We thank all reviewers for their comments. All reviewers think it is an interesting paper. R1's review summarizes our contribution well: "DLG is the first to shows a malicious player can recover private training data in collaborative learning scenario." Both R1 and R3 are positive overall in their comments (R1 "easy to read and well structured", "raises an important privacy issue", R3 "easy to read and understand", "it is surprising that obtaining the training datasets is possible by only utilizing the gradients"). For all typos/grammar mistakes, we have revised our writing accordingly.

R2: DLG may not work for accumulated gradients / Contrived settings. This is a misunderstanding: DLG is still effective in federated learning (Tab. 1). In the real-world case, a common workflow is to firstly deliver a pre-trained model to users' devices and fine-tune it by Federated Learning[1]. In this case, the gradient and learning rate are both small, thus the weight changes are small too. Thereby it can be approximated as multi-batch case and this is still possible to attack. Nowadays, noisy / sparse / accumulated gradients are just *optional* choices for training acceleration, but actually they are *essential* techniques to protect the training set. **Our work aims to raise people's awareness about the security of gradients**.

	Iterations=1	Iterations=2	Iterations=3	Iterations=4
MSE	3.3×10^{-6}	3.5×10^{-3}	3.0×10^{-3}	1.8×10^{-2}

Table 1: The effectiveness of DLG on federated learning for different communication frequency.

	Property Inference [26]	DLG
Eyeglasses	0.94	**1.00**
Asian	0.92	**1.00**

Table 2: AUC score on LFW dataset.

R3: Comparison with previous work. To the best of our acknowledge, **DLG is the first algorithm that performs pixel-level and token-level leakage** based on shared gradients. We have compared conventional synthetic outputs and our recovered results in the Fig. 4 in our paper. We also add a comparison on property inference task in Tab. 2: DLG is significantly better since it can directly obtain the raw training data. In the revision, we will add more comparisons.

R1, R2: Trade-off between accuracy and defendability. **R3:** The method is easy to defend. **R1:** Does 8-bit help? DLG is not easy to defend unless with a significant drop in accuracy. 8-bit gradient does not help either. We study the trade-off between accuracy and defendability in Tab. 3. It shows that **only when the defense strategy starts to degrade the accuracy then the deep leakage can be defended**.

	Original	**G**-10^{-4}	**G**-10^{-3}	**G**-10^{-2}	**G**-10^{-1}	**L**-10^{-4}	**L**-10^{-3}	**L**-10^{-2}	**L**-10^{-1}	**FP**-16	**8 bit**
Accuracy	76.3%	75.6%	73.3%	45.3%	≤1%	75.6%	73.4%	46.2%	≤1%	76.1%	53.7%
Defendability	–	✗	✗	✓	✓	✗	✗	✓	✓	✗	✓

Table 3: **G**: Gaussian noise, **L**: Laplacian noise, **FP**: Floating number. ✓ means it successfully defends against DLG while ✗ means fails to defend. The accuracy is evaluated on CIFAR-100, same as what we used in the paper.

R2: Do you use all trainable parameters of the ResNet as ∇W? Is the model W trained to convergence? Gradients of *all* trainable parameters are used as ∇W. It is important to clarify that DLG does *not* require the model trained to converge: **The attack can be performed at any moment during the training** (Tab. 4). Our results in paper are based on randomly initialized models.

Train Progress	0% epochs	30% epochs	70% epochs	100% epochs
MSE	5.7×10^{-6}	3.1×10^{-7}	4.4×10^{-6}	3.3×10^{-6}

Table 4: The MSE between leaked image and ground truth on different training stages. Pixel values are normalized to $[0, 1]$. The leaked image is nearly identical to original ones at each training phase.

R1: The concept of 'iterations' refers to the "n" in the for-loop in DLG algorithm, not the training iterations.

R2: Fig 5. Is the blue line "L2 distance" over all parameters and other lines over parameters of specific layers? No, the L2 distance (on the top of the figure) is measured between the leaked image and original ground truth image. Other lines are the distance between dummy gradients $\nabla W'$ and real gradients ∇W in each layer. We'll make it clear in the paper.

R2: Are qualitative results from a held-out test set? How/why were these images chosen? Yes, qualitative results are from a held-out test set. These images are randomly sampled and more examples have been already provided in the appendix. There is no cherry-picking.

R2: DLG becomes harder (needs more iterations) to attack when batch size increases." Wouldn't this also make for a good defense? How do the reconstruction results vary with batch size? In multi-batch examples (Fig 3 in paper and Line 1 in appendix), we only observe few artifact pixels compared with single-batch cases. Note DLG can be performed off-line as long as current model status and gradients are saved. Large-batch is not a good defense strategy since the information can be still leaked with more time and iterations.

R3: Time cost for L-BFGS reconstruction. Though L-BFGS takes more calculations for every single step, it is still faster (5 minutes) than other optimizers like SGD (30 minutes) on our hardware (Nvidia Tesla v100).

[1] Towards federated learning at scale: System design https://arxiv.org/abs/1902.01046