ABG: AN ADAPTIVE BLOCK-BASED GRADIENT INVERSION ATTACK IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning (**FL**) aims to protect training data by training models through gradient sharing without directly exchanging raw data. However, gradient inversion attacks (*GIAs*) demonstrate that shared gradients can still reveal sensitive client training data. In this paper, we proposes Adaptive Block-based Gradient inversion attack (***ABG***), an adaptive gradient inversion framework for high-fidelity image reconstruction. Unlike existing *GIAs* that rely on fixed reconstruction strategies and exhibit image-dependent reconstruction performance, ***ABG*** leverages the gradient mean to perform adaptive initialization and adaptively select the residual-layer weighting scheme. It further incorporates image encoding and decoding to reduce the pixel-level optimization space and performs block-wise gradient matching for ResNet architectures. Compared with the current **SOTA** approach, ***ABG*** improves PSNR by up to $\mathbf{5\,dB}$ while reducing reconstruction time by more than $\mathbf{12}\%$. Extensive experiments across diverse settings validate the effectiveness and stability of the proposed adaptive strategy, highlighting the practical privacy risks of gradient leakage in **FL**.

1 INTRODUCTION

Federated learning (**FL**) Konečnỳ et al. (2016) enables multiple clients to train a shared model without centralising their raw data. This data-local training paradigm reduces direct data exposure, but it does not prevent information leakage through the model updates exchanged during optimisation. An *honest-but-curious* server can still reconstruct client data from shared gradients through iterative optimization Zhu et al. (2019); Zhao et al. (2020); Geiping et al. (2020); Wang et al. (2020); Yin et al. (2021); Dong et al. (2021); Hatamizadeh et al. (2022); Yue et al. (2023); Hatamizadeh et al. (2023); Sun et al. (2024); Dao & Lee (2024); Liu et al. (2026), exposing sensitive information. Therefore, gradient inversion attacks (*GIAs*) have received increasing attention in **FL** research.

Iterative *GIAs* reconstruct client data by minimizing the discrepancy between the gradients of the reconstructed data $\mathbf{x}'$ and the ground-truth data $\mathbf{x}$, denoted as $\nabla W'$ and ∇W, respectively Zhu et al. (2019); Geiping et al. (2020); Yin et al. (2021); Liu et al. (2026). Reducing this discrepancy drives $\mathbf{x}'$ toward $\mathbf{x}$, thereby improving reconstruction accuracy. The current state-of-the-art (**SOTA**) method, Two-stage Layer-wise Gradient Inversion Attack (TLG) Liu et al. (2026), reduces the reliance on BN statistics and improves reconstruction efficiency under ResNet architectures He et al. (2016). However, existing methods may still exhibit substantial variation in reconstruction quality across different target images. This limitation indicates that image-dependent reconstruction stability remains insufficiently addressed. Therefore, a more adaptive *GIA* framework is needed to better characterize privacy vulnerabilities in **FL**.

Our objective is to accurately reconstruct client-side training images $\mathbf{x}$ with improved adaptivity, stability, and efficiency. To this end, we propose Adaptive Block-based Gradient inversion attack, denoted as ***ABG***. Unlike existing *GIA* methods that use fixed initialization strategies, such as Gaussian, random, or fixed-color initialization Yin et al. (2021); Hatamizadeh et al. (2023); Ye et al. (2024), ***ABG*** initializes image pixels using the gradient mean $\nabla\overline{W}$, enabling gradient-mean-guided adaptive initialization. It further applies bicubic downsampling and upsampling for image encoding and decoding during gradient matching, thereby reducing the optimization space of $\mathbf{x}'$ at each iteration. To improve matching stability and efficiency, ResNet He et al. (2016) is partitioned into multiple blocks, and gradients from selected layers are matched in a staged manner to update $\mathbf{x}'$. Although

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

residual connections preserve information about $\mathbf{x}$ through identity mappings, deeper residual layers may also retain noise-related information Zeiler & Fergus (2014); Yosinski et al. (2014); Krizhevsky et al. (2017). Therefore, ***ABG*** uses the gradient mean $\nabla\bar{W}$ to adaptively select the residual-layer weighting scheme. A fixed lower weight is applied to the selected residual layers, while the remaining unselected residual and convolutional layers retain a weight of 1. This gradient-mean-guided scheme selection adapts the matching process to image-dependent gradient characteristics and improves reconstruction stability across different target images. Finally, ***ABG*** introduces stage-specific regularization terms to accelerate convergence and improve reconstruction quality.

Experiments under diverse settings show that ***ABG*** outperforms TLG Liu et al. (2026) in reconstruction quality while reducing reconstruction time by more than 12%. Under large-batch settings, ***ABG*** still reconstructs high-resolution images with recognizable object structures and fine-grained details. The main contributions of this paper are summarized as follows:

- We propose an Adaptive Block-based Gradient inversion attack (***ABG***) framework that improves reconstruction adaptivity and stability across different target images while reducing reconstruction time.
- We introduce gradient-mean-guided adaptive initialization. In addition, by incorporating image encoding and decoding, ***ABG*** reduces the pixel-level optimization space and adapts the initialization to image-dependent gradient characteristics.
- We design a block-wise gradient matching mechanism for ResNet architectures and use the gradient mean $\nabla\bar{W}$ to adaptively select the residual-layer weighting scheme, thereby improving reconstruction stability and quality across different target images.
- Extensive experiments across multiple architectures, datasets, batch-size settings, and defense strategies validate the effectiveness and adaptive reconstruction capability of ***ABG***, while highlighting privacy risks in **FL**.

2 Related Work

Federated learning (**FL**) is a distributed learning paradigm designed to preserve data privacy and has attracted significant research attention. However, recent studies have revealed potential privacy vulnerabilities in **FL** Zhu et al. (2019); Zhao et al. (2020); Geiping et al. (2020); Wang et al. (2020); Yin et al. (2021); Dong et al. (2021); Hatamizadeh et al. (2022); Yue et al. (2023); Hatamizadeh et al. (2023); Zhu & Blaschko (2020); Chen & Campbell (2021); Liu et al. (2024); Ye et al. (2024); Carletti et al. (2025); Liu et al. (2026) . In particular, shared gradients during training can enable attackers to reconstruct ground-truth data, posing substantial privacy risks.

Since 2019, gradient inversion attacks (*GIAs*) have become an important topic in **FL** privacy research. Iterative *GIAs* are widely studied because they require fewer architectural assumptions and are broadly applicable to **FL**. DLG Zhu et al. (2019) first shows that client data can be reconstructed from shared gradients by optimizing dummy inputs and labels with an ℓ_2 distance objective. iDLG Zhao et al. (2020) infers labels from gradient information using cross-entropy loss, reducing label ambiguity but remaining limited when repeated labels appear in a batch. IG Geiping et al. (2020) introduces cosine similarity loss to align gradient directions and improve reconstruction quality. GradInversion Yin et al. (2021) incorporates $\mathbf{R_{BN}}$ and group regularization to enhance reconstruction fidelity and support multi-batch label recovery. However, its reliance on BN statistics introduces a strong assumption Chen & Campbell (2021); Wang et al. (2023); Guerraoui et al. (2024), and group regularization increases computational overhead. ROG Yue et al. (2023) reduces optimization complexity by encoding images into low-resolution representations and decoding them to the original resolution, but its evaluation is mainly limited to small-scale models. HFG Ye et al. (2024) performs hierarchical layer-wise matching from deep to shallow layers and uses $\mathbf{R_{BN}}$ to improve reconstruction fidelity. TLG Liu et al. (2026) using different stages to matching gradients but it exhibits unstable reconstruction performance across different raw images.

