

# What Your Features Reveal: Data-Efficient Black-Box Feature Inversion Attack for Split DNNs

## Supplementary Material

This supplementary material contains the following:

1. A theoretical analysis showing that the model can achieve high-fidelity reconstruction with a single inference step.
2. A variational interpretation of the two-stage objective, showing that it can be viewed as approximately maximizing a conditional ELBO under a frozen VAE decoder.
3. Implementation details for our proposed large vision–language model (LVLN)-based metrics and Object Reconstruction Rate (ORR) metrics.
4. Comprehensive experiments evaluating performance, generalization, robustness to noise, and data efficiency, with comparisons to prior work.
5. Ablation studies and feature visualizations that validate key architectural components and provide an intuitive understanding of the model’s workflow.

### 5. Theoretical Foundation of DIFM

**Theorem 1 (One-Step Inference Error).** *When LFSAM produces aligned features  $z_s$  that are sufficiently close to the target latent features  $z_x$ , the one-step Euler method  $\hat{z}_x = z_s + v_\theta(z_s, 0)$  achieves high-fidelity reconstruction.*

We consider the rectilinear path  $z_t = (1 - t) \cdot z_s + t \cdot z_x$  for  $t \in [0, 1]$ . The true vector field  $u_t$  that transports a point from  $z_s$  to  $z_x$  along this path is constant:

$$u_t = \frac{dz_t}{dt} = z_x - z_s \quad (11)$$

Our model  $v_\theta(z_t, t)$  is trained to approximate this field. The one-step Euler approximation starting from  $z_s$  (at  $t = 0$ ) with a step size of  $\Delta t = 1$  is:

$$\hat{z}_x = z_s + 1 \cdot v_\theta(z_s, 0) \quad (12)$$

The global reconstruction error  $e$  is the difference between this approximation and the true target  $z_x$ :

$$e = \|\hat{z}_x - z_x\|_2 \quad (13)$$

$$= \|(z_s + v_\theta(z_s, 0)) - z_x\|_2 \quad (14)$$

$$= \|v_\theta(z_s, 0) - (z_x - z_s)\|_2 \quad (15)$$

We define the model’s local approximation error at the starting point as  $\epsilon = \|v_\theta(z_s, 0) - u_0\|_2 = \|v_\theta(z_s, 0) - (z_x - z_s)\|_2$ , so that:  $e = \epsilon$ . This identity directly links the final reconstruction quality to the model’s local accuracy at the start of the integration.

During training, the FM loss  $\mathcal{L}_{\text{fm}}$  minimizes the expected approximation error over all times  $t$ . A well-converged model will thus have a small error on average. The LFSAM module is designed to produce an initial feature representation  $z_s$  that is already close to the target  $z_x$ . This has a crucial benefit: the magnitude of the target vector field,  $\|u_t\|_2 = \|z_x - z_s\|_2$ , becomes very small. This simplification enables  $v_\theta$  to achieve a very low approximation error  $\epsilon$ , resulting in a high-fidelity one-step reconstruction.

### 6. Variational Interpretation of the Two-Stage Objective

To further justify our two-stage objective, we provide a variational interpretation of FIA-Flow under a frozen VAE decoder. Specifically, we view FIA-Flow as approximately maximizing a conditional evidence lower bound (ELBO) of  $\log p(x | f)$ .

Given an intermediate feature  $f$  and latent variable  $z$ , we consider

$$p(x | f) = \int p_\theta(x | z) p_\psi(z | f) dz, \quad (16)$$

where  $p_\theta(x | z)$  is induced by the frozen decoder Dec and  $p_\psi(z | f)$  is the conditional latent distribution induced by LFSAM and DIFM. Introducing an auxiliary posterior  $q(z | x)$ , the standard conditional ELBO becomes

$$\log p(x | f) \geq \mathbb{E}_{q(z|x)} [\log p_\theta(x | z)] - \text{KL}(q(z | x) \| p_\psi(z | f)). \quad (17)$$

During training, the frozen encoder produces  $z_x = \text{Enc}(x)$ , which we treat as the mean of a simple posterior  $q(z | x) = \mathcal{N}(z_x, \sigma^2 I)$ . In the first stage, LFSAM outputs  $z_s = G_{\text{align}}(f)$ , and we approximate the conditional prior as  $p_\psi(z | f) \approx \mathcal{N}(z_s, \sigma^2 I)$ . Under these Gaussian assumptions, the KL term becomes

$$\text{KL}(q(z | x) \| p_\psi(z | f)) = \frac{1}{2\sigma^2} \|z_x - z_s\|_2^2, \quad (18)$$

which is exactly proportional to the latent alignment loss in Eq. (5) under the above simplifications. Therefore, minimizing Eq. (5) can be interpreted as approximately reducing the KL divergence between the encoder-induced posterior  $q(z | x)$  and the feature-induced conditional prior  $p_\psi(z | f)$ .

