L}_\alpha(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ *and the average loss of the endpoints* $\frac{1}{2}(\mathcal{L}(\boldsymbol{\theta}_0) + \mathcal{L}(\boldsymbol{\theta}_1))$ *is called the linear interpolation instability for the CLIP on OOD.*

Recall that *zero-shot* CLIP performs better on OOD tasks compared to the fine-tuned version of CLIP. Within the same settings of (Wortsman et al., 2022b;a), we are interested in exploring the linear path between *zero-shot* CLIP and fine-tuned CLIP. Therefore, we set $\boldsymbol{\theta}_0$ as *zero-shot* model.

Two parametrizations $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ have a **barrier** between them if the linear interpolation instability for ***sufficiently*** large δ, there exists an $\alpha \notin \{0, 1\}$ such that:

$$\sup_\alpha \mathcal{L}_\alpha(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; S_{\text{OOD}}) - \mathcal{L}(\boldsymbol{\theta}_0; S_{\text{OOD}}) \geq \delta > 0 \quad (1)$$

The value of δ can be empirically determined for each OOD task. Similarly, we state that linear interpolation or the `RFT` algorithm can achieve ***high gain accuracy*** if there exists an $\alpha \in [0, 1]$ such that:

$$\sup_\alpha \mathcal{A}_\alpha(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; S_{\text{OOD}}) - \mathcal{A}(\boldsymbol{\theta}_0; S_{\text{OOD}}) \geq \xi > 0 \quad (2)$$

where ξ is ***sufficiently*** large. Also, we define a linear path as having a *gain* if the *supremum* in Eq. 2 exists with $\xi > 0$. It is important to mention that a path is considered a ***failure mode*** if the *supremum* in Eq. 2 does not exist.

> **Hypothesis**: Interpolating linearly on a path that includes barriers can have adverse effects on generalization on OOD of models resulting from the interpolation.

Prior research has shown that the existence of a barrier between two CNN models can lead to failed interpolation and that when a barrier is present in a linear path, the interpolation process results in a ***failure mode*** with a gain of 0.00% (Vlaar & Frankle, 2022; Lucas et al., 2021; Entezari et al., 2022).

In Figure 2, we show despite all models exhibiting barriers in their loss landscapes, linear interpolation or `RFT` algorithm can *still* yield a model with ***slightly*** better performance on OOD tasks. For example, the *blue* model, despite experiencing an increasing loss, deviates slightly from the *zero-shot* point and nonetheless achieves a new performance level with a 1.39% gain. This phenomenon is similarly observed in the *red* model. Furthermore, an interesting phenomenon is noted in the ***black*** model, which attains ***high gain accuracy*** (2.11% gain) despite being derived from a linear interpolation that includes a barrier. Additionally, Figure 3 (right)

displays the quantified improvement achieved by each of the 70 CLIP models on ImageNet-A as an OOD task. Finally, in the context of CLIP for OOD settings, no *strong* correlation is observed between the **depth** of a path and generalization.

Finding: Applying linear interpolation along a path that includes barriers does *not* ***necessarily*** result in a ***failure mode***.

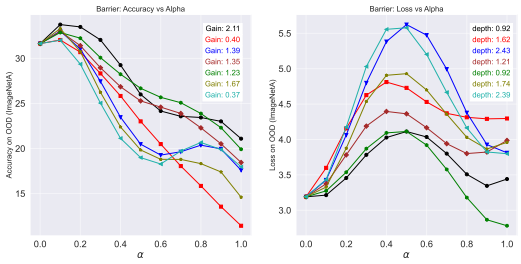

Figure 2: For 7 distinct fine-tuned CLIP models (each colors shows different CLIP models) on ImageNet (Deng et al., 2009), this plot demonstrates the accuracy and loss on ImageNet-A (Hendrycks et al., 2021) as an OOD task. For each model, we show the maximum accuracy gain achieved along a corresponding linear path. In the loss plot, we show **depth** as the largest barrier on the linear path starting from *zero-shot* model.