Recursive *GIAs* reconstruct $\mathbf{x}$ layer by layer using analytical gradients at each model layer Zhu & Blaschko (2020); Chen & Campbell (2021); Carletti et al. (2025); Kapoor & Kumar (2025). These methods avoid extensive iterative optimization and can be efficient. However, their dependence on layer-wise analytical inversion usually restricts them to simple architectures and batch size one,

limiting their applicability in practical **FL** deployments. To address these limitations, this work proposes a more general *GIA* framework that reduces reliance on strong assumptions while improving reconstruction efficiency and applicability.

3 METHODOLOGY

This section presents the design of ***ABG***. We first introduce the attack model and the objective function of iteration-based *GIAs*, then describe the overall framework of ***ABG***, and finally provide its complete objective formulation.

3.1 THREAT MODEL

We consider a standard **FL** process based on FedAvg Li et al. (2019); Collins et al. (2022); Li et al. (2025), where an *honest-but-curious* server acts as the potential attacker. This setting follows the threat model commonly adopted in *GIA* studies Zhu et al. (2019); Geiping et al. (2020); Yin et al. (2021); Ye et al. (2024); Liu et al. (2026). The server follows the **FL** protocol and does not modify local data or training outputs, but has access to the model architecture, model parameters, and client-uploaded gradients. In this framework, a central server and multiple clients collaboratively train a global image classification model f_θ. In each communication round, selected clients train local models using private samples $\mathbf{x}$ and labels $\mathbf{y}$. Each client computes a local gradient ∇W on its own device and uploads it to the server. The server aggregates the received gradients to update the global model f_θ, which is then broadcast to the selected clients for the next round.

The *honest-but-curious* server acts as the attacker and attempts to reconstruct $\mathbf{x}$ from the shared gradients ∇W_i. Under this threat model, the objective of iteration-based *GIAs* is formulated as

$$(\mathbf{x}', \mathbf{y}') = \arg \min_{(\mathbf{x}', \mathbf{y}')} \mathcal{L}\left(\nabla W', \nabla W\right), \tag{1}$$

$$\nabla W' = \nabla_\theta \mathcal{L}(f_\theta(\mathbf{x}'), \mathbf{y}'). \tag{2}$$

Here, $(\mathbf{x}', \mathbf{y}')$ denotes the reconstructed data and corresponding label, and $\mathcal{L}$ measures the discrepancy between the reconstructed gradient $\nabla W_i'$ and the shared gradient ∇W_i. Eq. 1 minimizes this discrepancy and defines the basic optimization objective of iteration-based *GIAs* Zhu et al. (2019); Zhao et al. (2020); Geiping et al. (2020); Wang et al. (2020); Yin et al. (2021); Dong et al. (2021); Hatamizadeh et al. (2023); Yue et al. (2023); Ye et al. (2024); Liu et al. (2024); Liu & Wang (2024); Carletti et al. (2025); Liu et al. (2026). Following prior studies Yin et al. (2021); Ye et al. (2024), image labels $\mathbf{y}'$ can be recovered with nearly 100% accuracy. Therefore, this work assumes accurate recovery of the labels corresponding to the target images, and ***ABG*** focuses on reconstructing $\mathbf{x}'$.

3.2 ADAPTIVE BLOCK-BASED GRADIENT INVERSION ATTACK (***ABG***) STRATEGY

To improve reconstruction adaptivity and stability, while reducing optimization overhead, we propose ***ABG***, an Adaptive Block-based Gradient inversion attack framework. Its core design uses the gradient mean to adapt image initialization and residual-layer weighting scheme selection to image-dependent gradient characteristics.

Random initialization Zhu et al. (2019); Zhao et al. (2020); Geiping et al. (2020); Yin et al. (2021); Ye et al. (2024); Liu et al. (2026) and fixed-color initialization, such as red or green Wang et al. (2020); Zhang et al. (2022); Liu et al. (2024); Liu & Wang (2024), can yield inconsistent reconstruction quality across target images. ***ABG*** addresses this limitation by initializing $\mathbf{x}'$ using the mean of non-negative entries in gradient ($\bar{\nabla W}$), thereby adapting initialization to the observed gradient distribution. Empirically, this mean is typically on the order of 10^{-3} or smaller, whereas image pixel values are generally constrained to $[0, 1]$. Therefore, the mean is scaled by 100 to obtain the initial pixel value. The initialized image is then downsampled to 121×121 to reduce the pixel-level optimization space. To reduce the pixel-level optimization space, the number of optimized variables, and reconstruction time, $\mathbf{x}'$ is upsampled to the resolution of $\mathbf{x}$, e.g., 224×224, using bicubic interpolation at each gradient-matching step for gradient computation and objective evaluation, while optimization remains in the low-resolution space.

During iterative reconstruction, ***ABG*** adopts a four-step matching strategy based on the ResNet block structure. Step I matches all gradients in Block4 during the first T_1 iterations. Step II matches

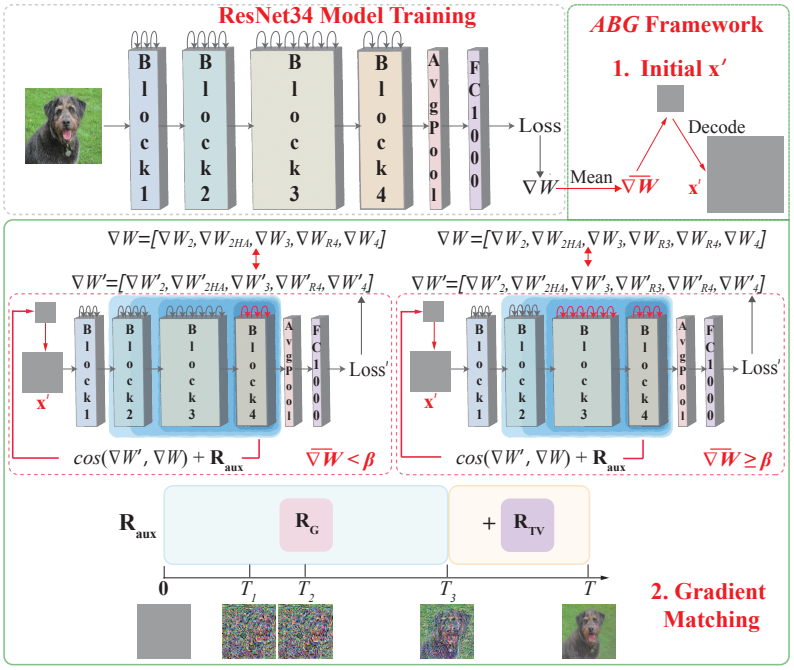


Figure 1: Overview of the ***ABG*** framework. ***ABG*** consists of two components: gradient-mean-based image initialization and gradient matching. The gradient matching process is further divided into two schemes according to $\nabla\bar{W}$. When $\nabla\bar{W} \geq \beta$, lower weights are assigned to the residual layers in Block4 and Block3; otherwise, lower weights are assigned only to the residual layers in Block4. The objective function introduces $\mathbf{R_{TV}}$ after the T_3-th iteration.