For the reconstruction term in (17), we assume a simple pixel-space likelihood paired with Dec. For example, a

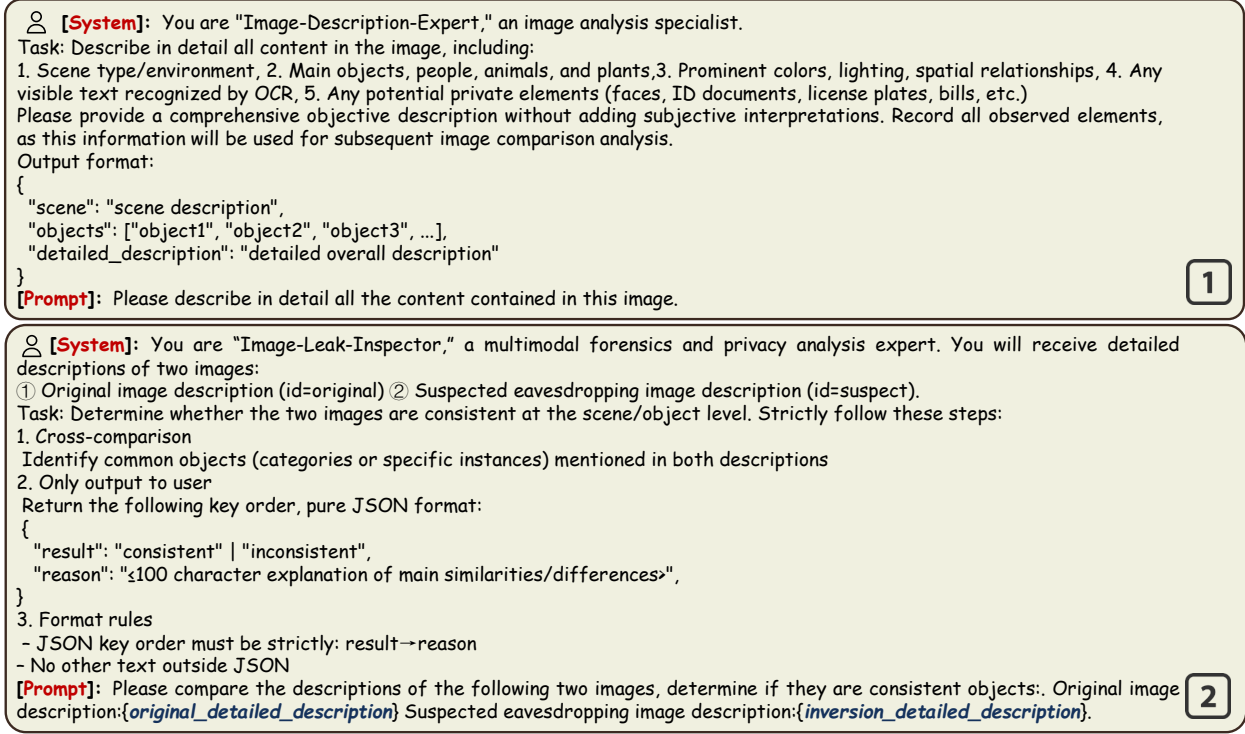


Figure 7. System and User Prompts for the “Image-Description-Expert” and “Image-Leak-Inspector” modules. These modules form the basis of our LVLM-PL and LVLM-C metrics, respectively.

Laplace noise model leads to

$$\log p_{\theta}(x | z) \propto -\|\text{Dec}(z) - x\|_1, \quad (19)$$

so maximizing  $\mathbb{E}_{q(z|x)}[\log p_{\theta}(x | z)]$  is equal to minimizing the  $\ell_1$  reconstruction loss in Eq. (6). In the second stage, the additional LPIPS term in Eq. (9) implies that the likelihood  $p_{\theta}(x | z)$  is modeled not only in the pixel-space but also in the feature-space, which improves perceptual fidelity by encouraging  $x'$  to be close to  $x$  both at the pixel level and in a learned feature representation.

DIFM further refines the conditional prior  $p_{\psi}(z | f)$  in latent space. Starting from  $z_s$ , the learned deterministic flow transports the distribution of  $z_s$  toward the distribution of  $z_x$ , and it moves the probability mass of  $p_{\psi}(z | f)$  towards  $q(z | x)$ , effectively mapping the initial simple distribution to a complex posterior approximation centered at  $\hat{z}_x$ . In the ELBO (17), this can be seen as further decreasing the KL term, while the image-space and perceptual losses in Eq. (9) improve the reconstruction term.

Overall, the two-stage objective (Eqs. (5)–(10)) can be interpreted as an implementable surrogate of the conditional ELBO in (17): LFSAM and DIFM gradually align  $p_{\psi}(z | f)$  with the encoder-induced  $q(z | x)$  in KL, and the reconstruction and perceptual losses approximate the negative log-likelihood under the frozen decoder. Consequently,

FIA-Flow produces reconstructions that are not arbitrary inversions of  $f$ , but approximate high-ELBO explanations of the joint observation  $(x, f)$  under the pretrained VAE.

## 7. Implementation Details for LVLM-based Evaluation Metrics

To quantify privacy leakage in the reconstructed images, we introduce two LVLM-based metrics: LVLM-Consistency (LVLM-C) and LVLM-Privacy-Leakage (LVLM-PL). The process starts with the model generating descriptions for the original image  $x$  and the inversion image  $x'$ . The description of the original image is defined as  $d_o = D(x)$ , while the description for the inversion image is  $d_i = D(x')$ . We set the *temperature* = 0.0 and a fixed random seed of 42 to reduce the randomness of LVLM.

