Appendix

We provide more experimental details in the following sections. First, we elaborate on person identification analysis to evaluate inversion strength in Section A. Then, we demonstrate efficacy of the proposed loss scheduling technique in Section B. We include additional supplementary images for ablation studies in Section C, and more visual examples for GradViT in Section E.

A. Person Identification

In this section, we study the likelihood of person identification as a function of the batch size, and also leverage a StyleGAN2 [18] network to improve the image fidelity. Specifically, we utilize an iterative refinement approach [1] based on a pre-trained StyleGAN2 for latent space optimization and finding the closest real image. Fig. S.3 illustrates the outputs of the latent optimization which uses a GradViT recovered image as an input.

To quantify person identification, we utilize the Image Identifiablity Precision (IIP) as studied by Yin *et al*. [38] to check on the level of data leakage across varying batch sizes. Specifically, we consider a total number of 15000 distinct subjects randomly selected from the MS-Celeb-1M dataset. The experiments are performed once for reconstructions of a given batch size. For IIP calculation, we extract deep feature embeddings using an ImageNet-1K pre-trained ResNet-50. To compute exact matches, we use k-nearest neighbor clustering to sort the closest training images to the reconstructions in the embedding space. The IIP score is computed as the ratio of number of exact matches to the batch size.

We use the outputs of GradViT and GradViT followed by StyleGAN2 for all person identification experiments (See Fig. S.1). We observe that for a batch size of 4, both models can accurately identify subjects with an IIP score of 100%. For a batch size of 8, GradViT and GradViT+StyleGAN2 yield IIP scores of 75% and 87.5% respectively. We observe that enhancements by latent code optimization result in improved facial recovery and hence increased IIP scores. IIP gradually decrease for both cases amid more gradient averaging at larger batch sizes.

B. Loss Scheduler

In Fig. S.2, we demonstrate the effect of our proposed loss scheduler on balancing the training between the gradient matching loss and the image prior. As observed by the progression of optimization from random noise to the final image, gradient matching phase obtains most of the semantics in the image by the mid-training. However, the recovered image lacks detailed information and suffers from low fidelity. By enabling the image prior loss and decreasing the contribution of gradient matching, the visual realism is significantly improved and more fine-grained detailed are recovered.

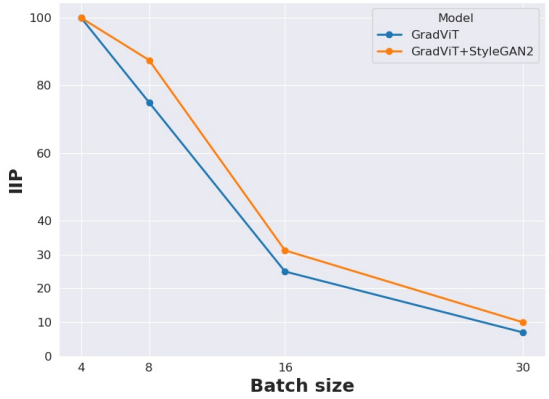

Figure S.1. Effect of batch size on IIP score for recovered images from MS-Celeb-1M dataset. Random guess probability is 0.007% as a reference.

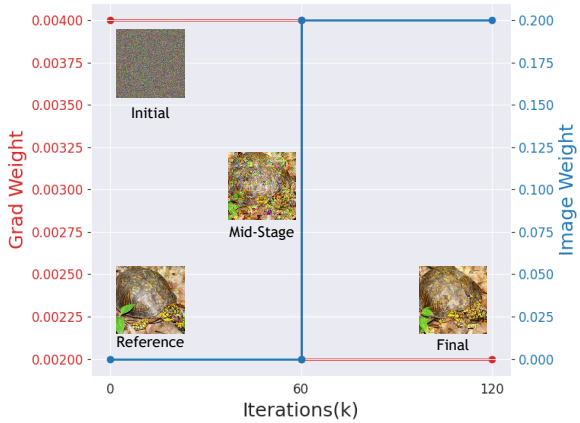

Figure S.2. Effect of loss scheduler on optimization progression of reconstructions.

C. Ablation Studies Additional Examples

We provide additional qualitative visualizations of data leakage analysis by layer-wise (Fig. S.4) and component-wise (Fig. S.5) analysis of ViT/B-16 [9] architecture. Inversion results obtained from batch of 8 images.

D. Limitations

GradViT remains computationally intensive. However, this offers security benefits in reality, given that the associated computation burden may hinder gradient inversion at scale.

E. More Inversion Examples

For ImageNet1K dataset, Fig. S.6 depicts additional recovered images from gradient inversion of vision transformers using GradViT for varying batch sizes. In addition,

Figure S.3. Step-by-step latent optimization of GradViT outputs. Recovered image shown on the right.

Figure S.4. Reconstructed images from layer-wise ablation.

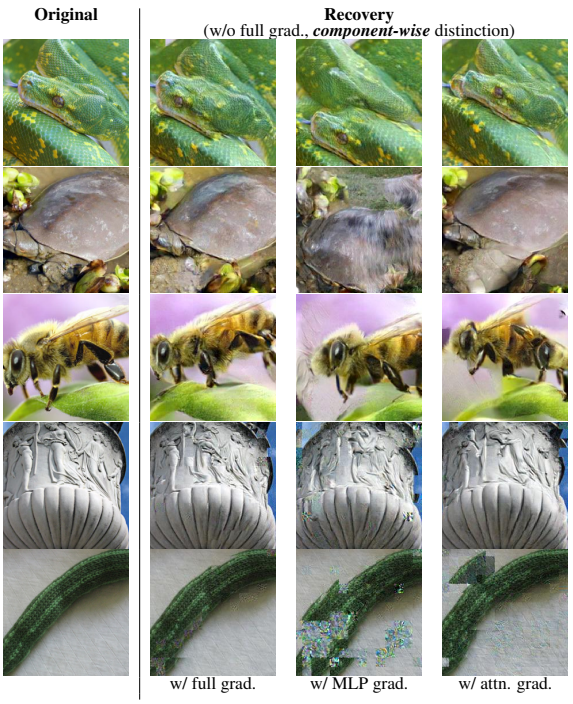

Figure S.5. Reconstructed images from component-wise ablation.