gradients in Block4 and Block3 during the first T_2 iterations. Step III further includes the latter half of Block2 during the first T_3 iterations. Step IV extends gradient matching to all gradients in Block2 from iteration T_3 to T. Since residual connections bypass convolutional layers, they may preserve both image-related information and noise, thereby affecting reconstruction quality. Accordingly, ***ABG*** uses the gradient mean $\nabla\bar{W}$ to adaptively select the residual-layer weighting scheme. When $\nabla\bar{W} < \beta$, a fixed weight of 0.1 is assigned to the residual layers in Block4; otherwise, the same weight is assigned to the residual layers in both Block4 and Block3.The weights of the remaining unselected residual layers and convolutional layers are kept as 1. This gradient-mean-guided scheme selection adapts the matching process to different target images and improves reconstruction stability.

3.3 Objective Formulation of *ABG*

Regularization terms have been shown to improve the reconstruction quality of $\mathbf{x}'$ Geiping et al. (2020); Yin et al. (2021); Dong et al. (2021); Yue et al. (2023); Ye et al. (2024); Liu et al. (2024); Carletti et al. (2025). To evaluate privacy vulnerabilities in **FL** without strong assumptions, we exclude BN information from the regularization design and impose only reasonable constraints. The resulting terms use only the gradients ∇W and labels $\mathbf{y}' = \mathbf{y}$ available to the server. The optimization objective of ***ABG*** for $\mathbf{x}'$ is defined as

$$\mathbf{x}' = \arg\min_{\mathbf{x}'} \left[\lambda_l \mathcal{L}(\nabla W', \nabla W) + \mathbf{R_{aux}}\right], \tag{3}$$

where $\mathbf{R_{aux}}$ denotes the regularization terms and λ_l is the weighting coefficient of $\mathcal{L}$. ***ABG*** adopts a cosine similarity loss Geiping et al. (2020); Zhang et al. (2022); Ovi & Gangopadhyay (2023); Liu & Wang (2024); Liu et al. (2024); Ye et al. (2024) to measure the alignment between $\nabla W'$ and ∇W:

$$\cos(\nabla W', \nabla W) = 1 - \frac{\langle \nabla W, \nabla W' \rangle}{\|\nabla W\| \|\nabla W'\|}. \tag{4}$$



(a) Raw (b) DLG (c) GP (d) HFG (e) TLG (f) ***ABG***

Figure 2: Reconstructed images on ImageNet datasets images under a batch size of 1. Each column represents a different reconstruction method and each row corresponds to a specific target image which class is 0, 50, 100 and 150 and used for evaluation.

In the early phase, the objective is to preserve the semantic correctness of the primary visual subject and avoid semantic reconstruction errors, such as reconstructing a dog image as a cat. Accordingly, a modified Gaussian kernel regularization term $\mathbf{R_G}$ Liu et al. (2026) is introduced:

$$\mathbf{R_G}(\nabla W') = \exp\left(-\frac{\|\nabla W'\|}{2}\right). \tag{5}$$

$\mathbf{R_G}$ accelerates the recovery of the main image structure and promotes spatial smoothness by suppressing sharp gradient variations. In the later phase, the optimization focus shifts to correcting background and color information in the reconstructed image. Accordingly, the total variation regularization term $\mathbf{R_{TV}}$ Rudin et al. (1992); Mahendran & Vedaldi (2015); Nguyen et al. (2015); Geiping et al. (2020); Yin et al. (2021); Liu et al. (2024); Ye et al. (2024) is introduced after T_3. The cosine similarity loss aligns $\nabla W'$ with ∇W, $\mathbf{R_G}$ suppresses sharp gradient variations, and $\mathbf{R_{TV}}$ reduces large differences between neighboring pixels. Together, these terms improve local consistency and overall fidelity. Therefore, the complete objective of ***ABG*** is formulated as:

$$\mathbf{x}' = \lambda_l \cos(\nabla W', \nabla W) + \mathbf{R_{aux}} \tag{6}$$

where

$$\mathbf{R_{aux}} = \begin{cases} \lambda_{\mathrm{G}}\mathbf{R_G}, & 1 \le t \le T_3, \\ \lambda_{\mathrm{G}}\mathbf{R_G} + \lambda_{\mathrm{TV}}\mathbf{R_{TV}}, & T_3 < t \le T, \end{cases} \tag{7}$$

where $T_3 \in [1, T]$, and λ_G and λ_{TV} denote the weighting coefficients of $\mathbf{R_G}$ and $\mathbf{R_{TV}}$, respectively. This objective suppresses noise and improves reconstruction fidelity, optimization stability, and efficiency. The execution pipeline is illustrated in Fig. 1, and the corresponding pseudocode is provided in the Appendix B.

4 Experiments

This section first describes the experimental settings and hyperparameters of ***ABG***, including datasets, hardware, and evaluation metrics. We then compare ***ABG*** with representative baselines and the current **SOTA** method in terms of visual reconstruction quality, quantitative metrics, and reconstruction time. Next, we evaluate its adaptive reconstruction performance under different batch sizes. Finally, we assess ***ABG*** under defense settings, including gradient pruning and differential privacy (DP), to examine whether its adaptive reconstruction capability remains effective under perturbed gradients.

Table 1: Metrics of reconstructed images in Fig. 2.