Monotonicity of loss landscapes. As we delve into the concept of interpolation within a path, including a barrier, it becomes imperative to examine interpolation along a trajectory that demonstrates monotonic behavior. Formally, we define a loss landscape as monotone if ***linear interpolation instability*** of loss being equals to $\frac{1}{2}|\mathcal{L}(\boldsymbol{\theta}_0, S_{\text{OOD}}) - \mathcal{L}(\boldsymbol{\theta}_1, S_{\text{OOD}})|$.

In this paper, we build upon the work of (Lucas et al., 2021; Vlaar & Frankle, 2022), which explored the presence or absence of the ***monotonic decrease*** (MLD) property for CNNs (ResNet, VGG) in test loss landscape for ID. They establish a link between ***successful*** linear interpolation and the violation of the MLD property in the loss landscape geometry. However, they argue that there is ***no correlation*** between the presence or absence of the MLD property and the ***generalization*** of CNNs. In this study, we investigate the MLD property within the context of VLM, such as CLIP, and specifically examine the loss landscape geometry of OOD samples. Contrary to the findings for CNNs, we demonstrate a ***significant correlation*** between the violation of ***monotonicity*** and ***generalization*** of CLIP.

Monotonicity for OOD task. When applying ***monotonic decreasing*** concept to OOD generalization with CLIP, an ***inverse*** representation of this measure is *also* required. This adjustment is needed because, as already mentioned, the *zero-shot* CLIP performs better on OOD tasks compared to the fine-tuned version of CLIP on specific tasks. Consequently, for accurate analysis, it is essential to consider a ***monotonic increase*** in loss *too*.

Case study 1: *monotonic decreasing* (MLD). In this case, we investigate whether a loss landscape $(\mathcal{L}_\alpha(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; S_{\text{OOD}}))$ with a MLD property can capture generalization on OOD tasks. As shown in Figure 3, we show that when CLIP models exhibit a MLD property, linear interpolation or the `RFT` algorithm can achieve ***high gain accuracy*** along these kinds of monotone path. Indeed, our empirical observations indicate that whenever a linear path possesses the MLD property, performing linear interpolation or using the `RFT` algorithm in the context of CLIP models on that path is reasonable. Our findings complement the studies conducted by (Vlaar & Frankle, 2022; Lucas et al., 2021; Goodfellow et al., 2015) regarding the ***success*** of linear interpolation and the existence of MLD in CNNs throughout the context of $\mathcal{L}_\alpha(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; S_{\text{OOD}})$ for CLIP.

Case study 2: *monotonic increasing* (MLI). In this case, we investigate whether a loss landscape $(\mathcal{L}_\alpha(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; S_{\text{OOD}}))$ with a MLI property can capture generalization on OOD tasks. As shown in Figure 3, when a linear path exhibits a MLI property, linear interpolation or the `RFT` algorithm can produce a new model that surpasses all models along that path. This result demonstrates that ***monotonic decreasing*** or MLD is not a ***necessary*** condition for achieving ***high gain accuracy*** with CLIP on OOD tasks.

Another crucial point derived from these two case studies is that, unlike CNNs, achieving ***high gain accuracy*** along a linear path *just* requires ***monotonicity***. Our results complement (Lucas et al., 2021) findings on the effect of MLD on ***successful*** interpolation, extending these insights to the generalizability of CLIP on OOD tasks. Specifically, *although* we confirm that the presence of the MLD property in loss landscape can be a reliable indicator for estimating CLIP's performance on OOD tasks *but* we find ***strong*** correlation between *pure* ***monotonicity*** and generalization. In Figure 3 (box-plot), we show that CLIP models with ***monotone*** loss landscape $(\mathcal{L}_\alpha(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1; S_{\text{OOD}}))$ achieve significantly ***high gain accuracy*** compared to models exhibiting barriers in their loss landscapes.

Role of various fine-tuning strategies. We aim to explore two critical components of CLIP fine-tuning. Firstly, we examine the impact of small and large learning rates on the shape of the linear path and the ***success*** of linear interpolation. Secondly, we will investigate the role of data augmentation during the fine-tuning of CLIP.

Hypothesis: The magnitude of the learning rate directly affects the geometry of the linear path.