The **LVLM-C** metric determines if the core semantic information has been leaked. The Image Leak Inspector module processes both descriptions,  $C = I(d_o, d_i)$ , to judge if the primary subject is consistent between them. This yields a binary value:

$$\text{LVLM-C} = \begin{cases} 1 & \text{if } C \text{ is consistent} \\ 0 & \text{if } C \text{ is inconsistent} \end{cases} \quad (20)$$

To measure the similarity of finer details, we use **LVLM-PL** metric, defined as the `BERTScore` [59] between the

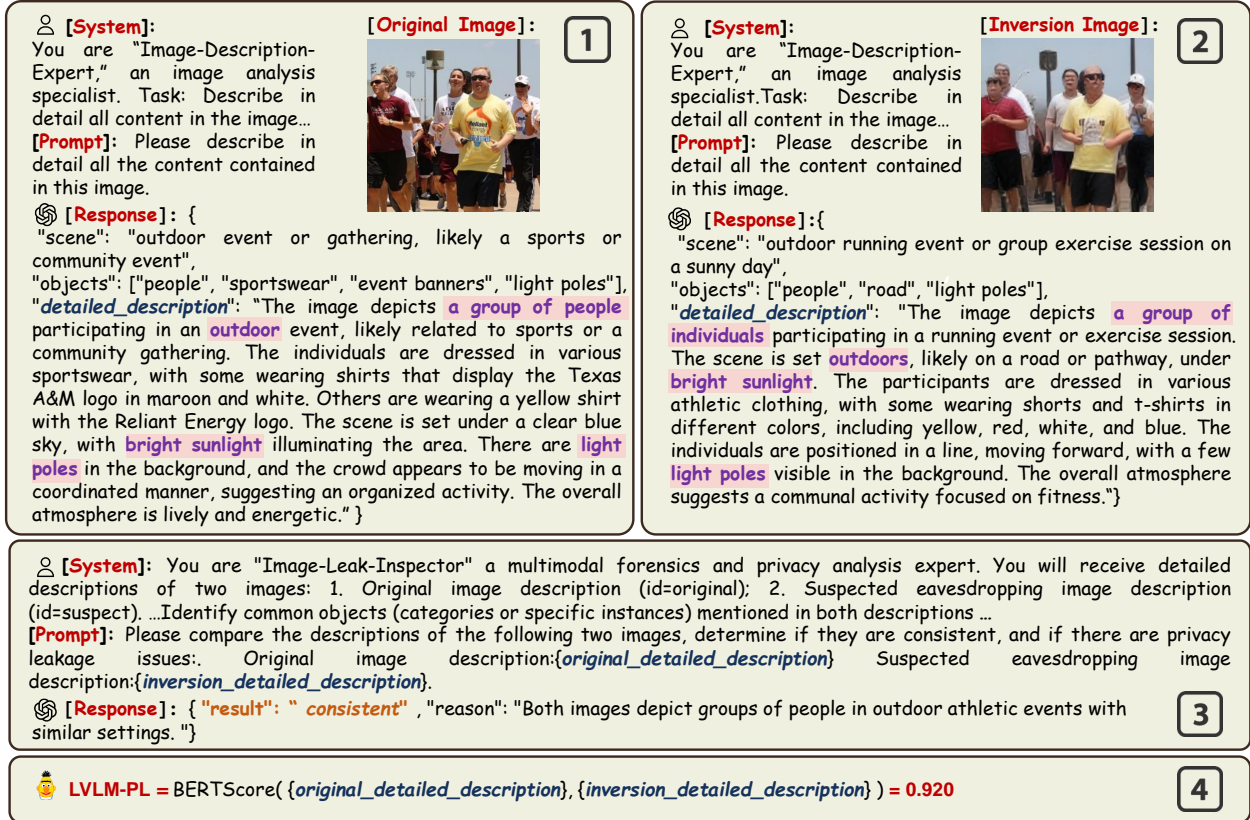


Figure 8. An illustration of LVLML-C and LVLML-PL evaluation. ① The LVLML is prompted to describe the original image. ② The LVLML is then prompted to describe the inversion image. ③ The LVLML compares these two descriptions to ascertain if the same object and scene are consistent. A consistent result yields an LVLML-C of 1. We highlight the same object between the two descriptions. ④ LVLML-PL is obtained by computing the BERTScore [59] between the two descriptions.

two text descriptions, providing a quantitative measure of their semantic resemblance:

$$\text{LVLML-PL} = \text{BERTScore}(d_o, d_i) \quad (21)$$

As shown in Figures 7 and 8, this methodology effectively captures information leakage. For example, the model generated semantically equivalent phrases such as “group of people” and “group of individuals,” and identified identical elements, such as “bright sunlight” and “light poles,” in both the original and the inverted image descriptions. High scores in both LVLML-C and LVLML-PL indicate that an attacker can extract significant private information, thereby validating our proposed FIA method.

Figures 9 and 10 further provide a comprehensive, comparative demonstration of our LVLML-based metrics in evaluating different FIA methods.