Figure	Metric				
	MSE ↓	PSNR ↑	LPIPS ↓	SSIM ↑	RT ↓
(b) DLG	0.05	13.52	0.85	0.07	1956
	0.11	9.64	0.80	0.03	1939
	0.09	10.61	0.80	0.05	1929
	0.06	12.04	0.71	0.06	1927
(c) IG	0.03	14.62	0.69	0.31	1512
	0.09	10.55	0.62	0.14	1511
	0.07	11.35	0.67	0.20	1521
	0.06	12.35	0.62	0.32	1519
(d) HFG	0.03	15.27	0.42	0.22	856
	0.05	12.83	0.35	0.25	864
	0.03	15.90	**0.25**	0.39	883
	0.07	11.77	0.61	0.18	845
(e) TLG	**0.01**	19.60	0.40	0.34	719
	0.03	14.62	0.32	0.29	713
	0.02	16.50	0.27	0.37	720
	0.04	14.28	0.58	0.27	719
(f) *ABG*	**0.01**	**20.82**	**0.34**	**0.42**	**641**
	0.02	**16.42**	**0.28**	**0.40**	**627**
	0.02	**17.97**	0.27	**0.42**	**645**
	0.01	**19.63**	**0.33**	**0.48**	**674**

4.1 Setup and Evaluation Metrics

Unless otherwise specified, ResNet34 He et al. (2016) is used as the **FL** training model. We evaluate ***ABG*** on CelebA Liu et al. (2015), ILSVRC 2012 Deng et al. (2009), and the WEB dataset introduced by HFG Ye et al. (2024). All images are resized to 224×224 pixels. The iteration thresholds are $T_1 = 3K$, $T_2 = 5K$, $T_3 = 10K$, and $T = 15K$. Image reconstruction is optimized using Adam with a learning rate of 0.3. The loss coefficients are set to $\lambda_l = 3$ and $\lambda_G = 5$, while the total variation coefficient is set to $\lambda_{TV} = 0.1$ after iteration $10K$. For adaptive gradient matching, ***ABG*** uses $\beta = 0.49$ as the threshold for adaptive residual-layer weighting and sets the selected residual-layer weight to 0.1. The initialized image is encoded to 121×121 and decoded to 224×224 using bicubic interpolation for gradient computation and objective evaluation. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

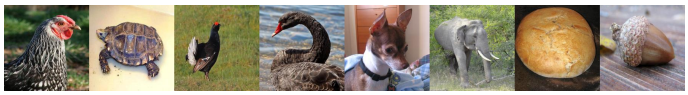
Reconstruction performance is evaluated using qualitative comparison, Mean Squared Error (MSE ↓), Peak Signal-to-Noise Ratio (PSNR ↑), Learned Perceptual Image Patch Similarity (LPIPS ↓), Structural Similarity Index Measure (SSIM ↑), and Reconstruction Time (RT ↓), measured in seconds. The arrows indicate the preferred direction of each metric.

4.2 Reconstruction with Batch Size = 1

We compare ***ABG*** with DLG Zhu et al. (2019), IG Geiping et al. (2020), HFG Ye et al. (2024), and the current **SOTA** *GIA* method TLG Liu et al. (2026). For fairness, the number of restarts in IG is set to 1, consistent with the other methods. All methods are evaluated with a batch size of 1.

To reduce image-specific variation, we use the first images from ImageNet classes 0, 50, 100, and 150. Qualitative and quantitative results are reported in Fig. 2 and Table 1. DLG mainly produces noise-like outputs and fails to recover coherent structures, while IG recovers coarse object shapes but exhibits substantial blurring and structural distortion. HFG improves object recognizability and localization, but visible color deviations and background artifacts remain. TLG further improves structural and color recovery over HFG; however, its reconstruction quality still varies across target images, including noticeable spatial misalignment in the third example.

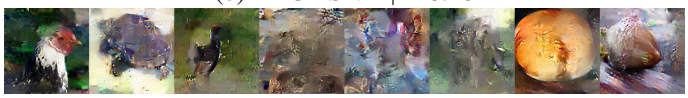
In comparison, ***ABG*** preserves object structures, spatial locations, and color distributions more consistently across different target images. As shown in Table 1, ***ABG*** achieves PSNR values of 20.82,



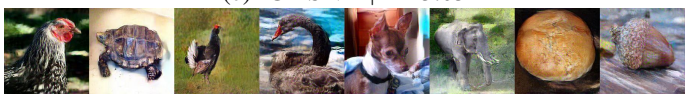
(a) Raw



(b) DLG PSNR ↑ = 8.48



(c) IG PSNR ↑ = 13.05



(d) HFG PSNR ↑ = 15.20



(e) TLG PSNR ↑ = 16.10



(f) ***ABG*** **PSNR ↑ = 16.78**

Figure 3: Reconstructed images on ImageNet under a batch size of 8. Each row corresponds to a different method. ***ABG*** achieves the best reconstruction performance.

16.42, 17.97, and 19.63, outperforming TLG by 1.22, 1.80, 1.47, and 5.35 dB, respectively. It also obtains the best or competitive results on the remaining metrics. These results indicate that the adaptive initialization and matching strategies of ***ABG*** improve reconstruction stability across image-dependent gradient distributions. Tables 1 shows that the reconstruction time of ***ABG*** remains around 640 seconds, reducing time by more than **12%** compared with TLG. Table 1 also shows reductions of approximately **67.7%** and **58.9%**, **27.4%** compared with DLG, IG and HFG, respectively. The reduction is mainly attributed to the block-wise matching strategy, which limits gradient matching to selected layers at different stages and reduces unnecessary optimization.

Due to space limitations, we further evaluate ***ABG*** on CelebA and the WEB dataset introduced by HFG in the Appendix C. The results across ImageNet, CelebA, and WEB demonstrate that ***ABG*** maintains stable reconstruction quality across heterogeneous target images, further validating its adaptive reconstruction capability.

4.3 Reconstruction with Batch Size = 8

As shown in Fig. 3, we evaluate reconstruction performance on ImageNet with a batch size of 8. DLG produces predominantly noise-like outputs and fails to recover recognizable object structures. IG recovers coarse semantic content but exhibits substantial blurring and structural distortion. HFG substantially improves object recognizability over DLG and IG, although visible background artifacts and color inconsistencies remain. TLG further reduces color distortion and improves structural recovery over HFG. However, its reconstruction quality still varies across target images; in particular, the third reconstructed object exhibits noticeable spatial misalignment. In comparison, ***ABG*** preserves object structures, spatial layouts, and color distributions more consistently across different target images. The reconstructed foreground objects are accurately localized, while background information and object boundaries remain comparatively stable. This consistency indicates that the adaptive initialization, matching, and residual-layer weighting strategies of ***ABG*** can accommodate image-dependent gradient characteristics under the batch-size-8 setting. Quantitatively, DLG, IG, HFG, and TLG achieve PSNR values of 8.48, 13.05, 15.20, and 16.10, respectively, whereas ***ABG*** achieves the highest PSNR of 16.78. The qualitative and quantitative results therefore demonstrate



(a) Raw (b) *cos* (c) $\mathbf{R_G}$ (d) $\mathbf{R_{TV}}$ (e) R2 (f) ***ABG***

Figure 4: The ablation results of ***ABG***. From left to right, the columns denote the regularization terms or design components added to each variant. R2 denotes weighting Block2 residual layers in ***ABG***.

Table 2: Metrics of reconstructed images in Fig. 4.