In order to evaluate this hypothesis, we first trained a model using a learning rate that was set to a low value. Afterward, while keeping all other factors the same, we systematically modified the learning rate, progressively raising it to two, four, and ultimately eight times the initial value. Subsequently, we trained new models for each of these changes.

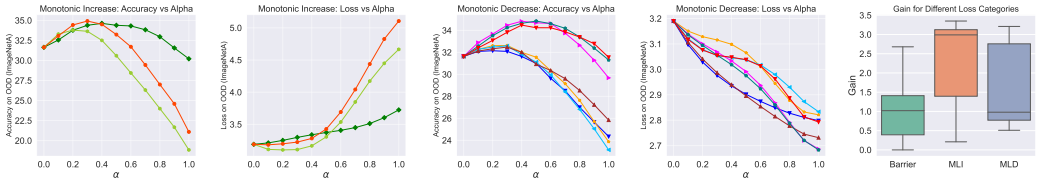

Figure 3: We show the effects of a monotonic increasing and monotonic decreasing correlation with the generalization of CLIP on ImageNet-A as an OOD task (See Appendix A for more models). In the box plot, we measure the ***high gain accuracy*** on OOD for 70 different fine-tuned CLIP models on ImageNet. This measurement is taken across the linear interpolation between each fine-tuned model and the *zero-shot* CLIP model.

In Table 1, we show that fine-tuning with a large learning rate can exhibit barriers in a linear path. Moreover, it is clearly obvious that in a path with barriers, RFT does not achieve ***high gain accuracy***. This result shows that RFT is *not* a ***generalizable*** algorithm for different settings, and its performance is related to the properties of the fine-tuned model. Indeed, smaller learning rates lead to smaller changes in the features of the *zero-shot* CLIP model which are more robust to distribution shifts since they were obtained from a much larger dataset than ImageNet (Andriushchenko et al., 2023a).

Finding: Fine-tuning CLIP with a high learning rate can increase ***linear interpolation instability*** and exhibit barriers simultaneously.

3. On the Role of Data Augmentation

In this section, we further investigate the effect of data augmentation on linear interpolation. In fine-tuning phase, we use minimum crop size in the data augmentation and optionally apply RandAugment (Cubuk et al., 2019), mixup (Zhang et al., 2018), or CutMix (Yun et al., 2019).

Hypothesis: Fine-tuning $\boldsymbol{\theta}_1$ *without* augmentation will lead to a ***significant increase*** in accuracy when using linear interpolation or the RFT algorithm.

In order to evaluate this hypothesis, we present a comparison where all parameters of a model are fixed, and two scenarios are considered: one model is fine-tuned *with* data augmentation, and the other *without* it. In Figure 5, it is evident that models without augmentation considerably achieve substantial ***high gain accuracy***. Controversy, linear interpolation, or the RFT algorithm fails to achieve ***high gain accuracy*** on the linear path between *zero-shot* and augmented fine-tune CLIP.

In Figure 4, we measure the ***gain accuracy*** on OOD for 70 different CLIP models on ImageNet-A (OOD). This demonstrates that, among the 70 models, those fine-tuned *without* augmentation achieve substantially ***high gain accuracy***.

Our findings confirm the results previously obtained by (Vlaar & Frankle, 2022), which indicate that reducing

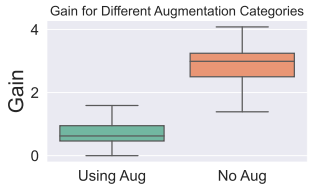

Figure 4: ***Gain accuracy*** on ImageNet-A for 70 different CLIP models.

the complexity of the task, such as by decreasing the amount of data or removing augmentation, can *enhance* the test accuracy of CNNs. In another context, (Andriushchenko et al., 2023b) observed that for ResNets, there is a strong correlation between sharpness (Foret et al., 2020) and OOD performance, but only within each subgroup of training parameters, such as augmentations and mixups. However, they do not observe a clear correlation between this phenomenon and ViTs (Dosovitskiy et al., 2021).