In Figure 9, we analyze the inversion results from various methods for an image depicting a group of people at an outdoor sports event. The generated descriptions and the resulting metrics clearly distinguish the quality of the reconstructions. For instance, methods like DIP and SG-

DIP produce vague or distorted images, leading the “Image Leak Inspector” to correctly identify them as “inconsistent” (LVLML-C = 0) and yielding lower LVLML-PL scores. In contrast, the high-fidelity image from our proposed FIA-Flow method prompts a detailed description that closely matches the original, resulting in a “consistent” judgment (LVLML-C = 1) and a high LVLML-PL score. This directly links the metric’s output to the success of the attack. In particular, high LVLML-C and LVLML-PL scores correspond to visually faithful reconstructions that preserve the primary objects and scenes. Furthermore, Figure 10 further highlights the effectiveness of LVLML-C and LVLML-PL metrics. While all inverted images of the ship are globally similar and might pass a simple classification task, they differ in fine-grained details. Our LVLML metric is sensitive to these semantic discrepancies, correctly flagging inaccurate reconstructions as inconsistent. Conversely, it confirms the presence of key elements like ‘dark ship’ and ‘rolling waves’ in high-fidelity reconstructions, which earn higher LVLML-C and LVLML-PL scores.

These examples highlight that our LVLML-based evalua-



Figure 9. Comparison of different FIA methods on LVLM-C and LVLM-PL metrics on the L4-2 layer of ResNet-50.

tion framework offers a robust and intuitive way to quantify the privacy risk posed by different FIA techniques, moving beyond pixel-level metrics to capture the leakage of meaningful semantic information.

## 8. Implementation Details of Object Reconstruction Rate

To quantify privacy leakage at the object level, we introduce the **Object Reconstruction Rate (ORR)**, which measures how many objects in the original image can be recovered in the inverted image, going beyond classical IQA metrics. ORR measures whether an inverted image unintentionally reveals semantic content present in the original image. To this end, we apply an off-the-shelf object detector pretrained on a large-scale dataset to both the original image  $x$  and the inverted image  $x'$ . The detector outputs a set of detections, where each detection is a triplet  $(b, c, s)$  consisting of

a bounding box  $b$ , a class label  $c$ , and a confidence score  $s$ . After filtering detections using a confidence threshold  $\tau_{\text{conf}}$ , we obtain two sets of box-label pairs:

$$B_x = \{(b, c) \mid (b, c, s) \in \text{Det}(x), s > \tau_{\text{conf}}\}, \quad (22)$$

$$B_{x'} = \{(b, c) \mid (b, c, s) \in \text{Det}(x'), s > \tau_{\text{conf}}\}, \quad (23)$$

where  $\text{Det}(\cdot)$  denotes the set of raw detector outputs.

The ORR is defined as the fraction of objects in  $B_x$  that find a matching object in  $B_{x'}$ . A match is considered successful if (i) the class labels are identical ( $c_i = c_j$ ), and (ii) the Intersection over Union (IoU) between the two bounding boxes is at least  $\tau_{\text{IoU}}$ . The ORR can be calculated as:

$$\text{ORR}_{\tau_{\text{IoU}}} = \frac{1}{|B_x|} \sum_{(b_i, c_i) \in B_x} \mathbb{I} \left[ \max_{(b_j, c_j) \in B_{x'}, c_j = c_i} \text{IoU}(b_i, b_j) \geq \tau_{\text{IoU}} \right], \quad (24)$$



Figure 10. Comparison of different FIA methods on LVLML-C and LVLML-PL metrics on the F3-2 layer of Swin-B.

where  $\mathbb{I}[\cdot]$  is the indicator function, which returns 1 if the condition is true and 0 otherwise. If no objects are detected in the original image (i.e.,  $|B_x| = 0$ ), we exclude this image from ORR computation.

## 9. Related Work

### 9.1. Split DNN for Efficient Inference

To address the significant computational demands of DNN on edge devices, the Split DNN approach was introduced, which partitions a DNN between an edge device and a server to reduce the computational load. In this paradigm, the edge device needs to send intermediate features to the server for final processing. Neurosurgeon [15] automatically finds the optimal split point for different devices and models. BranchyNet [48] integrates early exit points at various depths of a DNN. This allows the network to adaptively terminate the inference process for simpler inputs that can be classified with high confidence in earlier stages, thus

reducing the average computational cost. More recently, ASAP [35] demonstrated that an asynchronous cloud edge inference model for split DNNs can lead to higher throughput in many modern architectures.

### 9.2. Inversion Attacks in Split DNN

Although Split DNN can reduce device computation, the transmission of intermediate features creates a severe privacy risk. This enables FIA [7, 33, 58], a direct threat where adversaries aim to reconstruct the exact user input from these features. Unlike the classic MIA [14, 31, 55, 60] methods, which typically recover class-representative images from final predictions, FIA targets specific private inputs.