Figure	Metric			
	MSE ↓	PSNR ↑	LPIPS ↓	SSIM ↑
(b) *cos*	0.13	8.87	0.72	0.04
	0.14	8.41	0.78	0.03
(c) $\mathbf{R_G}$	0.04	13.51	0.30	0.14
	0.14	8.50	0.78	0.03
(d) $\mathbf{R_{TV}}$	0.05	13.05	0.48	0.24
	0.05	13.39	0.57	0.20
(e) R2	**0.02**	17.44	0.24	0.43
	0.05	13.04	0.60	0.18
(f) ***ABG***	**0.02**	**17.87**	**0.22**	**0.44**
	0.01	**19.50**	**0.31**	**0.41**

that ***ABG*** maintains more stable and accurate reconstruction across heterogeneous target images, further validating its adaptive reconstruction capability under larger-batch conditions.

4.4 Ablation Results

As shown in Fig. 4 and Table 2, using only the cosine similarity loss (*cos*) produces noisy reconstructions with incomplete structures. Adding $\mathbf{R_G}$ improves structural recovery, while $\mathbf{R_{TV}}$ further suppresses artifacts and improves visual consistency. The R2 additionally applies the reduced residual-layer weight to Block2, but does not consistently improve reconstruction and reduces the PSNR of the third image to 13.04. In contrast, ***ABG*** uses the gradient mean $\nabla\bar{W}$ to adaptively select whether the reduced residual-layer weight is applied to Block4 only or to both Block4 and Block3. It achieves the highest PSNR values of 17.87 and 19.50 for the three target images, together with the best or comparable results for the remaining metrics. These results validate the effectiveness of gradient-mean-guided adaptive initialization and residual-layer weighting-scheme selection in improving reconstruction quality and stability across different target images. Since HFG Ye et al. (2024) and TLG Liu et al. (2026) have shown that including Block1 can degrade reconstruction, and initialization strategies have been extensively studied, we report these additional ablations, together with the encoding-decoding ablation, in the Appendix D due to space limitations.

4.5 Adaptive Residual Layer

As shown in Fig. 5, we evaluate the residual-layer weighting strategy using the first ImageNet image sampled at intervals of 25 classes. The PSNR difference between weighting Block4 (R4, blue circles in Fig. 5) and weighting both Block4 and Block3 (R4+R3, orange triangles in Fig. 5) varies across target images, indicating that a fixed weighting scheme is not consistently optimal. To address this variation, ***ABG*** uses the gradient mean $\nabla\bar{W}$ (red squares in Fig. 5) as the adaptive criterion for scheme selection. Based on the empirical results, the threshold is set to $\beta = 0.49$, indicated by the red dashed line in Fig. 5. When $\nabla\bar{W} < \beta$, ***ABG*** assigns a weight of 0.1 to the residual layers in Block4; otherwise, the same weight is assigned to the residual layers in both Block4 and Block3. The remaining convolutional-layer weights are kept as 1. This gradient-mean-guided se-

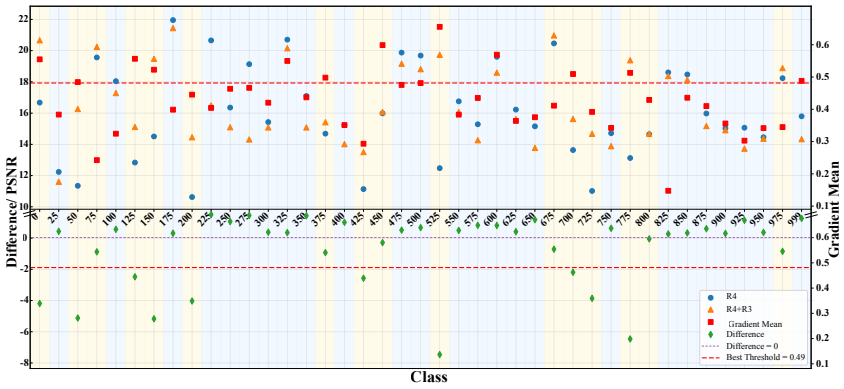


Figure 5: Relationship between the gradient mean and residual-layer weighting in ResNet-based reconstruction. Blue circles and orange triangles denote the PSNR values after weighting Block4 (R4) and in both Block4 and Block3 (R4+R3), respectively. Green diamonds show the PSNR difference, and red squares indicate the gradient mean. The red dashed line marks the threshold 0.49. Blue and yellow backgrounds indicate cases where R4 and R4+R3 achieve higher PSNR, respectively.



(a) Raw (b) 18 (c) 50 (d) 32 (e) 90% (f) 0.1

Figure 6: Reconstruction results of ***ABG*** on ImageNet under different experimental settings. (b) and (c) correspond to ResNet18 and ResNet50, respectively; (d) uses a batch size of 32; (e) applies 90% gradient pruning; and (f) applies differential privacy with Gaussian noise ($\sigma = 0.1$).

lection enables ***ABG*** to adapt residual-layer weighting to image-dependent gradient characteristics, improving reconstruction stability across different target images. Complete visual reconstruction results are provided in the Appendix E.

4.6 Results in Additional Scenarios

To evaluate the generalizability of ***ABG*** and its robustness against existing defense strategies, we conduct experiments under different model architectures, a large-batch setting, and defense settings, including gradient pruning and differential privacy. The results are shown in Fig. 6. The experimental results show that ***ABG*** can effectively recover the original images under ResNet18 and ResNet50, a batch size of 32, 90% gradient pruning, and differential privacy with Gaussian noise injection ($\sigma = 0.1$). These results indicate that ***ABG*** maintains strong reconstruction capability across different architectures, batch-size settings, and defense scenarios, highlighting the urgency of developing stronger defense mechanisms for **FL**.

5 Conclusion

In this paper, we propose ***ABG***, an Adaptive Block-based Gradient inversion attack framework for federated learning. ***ABG*** leverages the gradient mean to adaptively initialize reconstructed images and incorporates image encoding and decoding to reduce the pixel-level optimization space. It further performs block-wise gradient matching for ResNet architectures and adaptively assigns residual-layer weights according to the gradient mean, thereby improving reconstruction stability across different target images. Stage-specific regularization terms are introduced to improve structural recovery and visual-detail refinement. Extensive experiments across different architectures, datasets, batch-size settings, and defense strategies validate the effectiveness and adaptivity of ***ABG***, highlighting the persistent privacy risks posed by gradient leakage in **FL**.

AI USE STATEMENT

In this work, we used generative AI tools to refine the manuscript and identify recent relevant literature in this research area. Generative AI tools are also used to assist in improving the presentation of Section 2. We use generative AI tools to explore preliminary approaches and assist with code generation. The authors review, verify, and test all AI-generated code for correctness before incorporating it into this work. In Section 4.5, generative AI tools assist with data processing and the revision of Fig. 5. We have not used generative AI tools to generate datasets. The use of AI to prove mathematical propositions is not applicable to this study. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Vincenzo Carletti, Pasquale Foggia, Carlo Mazzocca, Giuseppe Parrella, and Mario Vento. SoK: Gradient inversion attacks in federated learning. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 6439–6459, 2025.

Cangxiong Chen and Neill DF Campbell. Understanding training-data leakage from gradients in neural networks for image classification. *arXiv preprint arXiv:2111.10178*, 2021.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Fedavg with fine tuning: Local updates lead to representation learning. *Advances in Neural Information Processing Systems*, 35:10572–10586, 2022.