Fig. S.7 demonstrates additional reconstructions from gradient inversion of FaceTransformer [43] model using GradViT for different batch sizes.

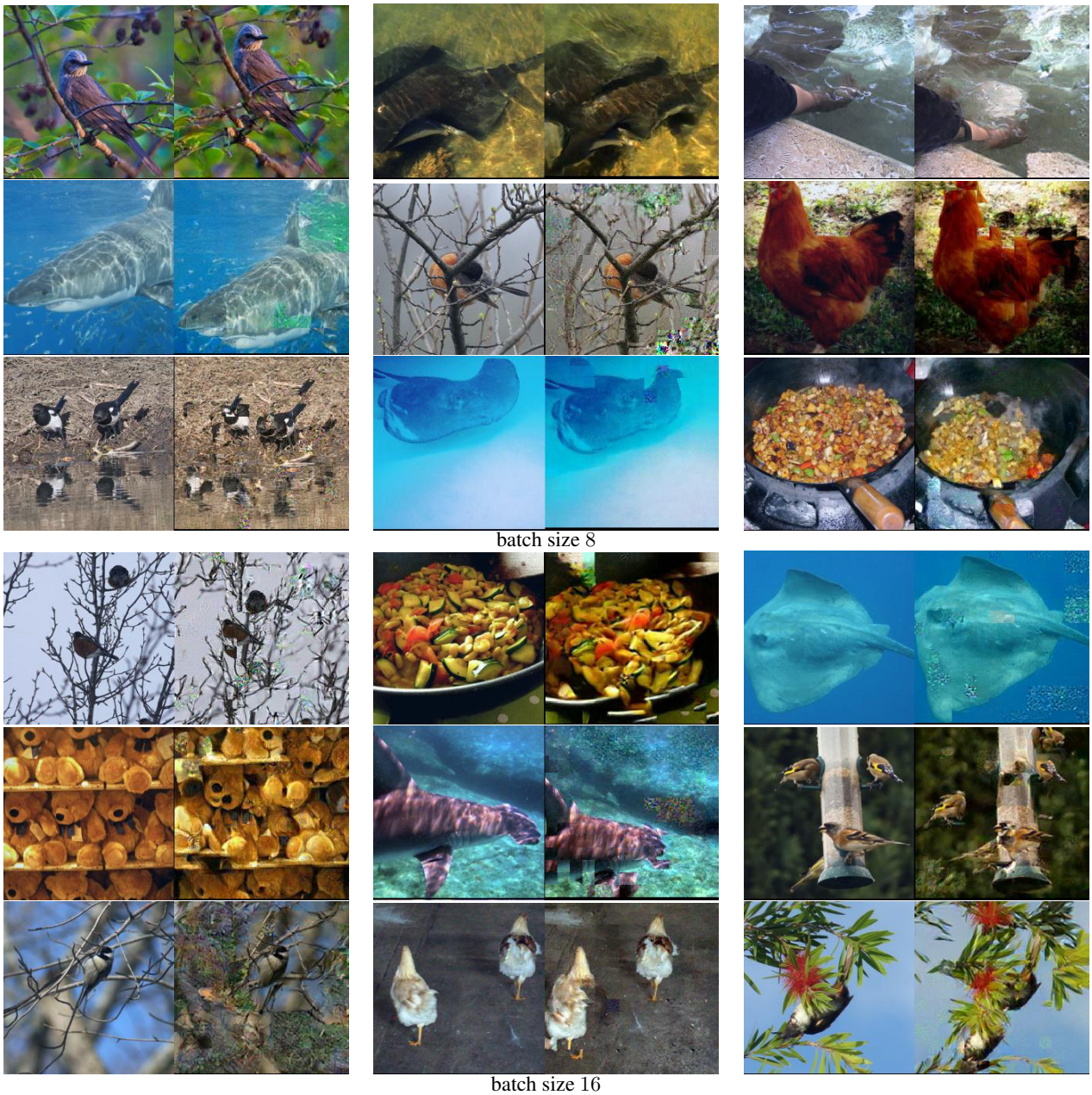

Figure S.6. Inverting ViT-B/16 gradients on the ImageNet-1K validation set. Pair of (left) original sample and its (right) recovery.

F. Face Reconstruction Quantitative Analysis

Table. S.1 presents the quantitative benchmarks of image reconstruction quality for gradient inversion from batch sizes of 4 and 8 using images in MS-Celeb-1M dataset. As expected, for all image reconstruction metrics, the reconstruction quality decreases with increasing batch size.

Batch Size	**Image Reconstruction Metric**		
	PSNR ↑	FFT_{2D} ↓	LPIPS ↓
4	27.370	0.001	0.030
8	23.313	0.008	0.101

Table S.1. Quantitative benchmarks of image reconstruction quality from batch sizes of 4 and 8 images in MS-Celeb-1M dataset.

Figure S.7. Additional examples of information leakage when inverting FaceTransformer gradients on the MS-Celeb-1M validation set. Each block containing a pair of (left) original sample and its (right) reconstruction by GradViT.