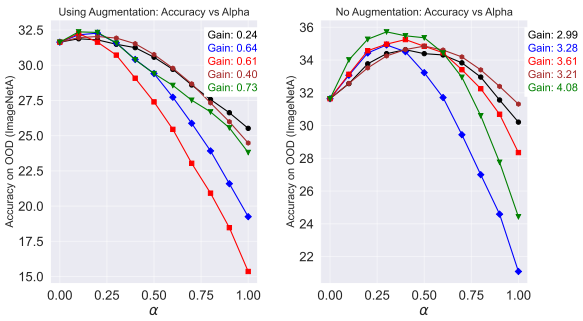

Figure 5: We present a comparison where all parameters of a model are fixed, and two scenarios are considered: one model is fine-tuned *with* data augmentation and the other *without* it (each color shows different CLIP models).

Failure mode. We demonstrate that the *combined* use of a high learning rate and data augmentation can ***adversely*** affect the ***success*** of linear interpolation. This combination can lead to a ***failure mode*** in the RFT algorithm or linear interpolation. In Figure 6, we demonstrate instances of CLIP models where the combination of high learning rates and

augmentation has an *adverse* effect on linear interpolation. In this particular situation, linear interpolation reveals to be *completely* ***worthless*** in producing superior models.

Figure 6: We show the effect of the superposition of both learning rate magnitude and the presence or absence of augmentation during the fine-tuning of CLIP. Sometimes, it can result in a ***failure mode***.

Finding: Linear interpolation or the `RFT` algorithm achieves significant ***high gain accuracy*** when θ_1 is fine-tuned *without* ***augmentation***.

4. Conclusion

In conclusion, our study demonstrates that linear interpolation plays a critical role in enhancing the generalization capabilities of CLIP models for OOD tasks. We establish that fine-tuning CLIP without dataset augmentation and using appropriate learning rates significantly improves OOD accuracy. Notably, a non-monotonic loss landscape, often caused by higher learning rates, diminishes CLIP's generalization effectiveness. Overall, our findings bridge theoretical assumptions about linear interpolation in CNN models to practical applications in CLIP. Notably, this work is the first to investigate the generalization and interpretability of CLIP (VLM) models using mode connectivity and interpolation, providing new insights into their behavior and potential for robust application across diverse tasks.

5. Acknowledgements

Alireza Abdollahpourrostam carried out part of the research presented in this paper during an internship at LTS4, with partial support from Pascal Frossard. The authors wish to thank Guillermo Ortiz-Jimenez for his valuable discussions during the preparation of this paper.

References

Andriushchenko, M., Croce, F., Müller, M., Hein, M., and Flammarion, N. A modern look at the relationship between sharpness and generalization, 2023a. 4

Andriushchenko, M., Croce, F., Müller, M., Hein, M., and Flammarion, N. A modern look at the relationship between sharpness and generalization, 2023b. 4

Chatterji, N. S., Neyshabur, B., and Sedghi, H. The intriguing role of module criticality in the generalization of deep networks. *ICLR*, 2020. 1

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space, 2019. 4

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. https://ieeexplore.ieee.org/abstract/document/5206848. 3

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. https://openreview.net/forum?id=YicbFdNTTy. 4

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning (ICML)*, 2018. https://arxiv.org/abs/1803.00885. 1

Entezari, R., Sedghi, H., Saukh, O., and Neyshabur, B. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations (ICLR)*, 2022. https://arxiv.org/abs/2110.06296. 1, 2

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 4

Frankle, J., Dziugaite, G. K., Roy, D., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning (ICML)*, 2020. https://proceedings.mlr.press/v119/frankle20a.html. 2

Goodfellow, I., Vinyals, O., and Saxe, A. Qualitatively characterizing neural network optimization problems. *ICLR*, 2015. 3

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples, 2021. 3

Lucas, J., Bae, J., Zhang, M. R., Fort, S., Zemel, R., and Grosse, R. Analyzing monotonic linear interpolation in neural network loss landscapes, 2021. https://arxiv.org/abs/2104.11044. 1, 2, 3