Thi-Nga Dao and Hyungjune Lee. Encgradinversion: Image encoding and gradient-inversion-based batch attack in federated learning. *IEEE Internet of Things Journal*, 12(4):3858–3870, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Xin Dong, Hongxu Yin, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. Deep neural networks are surprisingly reversible: a baseline for zero-shot inversion. *arXiv preprint arXiv:2107.06304*, 2021.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients - how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 16937–16947. Curran Associates, Inc., 2020.

Rachid Guerraoui, Rafael Pinot, Geovani Rizk, John Stephan, and François Taiani. Overcoming the challenges of batch normalization in federated learning. *arXiv preprint arXiv:2405.14670*, 2024.

Ali Hatamizadeh, Hongxu Yin, Holger R Roth, Wenqi Li, Jan Kautz, Daguang Xu, and Pavlo Molchanov. Gradvit: Gradient inversion of vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10021–10030, 2022.

Ali Hatamizadeh, Hongxu Yin, Pavlo Molchanov, Andriy Myronenko, Wenqi Li, Prerna Dogra, Andrew Feng, Mona G Flores, Jan Kautz, Daguang Xu, and Holger R. Roth. Do gradient inversion attacks make federated learning unsafe? *IEEE Transactions on Medical Imaging*, 42(7):2044–2056, 2023. doi: 10.1109/TMI.2023.3239391.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ayshika Kapoor and Dheeraj Kumar. Federated learning for urban sensing systems: A comprehensive survey on attacks, defences, incentive mechanisms, and applications. *IEEE Communications Surveys & Tutorials*, 27(2):1293–1325, 2025. doi: 10.1109/COMST.2024.3434510.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Rui Li, Hai Wang, Qiang Lu, Jie Yan, Shuo Ji, and Yuhui Ma. Research on medical image classification based on improved fedavg algorithm. *Tsinghua Science and Technology*, 30(5):2243–2258, 2025.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

Can Liu and Jin Wang. Mgic: A multi-label gradient inversion attack based on canny edge detection on federated learning. *arXiv preprint arXiv:2403.08284*, 2024.

Can Liu, Jin Wang, , Yipeng Zhou, Yachao Yuan, Quanzheng Sheng, and Kejie Lu. AFGI: Towards accurate and fast-convergent gradient inversion attack in federated learning, 2024. URL `https://arxiv.org/abs/2403.08383`.

Can Liu, Jin Wang, Yachao Yuan, and Haozhe Wang. Tlg: Two-stage layer-wise gradient inversion attack in federated learning. In *2026 56th Annual IEEE International Conference on Dependable Systems and Networks (DSN)*, pp. 641–652. IEEE, 2026.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Pretom Roy Ovi and Aryya Gangopadhyay. A comprehensive study of gradient inversion attacks in federated learning and baseline defense strategies. In *2023 57th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, 2023.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Yu Sun, Gaojian Xiong, Xianxun Yao, Kailang Ma, and Jian Cui. Gi-pip: Do we require impractical auxiliary dataset for gradient inversion attacks? In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4675–4679. IEEE, 2024.

Yanmeng Wang, Qingjiang Shi, and Tsung-Hui Chang. Why batch normalization damage federated learning on non-iid data? *IEEE transactions on neural networks and learning systems*, 36(1): 1692–1706, 2023.

Yijue Wang, Jieren Deng, Dan Guo, Chenghong Wang, Xianrui Meng, Hang Liu, Caiwen Ding, and Sanguthevar Rajasekaran. Sapag: A self-adaptive privacy attack from gradients. *arXiv preprint arXiv:2009.06228*, 2020.

Zipeng Ye, Wenjian Luo, Qi Zhou, and Yubo Tang. High-fidelity gradient inversion in distributed learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19983–19991, 2024.

Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M. Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16337–16346, 2021.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Kai Yue, Richeng Jin, Chau-Wai Wong, Dror Baron, and Huaiyu Dai. Gradient obfuscation gives a false sense of security in federated learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 6381–6398, 2023.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Rui Zhang, Song Guo, Junxiao Wang, Xin Xie, and Dacheng Tao. A survey on gradient inversion: Attacks, defenses and future directions. *arXiv preprint arXiv:2206.07284*, 2022.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

Junyi Zhu and Matthew Blaschko. R-gap: Recursive gradient attack on privacy. *arXiv preprint arXiv:2010.07733*, 2020.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.



(a) Raw (b) HFG (c) TLG (d) ***ABG*** (e) Raw (f) HFG (g) TLG (h) ***ABG***

Figure 7: Reconstruction results on CelebA and WEB images under a batch size of 1. The selected samples include CelebA images with labels 0 and 1 and WEB images with labels 0 and 100.

Table 3: Metrics of reconstructed images in Fig. 7.

Figure		**Metric**				
		MSE ↓	**PSNR** ↑	**LPIPS** ↓	**SSIM** ↑	**RT** ↓
HFG	(b)	0.02	16.40	0.18	0.45	888
		0.04	14.53	0.26	0.33	889
	(f)	**0.02**	16.48	0.22	0.50	899
		0.04	14.26	0.31	0.36	894
TLG	(c)	0.02	16.34	0.16	0.56	716
		0.02	16.83	0.22	0.44	719
	(g)	**0.02**	16.85	0.16	0.58	718
		0.03	15.25	0.30	0.41	713
ABG	(d)	**0.01**	**19.02**	**0.12**	**0.69**	**641**
		0.02	**17.08**	**0.18**	**0.49**	**640**
	(h)	**0.02**	**17.65**	**0.14**	**0.64**	**631**
		0.01	**18.34**	**0.20**	**0.58**	**637**

A CODE OF *ABG*

Our code is available in `https://anonymous.4open.science/r/ABG-7779`.

B PSEUDOCODE OF *ABG*

Algorithm 1 presents the complete optimization procedure of ***ABG***. Given the shared gradient ∇W, the algorithm first computes the mean $\nabla \bar{W}$ of the non-negative entries in ∇W. This gradient statistic serves two purposes in ***ABG***: adaptive image initialization and adaptive selection of the residual-layer weighting scheme. Specifically, $\nabla \bar{W}$ is scaled to the image pixel range and used to initialize the low-resolution reconstruction $\mathbf{x}^*$. Compared with a fixed or random initialization, this design allows the initial reconstruction to depend directly on the observed gradient characteristics of each target image.

The gradient mean $\nabla \bar{W}$ is further used to select the residual-layer weighting scheme. When $\nabla \bar{W} < \beta$, the residual layers in Block4 are selected and assigned the fixed weight w_r. Otherwise, the residual layers in both Block4 and Block3 are assigned $w_r = 0.1$. The weights of the remaining unselected residual layers and convolutional layers are kept as 1. Therefore, the value of w_r itself is fixed throughout reconstruction, whereas the set of residual layers to which w_r is applied is adaptively determined by $\nabla \bar{W}$. This distinction allows ***ABG*** to adjust the gradient contributions of different residual blocks according to image-dependent gradient characteristics without introducing additional learnable weighting parameters.

