We would like to thank all the reviewers for recognizing the contributions of our work and providing valuable feedback. Below are our responses to the comments.

1 To Reviewer #1

To the comment "…some more direct evidence to support the key arguments…": we follow the kind suggestion from the reviewer and compare the cosine similarity between prior and true gradients in absolute values with that evaluated using Gaussian vectors. We tested on 100 CIFAR-10 images with a WRN as the victim model, and the average results are: 0.152 for ResNet-56, 0.154 for ResNet-32, 0.130 for VGG-19, 0.141 for VGG-16, and 0.014 for random Gaussian vectors. Apparently, prior gradients calculated on these reference models are an order of magnitude more similar to the true gradient of WRN, than random Gaussian vectors.

To the comment "…Figure 1 and subspace attack…": we would like to explain that our idea stems from performing zeroth-order optimizations in linear subspaces with reduced dimensionalities, as introduced in the first half of Section 3 in our paper. Based on prior evidence, we hypothesize that different DNN models may share similar input-gradient (subspaces) and it can be more principled to substitute the random basis vectors with prior gradients (line 116-119). At first, these basis vectors are set to be *fixed* over the entire optimization procedure, that said, only the input-gradient of reference models with respect to the clean image $\mathbf{x}$ is utilized. Apparently, then a natural extension/improvement is to *adaptively* calculate the prior gradients (of the reference models) with respect to the current estimation of adversarial example $\mathbf{x}_{\text{adv}}$, i.e., our solution introduced in the second half of Section 3. See Figure 1 for an illustrative comparison of different choices of basis vectors. Hope this also addresses the minor concern regarding "adaptive". We will revise the paper accordingly and add more explanations to enhance the clarity and readability.

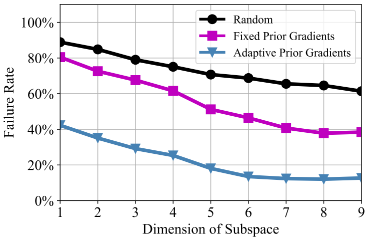

Figure 1: Choices of basis vectors.

To the comment "…only if we know the data distribution first": we now apply our method under more severe domain shifts to address your concern. We attempt to train reference models in significantly different domains from that of the target. Specifically, we use reference models trained on 1) ImageNet and 2) noisy CIFAR-10.1 images (with additive Gaussian noise and $\sigma = 10/255$) respectively to attack a victim WRN on CIFAR-10, and still obtain $\sim$33% and $\sim$30% reductions in query on the base of Bandits-TD, both with lower failure rates.

To the comment "…why it is coordinate descent": we use coordinate descent there to indicate optimization procedures that search along the direction of one basis vector at each iteration, in contrast to the procedures whose update directions do not necessarily align with single basis vectors. Note that, unlike in a Cartesian coordinate system whose basis vectors are orthogonal one-hot vectors, ours are some prior gradients thus it is slightly different from the scenario of ZOO.

2 To Reviewer #2

We appreciate the suggestion about introducing a new threat model, and we agree that it is realistic to consider the cost of training reference models when performing attacks. We will carefully explain and comment on the suggested threat model in an updated version of this paper for comparison fairness. Yet, here we would also like to mention softly that it might be a bit subjective to evaluate the cost of the reference models, since querying a victim model can be sensitive [1] and costly (in money), depending on where it is hosted. Specifically, when targeting at a system which is expensive to query, an adversary may still tend to train reference models even if the number of target points is small.

3 To Reviewer #3

To the comment "…may not have access to such large amount of auxiliary images…": we explain that our method does not always require a large number of images to train reference models. It is shown in Table 2 in our paper how the number of auxiliary images would affect the attack failure rate and query efficiency. While more auxiliary images are always preferred to enhance the reference models and further reduce the query complexity for our method, thousands (or even hundreds of) images can still be beneficial to achieve superior performance to previous state-of-the-arts.

To the comment "…similar ideas are available online…": we appreciate the pointer to this contemporaneous work. It seems that the work is available online after the conference submission deadline, and we shall discuss about it in an updated version. Our experimental results are more significant than theirs in two aspects: 1) we utilize far less training images (only 75K) for ImageNet that are strictly unseen by the victim model to obtain our reference models while the standard training set (with 1.2M images) is adopted in the contemporaneous work, which means our training cost is also much cheaper, 2) we demonstrate that our method outperforms existing competitors in both targeted and untargeted settings while it only considers the untargeted setting, (note that as pointed in some papers [2], the ineffectiveness under a targeted setting can be a main drawback in transfered-based attacks).

References

[1] S. Chen, N. Carlini, and D. Wagner. Stateful detection of black-box adversarial attacks. *arXiv preprint arXiv:1907.05587*, 2019.

[2] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.