We are very grateful to the positive feedback and constructive comments. Minor comments will be addressed accordingly and our responses to the major comments are provided as follows.

Visualization and explanation: We have provided some more visualization and explanation results in the below figure. It can be seen that the no-box adversarial examples crafted on the auto-encoding models are indeed likely to be different from the adversarial examples crafted on a pre-trained ResNet-50 (i.e., Beyonder) as in the white-box/black-box settings. Particularly, visual artifacts (somewhat like moiré patterns) may present in the no-box adversarial examples under ϵ=0.1. Illustration results under $\epsilon = 0.08$ and $\epsilon = 0.03$ (where the perturbations are more imperceptible) will also be given, in an updated version of the paper.

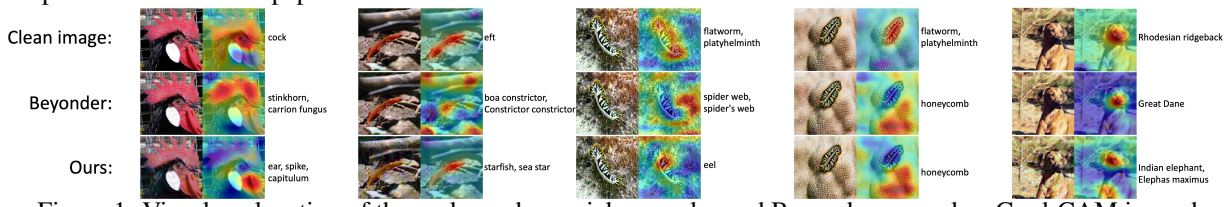

Figure 1: Visual explanation of the no-box adversarial examples and Beyonder examples. Grad-CAM is used.

To Reviewer #1:

1) Regarding our claim of "training on different data from that of the victim models further leads to low attack success rates", we meant to say that the adversarial example crafted on models (with loss and accuracy shown in Figure 1 in the paper) trained using a conventional supervised manner transferred poorly to the victim models. The experimental results were given in Table 1 in the paper (reported as the "naive†" method). We will revise the related content for clarify.

2) **Q:** I am a little bit curious about how other transferable attacks can help here (e.g., replace ILA with [48,56]). **A:** We tested with TAP [56] and observed that our mechanisms also worked well. Specifically, our prototypical mechanism led to an average victim accuracy of 28.82% on ImageNet with TAP, which is remarkably superior to naive† (77.39%).

3) **Q:** Diagnoses on the adversarial perturbations generated by attacking auto-encoders? Is $\epsilon = 0.1$ noticeable? **A:** We have provided more visualizations of the generated adversarial examples in this response letter. ℓ_∞ perturbations under a constraint of ϵ=0.1 may be perceptible on some images, therefore we have also reported attack performance under ϵ=0.08 in the supplementary material of the paper and ϵ=0.03 in our response to Reviewer #3. Our method still works significantly better than the concerned supervised and unsupervised baselines in these settings.

To Reviewer #2:

1) **Q:** Table 1 does not contain any sort of error bar. **A:** We here report the standard deviation of multiple training and evaluations in the below table, and it can now be inferred which models are more susceptible to our no-box attacks.

Table 1: The standard deviation of the prediction accuracy on adversarial examples on ImageNet.

Method	Sup.	VGG-19	Inception v3	ResNet	DenseNet	SENet	WRN	PNASNet	MobileNet	Average
Naïve‡	✗	4.59%	5.50%	4.98%	5.08%	5.37%	4.80%	5.17%	3.73%	4.90%
Jigsaw	✗	3.05%	4.66%	4.52%	3.83%	4.70%	3.80%	4.54%	2.47%	3.95%
Rotation	✗	3.46%	4.36%	3.90%	3.97%	3.81%	4.94%	4.41%	2.62%	3.93%
Naïve†	✓	5.66%	5.61%	5.32%	6.59%	4.42%	5.22%	4.55%	6.55%	5.49%
Prototypical	✓	2.61%	3.47%	3.48%	3.55%	4.03%	2.92%	3.71%	2.28%	3.26%
Prototypical*	✓	2.01%	3.03%	3.35%	2.60%	3.29%	3.39%	4.22%	1.70%	2.95%

2) **Q:** Whether results hold beyond the ℓ_∞ norm? **A:** We further considered ℓ_2 attacks. Specifically, by restricting the ℓ_2 norm of the perturbations to be not greater than a common threshold, our prototypical mechanism led to a significantly lower average prediction accuracy (59.48%) of the victim models, in comparison to the supervised baseline (81.37%).

To Reviewer #3:

1) **Q:** With smaller ϵ values, the success rates of no-box attacks diminishes. **A:** Indeed, under stricter constraints, the performance of all the concerned attacks degrades. Yet, our prototypical mechanism still led to a much lower average prediction accuracy (77.54%) of the victim models, than that of the supervised baseline (84.33%), under ϵ=0.03.

2) **Q:** Lower test loss but lower test accuracy at first several hundred iterations. **A:** With very limited training data, the supervised models became increasingly more confident during the first hundreds of training iterations, therefore several incorrect predictions can lead to an increase of test loss (together with improvement in test accuracy though). Modeling of the internal data distribution acted as a regularization and it actually alleviated this problem to some extent.

3) We followed the suggestion and tested on an adversarially trained ResNet provided by Madry et al.. The superiority of our prototypical mechanism to its competitor held, leading to a lower victim accuracy (39.24% vs 54.12%).

4) We have performed an ablation study on n in Section A in the supplementary material of the submission. It can be seen that by further increasing n to 40, the prototypical mechanism achieved even better attack performance, while the performance of rotation/jigsaw models degraded with larger n, probably on account of slower training convergence.

To Reviewer #4: Thanks for the very positive feedback! We will discuss the mentioned related work.