C RECONSTRUCTION WITH BATCH SIZE = 1

We further evaluate ***ABG*** on CelebA and the WEB dataset introduced by HFG. The results are reported in Fig. 7 and Table 3. On CelebA, ***ABG*** preserves facial structure, color, and local details more accurately than HFG and TLG. On the WEB dataset, HFG and TLG exhibit visible color

Algorithm 1 Adaptive Block-based Gradient Inversion Attack (***ABG***)

Input: shared gradient ∇W, gradient-mean threshold β, iteration boundaries T_1, T_2, T_3, T, residual-layer weight w_r, residual-layer $\mathrm{Res}(\cdot)$, mean function $\mathrm{mean}(\cdot)$, bicubic interpolation function $\mathrm{Bicubic}(\cdot)$
Output: reconstructed image $\mathbf{x}'$
$\bar{\nabla W} \leftarrow \mathrm{mean}(v \in \nabla W \mid v \geq 0)$
$\mathbf{x}^* \leftarrow 100 * \bar{\nabla W} \cdot \mathbf{1}_{121\times121\times3}$ {Adaptive initialization}
if $\bar{\nabla W} < \beta$ **then**
  $\mathcal{R} \leftarrow \mathrm{Res}(B_4)$
else
  $\mathcal{R} \leftarrow \mathrm{Res}(B_4) \cup \mathrm{Res}(B_3)$
end if
{$\mathcal{R}$ denotes residual layers assigned weight w_r}
for $t = 1$ to T **do**
  $\mathbf{x}' \leftarrow \mathrm{Bicubic}(\mathbf{x}^*, 224 \times 224)$
  $\nabla W' \leftarrow \nabla_\theta \mathcal{L}(f_\theta(\mathbf{x}'), \mathbf{y}')$
  if $1 \leq t \leq T_1$ **then**
    $\mathcal{S}_t \leftarrow B_4$
  else if $T_1 < t \leq T_2$ **then**
    $\mathcal{S}_t \leftarrow B_4 \cup B_3$
  else if $T_2 < t \leq T_3$ **then**
    $\mathcal{S}_t \leftarrow B_4 \cup B_3 \cup LB_2$
  else
    $\mathcal{S}_t \leftarrow B_4 \cup B_3 \cup B_2$
  end if
  $\mathcal{L}_{match} \leftarrow 0$
  for each layer $l \in \mathcal{S}_t$ **do**
    if $l \in \mathcal{R}$ **then**
      $\omega_l \leftarrow w_r$
    else
      $\omega_l \leftarrow 1$
    end if
    $\mathcal{L}_{match} \leftarrow \mathcal{L}_{match} + \omega_l \left(1 - \frac{\langle \nabla W_l', \nabla W_l \rangle}{\|\nabla W_l'\| \|\nabla W_l\|}\right)$
  end for
  if $t \leq T_3$ **then**
    $\mathbf{R_{aux}} \leftarrow \lambda_G \mathbf{R_G}$
  else
    $\mathbf{R_{aux}} \leftarrow \lambda_G \mathbf{R_G} + \lambda_{TV} \mathbf{R_{TV}}$
  end if
  $\mathcal{J}(\mathbf{x}') \leftarrow \lambda_l \mathcal{L}_{match} + \mathbf{R}_{aux}$
  Update $\mathbf{x}^* \leftarrow \arg\min_{\mathbf{x}'} \mathcal{J}(\mathbf{x}')$ {Optimize in low-resolution space}
end for
$\mathbf{x}' \leftarrow \mathrm{Bicubic}(\mathbf{x}^*, 224 \times 224)$
return $\mathbf{x}'$

deviations and local structural artifacts, whereas ***ABG*** maintains more consistent foreground and background information. Quantitatively, ***ABG*** achieves the highest PSNR on all four CelebA and WEB samples, with gains of 2.68, 0.25, 0.80, and 3.09 dB over TLG. It also reduces the average reconstruction time by approximately 11.1% compared with TLG. Overall, the results across ImageNet, CelebA, and WEB demonstrate that ***ABG*** maintains stable reconstruction quality across heterogeneous target images, further validating its adaptive reconstruction capability.

After initialization and weighting-scheme selection, ***ABG*** performs block-wise gradient matching. At each iteration, the current low-resolution reconstruction $\mathbf{x}^*$ is first upsampled by bicubic interpolation to the resolution of the target image, e.g., 224×224, producing $\mathbf{x}'$ for gradient computation and objective evaluation. The reconstructed gradient $\nabla W'$ is then obtained from $\mathbf{x}'$ and $\mathbf{y}'$. Instead of matching all model gradients throughout the entire optimization process, ***ABG*** progressively ex-



(a) Raw (b) Ran (c) $B1$ (d) $w/o\,c$ (e) ***ABG***

Figure 8: Ablation results of ***ABG***. Ran denotes random initialization, B1 denotes additional Block1 gradient matching, and $w/o\,c$ denotes image without encoding and decoding.

pands the set of matched layers according to the iteration boundaries T_1, T_2, and T_3. Specifically, Step I matches gradients from Block4 during the first T_1 iterations. Step II extends the matching range to Block4 and Block3 until T_2. Step III further incorporates the latter half of Block2 until T_3. In Step IV, the matching range is extended to all gradients in Block2 and maintained until the final iteration T. This coarse-to-fine matching procedure first exploits gradients from deeper blocks and progressively incorporates additional layers as optimization proceeds, avoiding the computational overhead of matching all available gradients at every iteration.

For each matched layer, its contribution to the gradient matching loss is determined by the residual-layer weighting scheme selected using $\nabla\bar{W}$. The selected residual layers are weighted by w_r, whereas the remaining matched layers retain a weight of 1. The resulting weighted gradient matching loss is combined with the stage-specific regularization terms defined in the main text. In the early optimization stage, $\mathbf{R}_G$ is used to facilitate structural recovery. After T_3, $\mathbf{R}_{TV}$ is additionally introduced to refine local spatial consistency and image details. Although gradient computation and objective evaluation are performed using the upsampled image $\mathbf{x}'$, parameter updates are applied only to the low-resolution representation $\mathbf{x}^*$. This encoding–decoding procedure reduces the pixel-level optimization space and the number of variables updated at each iteration. Consequently, ***ABG*** combines gradient-mean-guided adaptation, multi-step block-wise matching, and low-resolution optimization within a unified reconstruction procedure. After T iterations, the optimized $\mathbf{x}^*$ is upsampled using bicubic interpolation to obtain the final reconstructed image $\mathbf{x}'$.

D Additional Ablation Results

Figure 8 and Table 4 provide additional ablation results for ***ABG***. Ran replaces the gradient-mean-based initialization with random initialization, B1 additionally includes Block1 in gradient matching, and coding evaluates the image encoding and decoding mechanism.