Neyshabur, B., Sedghi, H., and Zhang, C. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. https://arxiv.org/abs/2008.11687. 1

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. 1

Vlaar, T. J. and Frankle, J. What can linear interpolation of neural network loss landscapes tell us? In *International Conference on Machine Learning*, pp. 22325–22341. PMLR, 2022. 1, 2, 3, 4

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022a. https://arxiv.org/abs/2203.05482. 2

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Gontijo-Lopes, R., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. Robust fine-tuning of zero-shot models, 2022b. 1, 2

Wortsman, M., Ilharco, G., Li, M., Kim, J. W., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. Robust fine-tuning of zero-shot models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022c. https://arxiv.org/abs/2109.01903. 1

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019. 4

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization, 2018. 4

A. Monotonocity of loss landscape geometry

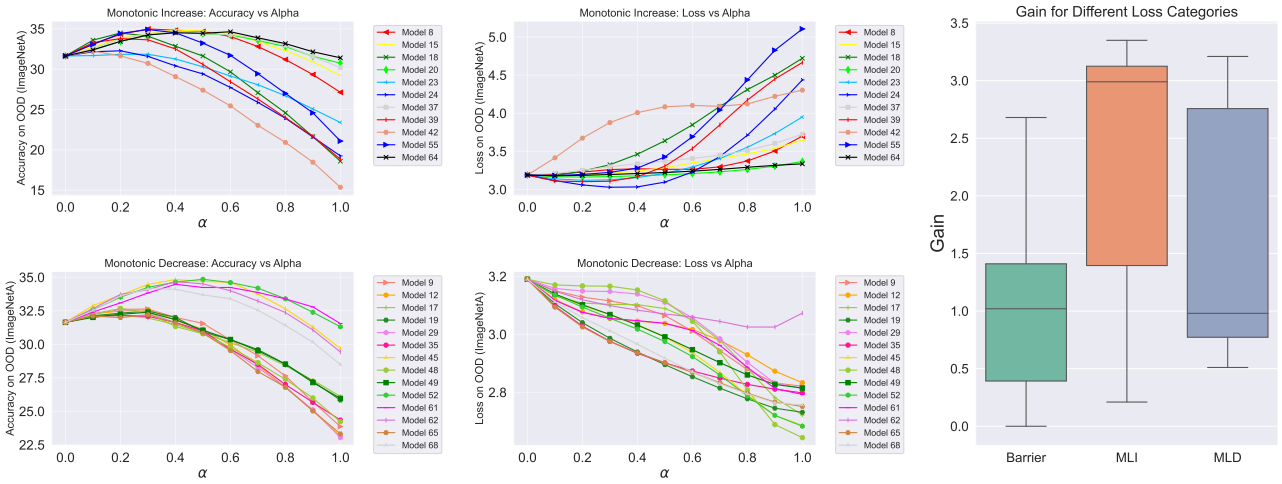

Figure 7: In this plot, we show the effects of a monotonic increasing (***top***) and monotonic decreasing (***bottom***) correlation with the generalization of CLIP on ImageNet-A as an OOD task for 70 CLIP models. In the right plot, we measure the ***high gain accuracy*** on OOD for these models. This measurement is taken across the linear interpolation between each fine-tuned model and the *zero-shot* CLIP model.

B. On the Role of Learning Rate

Table 1: Fine-tuning CLIP on ImageNet with different learning rates. We show test accuracy on ImageNet and its variant ImageNet-A as OOD.

LR	ImageNet (%)	OOD (%)	RFT	Gain (↑)%
LR = 1.00×10^{-5}	0.77	0.27	✔	3.7
LR = 2.00×10^{-5}	0.77	0.21	✗	$0.01 \approx 0.00$
LR = 4.00×10^{-5}	0.76	0.17	✗	$0.03 \approx 0.00$
LR = 7.00×10^{-5}	0.76	0.11	✗	$0.04 \approx 0.00$
LR = 7.25×10^{-5}	0.78	0.14	✗	0.00
LR = 8.00×10^{-5}	0.77	0.13	✗	0.00
LR = 7.92×10^{-6}	0.77	0.29	✔	3.21
LR = 3.61×10^{-6}	0.77	0.25	✔	3.04