More recently, diffusion models (DMs) have enabled a new generation of FIA with far superior image quality. Despite this, state-of-the-art approaches [58] exhibit a critical flaw: they rely on precise text prompts corresponding to the input image to guide the generation process. This require-

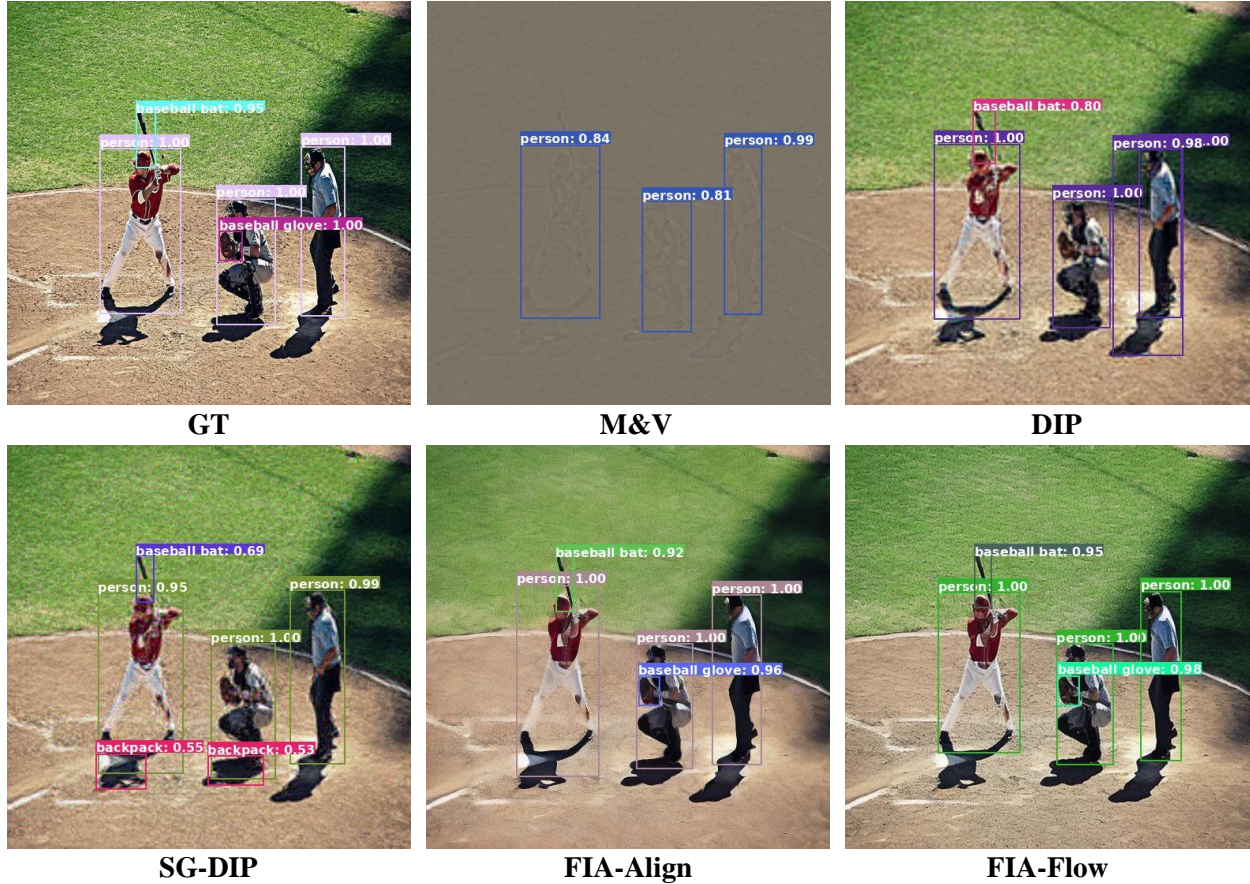


Figure 11. A visualization comparison result of the L1-2 layer of ResNet-50 on the COCO dataset. The bounding box is obtained by the same Faster-RCNN with the ResNet-50 backbone. The bounding boxes from FIA-Flow most closely match the GT.

ment is impractical in most real-world scenarios where such auxiliary information is unavailable. This motivates our work, where we propose a powerful DM-based FIA that operates without the need for these strong, unrealistic priors.

### 9.3. Diffusion Models

As discussed in the previous section, DM has become the state-of-the-art in generative modeling [9]. The concept of DM originates from non-equilibrium thermodynamics [12] and was first introduced to machine learning by [47] to model datasets. Ho et al. proposed the diffusion probabilistic model (DDPM) [9], which uses a simplified training objective and a powerful U-Net architecture [42]. To reduce computational costs, the Latent Diffusion Model (LDM) [41] was introduced to shift the denoising process from the pixel domain to a latent space provided by a pre-trained Variational Autoencoder (VAE) [20]. Recently, FM [28] has offered a more general and powerful perspective, which trains Continuous Normalizing Flows by directly regressing the vector field connecting data and noise, opening new avenues for improving model performance and efficiency.

## 10. Extension Experiments

### 10.1. Evaluation on the COCO Dataset

To evaluate generalization on a more complex and diverse dataset, we conduct experiments on the MS COCO-2017 validation set [27]. Specifically, we use a pre-trained Faster R-CNN [39] detector with a ResNet-50 backbone on both the original image  $x$  and the inverted image  $x'$ . We set  $\tau_{conf} = 0.5$  and report  $ORR_{0.5}/ORR_{0.75}$  with  $\tau_{IoU} = 0.5/0.75$ , respectively. The results are shown in Table 8.