Compared with the complete ***ABG***, random initialization produces less stable reconstructions with degraded structural consistency, particularly for the second and third target images, where the recovered object structures are less complete. The B1 variant preserves the first two targets relatively well, but the third reconstruction exhibits substantial structural degradation, indicating that additionally matching Block1 gradients is not consistently beneficial across target images. The $w/o\,c$ variant recovers the primary objects more reliably than Ran and B1, but its reconstructions still exhibit reduced structural fidelity and local detail compared with the complete method. These observations support the use of gradient-mean-based initialization, the exclusion of Block1 from gradient matching, and the encoding-decoding mechanism.

In contrast, the complete ***ABG*** combines gradient-mean-guided adaptive initialization, block-wise gradient matching, image encoding and decoding, and adaptive selection of the residual-layer

Table 4: Metrics of reconstructed images in Fig. 8.

Figure	Metric			
	MSE ↓	PSNR ↑	LPIPS ↓	SSIM ↑
(b) Ran	0.04	13.63	0.63	0.17
	0.07	11.83	0.58	0.13
	0.07	11.48	0.70	0.12
(c) $B1$	**0.01**	21.08	0.18	0.44
	0.03	16.23	0.27	0.34
	0.05	12.64	0.63	0.18
(d) $w/o\,c$	0.02	18.06	0.21	0.37
	0.02	16.53	0.21	0.40
	0.02	16.51	**0.31**	0.36
(e) ***ABG***	**0.01**	**21.46**	**0.17**	**0.46**
	0.02	**17.87**	**0.22**	**0.44**
	0.01	**19.50**	**0.31**	**0.41**

weighting scheme according to $\nabla\bar{W}$. It achieves the highest PSNR values of 21.46, 17.87, and 19.50 for the three target images, together with the best overall MSE, LPIPS, and SSIM results. These results confirm that the complete design of ***ABG*** provides the most stable and accurate reconstruction among the evaluated variants.

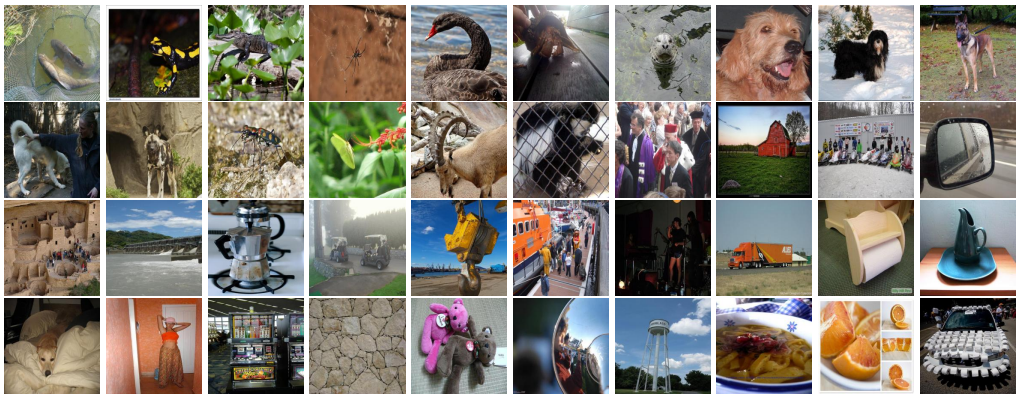
E THE VISUAL RESULTS OF THE ADAPTIVE RESIDUAL LAYER WEIGHTING

As shown in Fig. 9, we reconstruct the first ImageNet image sampled at intervals of 25 classes to evaluate different residual-layer weighting schemes across diverse target images. The two fixed schemes, R4 and R4+R3, exhibit image-dependent and complementary reconstruction performance. Applying the reduced weight only to the residual layers in Block4 (R4) better preserves object structure, color, or background information for some targets, whereas applying the reduced weight to the residual layers in both Block4 and Block3 (R4+R3) performs better for others.

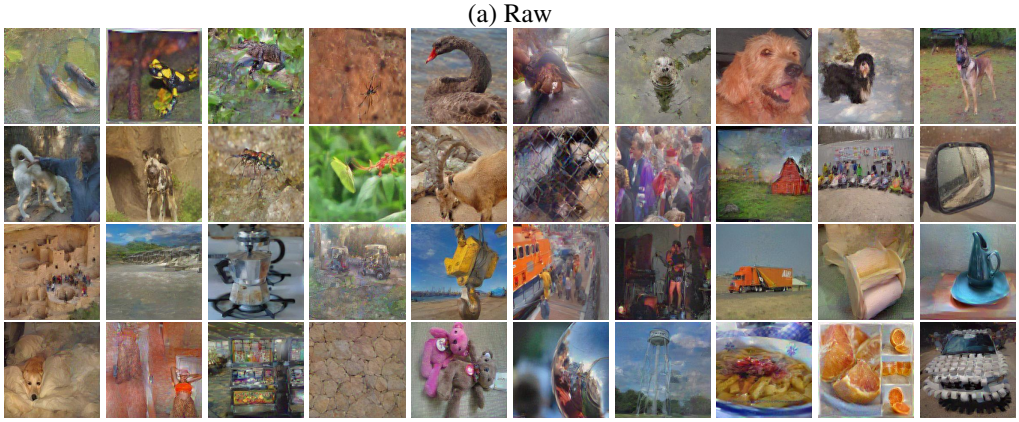
For example, in the third image of the first row, the R4 reconstruction exhibits a noticeable spatial displacement of the primary object, whereas R4+R3 produces a more accurate reconstruction. In contrast, in the second image of the second row, R4+R3 introduces duplicated foreground structures, while R4 better preserves both the object structure and its spatial location. Similar differences are observed in the remaining rows: R4 better preserves background information and spatial layout for some targets, whereas R4+R3 reduces structural distortion or ghosting artifacts for others. These sample-level differences indicate that extending the reduced residual-layer weight from Block4 to Block3 does not consistently improve reconstruction across all target images. Therefore, neither R4 nor R4+R3 alone provides a uniformly optimal weighting configuration.

To address this variation, ***ABG*** uses the gradient mean $\nabla\bar{W}$ as the criterion for adaptively selecting whether the reduced residual-layer weight is applied to Block4 only or to both Block4 and Block3. This design avoids using a single residual-layer weighting configuration for all reconstructions and allows the gradient matching process to adapt to image-dependent differences.

Across the evaluated samples, the complete ***ABG*** consistently selects the more appropriate weighting configuration and achieves the best overall reconstruction performance. Compared with either fixed scheme, the resulting reconstructions more consistently preserve object structures, spatial locations, background information, and color distributions, while reducing cases of ghosting, structural degradation, and color deviation. These results provide additional evidence that gradient-mean-guided adaptive selection of the residual-layer weighting scheme is important for maintaining stable reconstruction quality across different target images.



(a) Raw



(b) R4



(c) R4+R3

Figure 9: Reconstructed images of ***ABG***. The figures are selected the first image from ImageNet at intervals of 25 classes.