C. On the Role of Data Augmentation

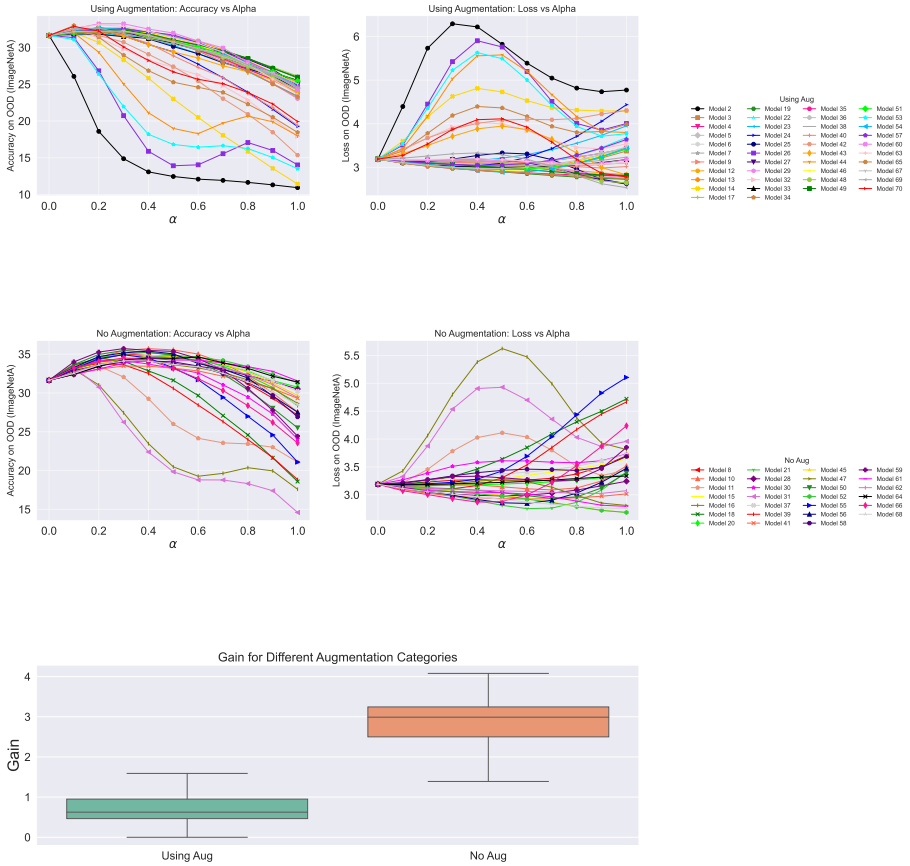

Figure 8: In this plot, we show the effects of fine-tuning *with* augmentation (***first row***) and *without* augmentation (***second row***) correlation with the generalization of CLIP on ImageNet-A as an OOD task for 70 CLIP models. In the ***bottom*** box plot, we measure the ***high gain accuracy*** on OOD for these models. This measurement is taken across the linear interpolation between each fine-tuned model and the *zero-shot* CLIP model.

In this section, we provide further details and additional results supporting our findings on the impact of data augmentation during fine-tuning on the performance of CLIP models in out-of-distribution (OOD) tasks. Our primary observation is that the presence of data augmentation during the fine-tuning phase negatively affects the efficacy of linear interpolation and the Robust Fine-Tuning (RFT) algorithm. Specifically, we found that models fine-tuned with data augmentation do not exhibit significant accuracy improvements along the linear path, whereas models fine-tuned without data augmentation do.

To substantiate these findings, we conducted multiple experiments with varying sets of parameters and different model configurations. Figures in 8 illustrate that fine-tuning with data augmentation consistently results in worse-performing models along the linear interpolation path. These results are consistent across different training runs and parameter settings, highlighting the robustness of our observations.

These supplementary results provide a comprehensive view of the nuanced dynamics between data augmentation and fine-tuning strategies, reinforcing our main findings that highlight the superior performance of CLIP models fine-tuned without data augmentation on OOD tasks.