Figure 11 provides a qualitative visualization of the inversion results with bounding boxes from the ResNet-50 L1-2 layer. Our FIA-Flow method yields a reconstruction with high perceptual quality and semantic integrity. While prior methods like DIP and SG-DIP can only recover a coarse layout with noticeable artifacts, our approach reconstructs fine-grained details, such as the players' poses and the texture of the baseball field, with high fidelity. The inverted image is accurate enough that an off-the-shelf detector correctly identifies all ground-truth objects (person, baseball bat, baseball glove) with high confidence, in sharp

Table 8. The performance comparison with different FIA methods on the COCO dataset. Bold indicates the best result of all methods.

Model	Layer	Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	NIQE $\downarrow$	MANIQA $\uparrow$	ORR <sub>0.5</sub> $\uparrow$	ORR <sub>0.75</sub> $\uparrow$
ResNet-50	L1-2	M&V	13.34	0.503	0.700	5.892	0.5191	3.30	2.20
		DIP	23.98	0.650	0.332	5.250	0.4464	44.94	33.40
		SG-DIP	25.66	0.698	0.284	5.004	0.4834	50.41	39.75
		FIA-Align	<b>27.03</b>	0.745	0.195	5.184	0.5981	56.02	45.84
		FIA-Flow	27.00	<b>0.748</b>	<b>0.115</b>	<b>3.699</b>	<b>0.6626</b>	<b>69.00</b>	<b>59.33</b>
	L4-2	M&V	13.26	0.466	0.890	8.972	0.4655	0.03	0.00
		DIP	13.08	0.415	0.802	7.018	0.2182	4.54	1.62
		SG-DIP	11.20	0.280	0.847	5.522	0.2992	0.50	0.23
		FIA-Align	<b>19.40</b>	<b>0.551</b>	0.711	11.192	0.2881	5.73	2.84
		FIA-Flow	19.12	0.526	<b>0.451</b>	<b>4.841</b>	<b>0.5410</b>	<b>23.64</b>	<b>14.87</b>

Table 9. The performance comparison with different FIA methods on the F-4 layer of AlexNet. Bold indicates the best result of all methods.

Model	Layer	Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Acc $\uparrow$	LVLM-C $\uparrow$	LVLM-PL $\uparrow$	NIQE $\downarrow$	MANIQA $\uparrow$
AlexNet	F-4	M&V	13.57	0.505	0.584	0.4	7.1	0.888	6.342	0.4823
		DIP	17.18	0.490	0.664	9.9	15.0	0.880	5.835	0.2826
		SG-DIP	21.43	0.526	0.564	20.7	30.2	0.904	5.693	0.3563
		FIA-Align	25.56	0.701	0.395	24.6	33.6	0.907	8.727	0.3684
		FIA-Flow	<b>25.72</b>	<b>0.706</b>	<b>0.202</b>	<b>63.9</b>	<b>49.6</b>	<b>0.920</b>	<b>5.251</b>	<b>0.5731</b>

contrast to the baselines. This is in stark contrast to competing methods; for example, M&V fails to produce any recognizable structure, and SG-DIP hallucinates non-existent objects like a “backpack”. This visual evidence corroborates our quantitative findings, highlighting that the images inverted by FIA-Flow retain sufficient semantic information to be effectively used by other downstream models, thus exposing a significant privacy risk.

Our FIA-Flow method consistently achieves state-of-the-art performance. FIA-Flow, trained only on ImageNet, generalizes strongly to the COCO validation set without any per-sample optimization on COCO features. It achieves the best or near-best scores across LPIPS, NIQE, MANIQA, and ORR, indicating both high perceptual quality and strong semantic fidelity. For the ResNet-50 L1-2 layer, it produces reconstructions of high perceptual quality, evidenced by the best scores in LPIPS, NIQE, and MANIQA. Meanwhile, FIA-Flow achieves an ORR<sub>0.5</sub> of 69.00% and ORR<sub>0.75</sub> of 59.33%. This significantly exceeds all other methods, proving that the inversion images retain not just visual appeal but also critical structural and semantic information for downstream machine analysis.

The advantage of our alignment-refinement strategy becomes even more pronounced when attacking deeper, more abstract features. On the challenging L4-2 layer, where most methods fail to recover meaningful information, FIA-Flow maintains robust performance. It achieves an ORR<sub>0.5</sub> of 23.64%, whereas FIA-Align’s performance plummets to 5.73% and other methods become entirely ineffective. These results confirm that FIA-Flow generalizes exceptionally well. More importantly, they reveal a severe privacy

threat: the inverted images are of such high quality that they can be successfully used by other sophisticated models, leading to a level of information leakage far greater than what conventional IQA metrics can capture.

## 10.2. Additional Results on Diverse Victim Layer

To further investigate the privacy leakage risk across different layers, we have included more FIA comparison results. As shown in Table 9, we present the FIA results for the *features.4* (F-4) layer of AlexNet [21]. Our results demonstrate that FIA-Flow achieves the best performance across all evaluation metrics, particularly in terms of reconstruction quality. Specifically, on the F-4 layer, it achieves the highest PSNR and SSIM, along with the lowest LPIPS. In addition, FIA-Flow achieves an Acc of 63.9% on the classification task, significantly outperforming other methods. These results collectively demonstrate that the F-4 layer of AlexNet encodes rich information from the input image, and the powerful reconstruction capability of FIA-Flow further highlights the serious privacy leakage risk present in these deep features.

## 10.3. Robustness to Noisy Intermediate Features

To simulate real-world wireless transmission errors and deliberate feature perturbation for privacy [11, 43], we investigate the robustness of different FIA methods to noisy intermediate features. Unlike the defense methods discussed in the main paper, which assume the adversary is aware of the deployed defense. This experiment assesses the model’s generalization ability to unknown noise distributions. Specifically, we evaluate models trained on pristine

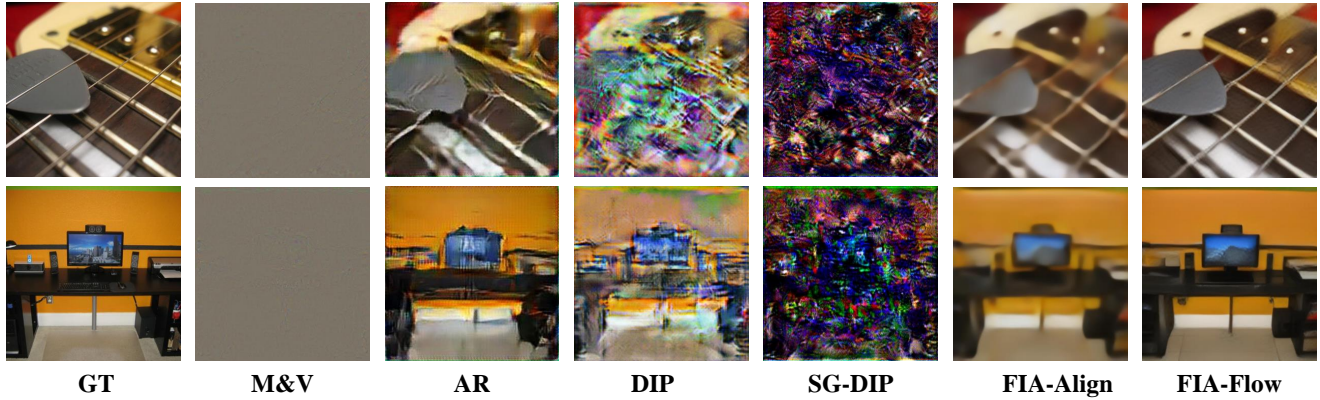


Figure 12. Visualization result for the F-10 layer of AlexNet.



Figure 13. Visualization result for the L1-2 layer of ResNet-50.

features directly on inputs corrupted by Gaussian noise of varying magnitudes, without any fine-tuning.

We evaluate FIA on ResNet-50 intermediate features corrupted with Gaussian noise of variance 0.1 and 0.3 (see Table 10). Across different layers and noise levels, FIA-Align often achieves higher PSNR and SSIM, whereas FIA-Flow consistently attains better perceptual scores (LPIPS, MANIQA). As the noise variance increases, performance degrades for all methods, but FIA-Flow retains a clear advantage in perceptual quality. For Acc, our FIA-Flow secures the top performance in most cases. Even under the most challenging conditions (L4-2 layer, Var=0.3), it achieves a notable 25.7% accuracy. Importantly, FIA-Flow was never trained on noise, unlike DIP and SG-DIP, which are specifically designed for it. Despite this, its performance significantly outperforms SG-DIP (6.8%), confirming the model’s strong generalization and effectiveness against unseen noise. FIA-Flow not only produces images of higher perceptual quality in noisy conditions but also proves exceptionally effective at preserving the core semantic infor-

mation essential for downstream tasks, highlighting its enhanced robustness for FIA under adverse conditions.

#### 10.4. More Visualization Comparison

To visualize the comparative results of our FIA method on more layers, we present the visualization results for the F-10 of AlexNet and the L1-2 layer of ResNet-50 in Figures 12 and 13. As shown, FIA-Flow consistently produces high-fidelity and semantically coherent visualizations across different layers. Compared to M&V and AR, it achieves higher reconstruction quality while avoiding the noisy artifacts often seen in optimization-based methods like DIP and SG-DIP. Moreover, while FIA-Align performs well, FIA-Flow is more effective at preserving spatial and structural integrity, yielding more natural and less distorted outputs. This demonstrates FIA-Flow’s enhanced capability to reconstruct the input from deep features accurately.

Table 10. The performance comparison of FIA against Gaussian noise on intermediate features from ResNet-50. Var denotes the variance of the added Gaussian Noise. Bold indicates the best result of all methods.

Model	Layer	Var	Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Acc $\uparrow$	NIQE $\downarrow$	MANIQA $\uparrow$
ResNet-50	L1-2	0.1	DIP	25.60	0.700	0.247	59.6	5.412	0.450
			SG-DIP	26.99	0.716	0.287	58.3	4.982	0.4403
			FIA-Align	<b>29.16</b>	<b>0.798</b>	0.166	63.3	5.034	0.5468
			FIA-Flow	29.15	0.797	<b>0.117</b>	<b>70.6</b>	<b>4.458</b>	<b>0.6004</b>
	L1-2	0.3	DIP	24.63	0.671	0.311	53.6	5.341	0.3975
			SG-DIP	24.04	0.595	0.568	34.4	5.779	0.3448
			FIA-Align	<b>26.16</b>	<b>0.736</b>	<b>0.256</b>	49.8	<b>4.668</b>	0.4711
			FIA-Flow	25.92	0.720	0.281	<b>59.3</b>	5.051	<b>0.5189</b>
	L4-2	0.1	DIP	13.96	0.461	0.682	33.3	7.301	0.2757
			SG-DIP	11.69	0.310	0.778	7.1	<b>5.647</b>	0.3166
			FIA-Align	<b>20.27</b>	<b>0.602</b>	0.645	4.3	11.314	0.2931
			FIA-Flow	19.97	0.573	<b>0.405</b>	<b>35.6</b>	5.670	<b>0.5047</b>
L4-2	0.3	DIP	13.87	0.460	0.691	<b>32.9</b>	7.325	0.2731	
		SG-DIP	11.44	0.302	0.793	6.8	5.629	0.3091	
		FIA-Align	<b>19.58</b>	<b>0.592</b>	0.657	3.3	11.261	0.2739	
		FIA-Flow	18.97	0.547	<b>0.449</b>	25.7	<b>5.860</b>	<b>0.4960</b>	

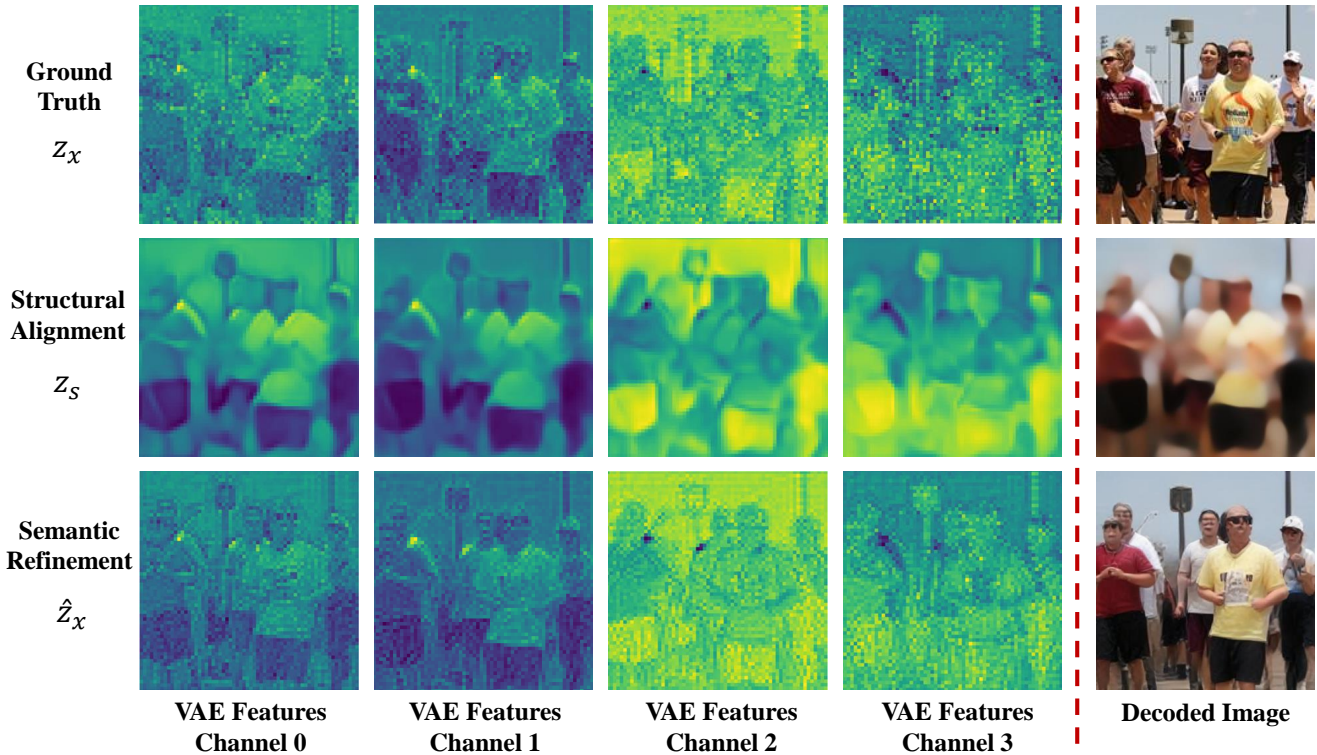


Figure 14. Visualization of the alignment-refinement process in FIA-Flow. Top row: VAE features for the ground truth  $z_x$ . Middle row: features after structural alignment by LFSAM  $z_s$ . Bottom row: features after semantic refinement by DIFM  $\hat{z}_x$ .

### 10.5. Internal Feature Visualization of FIA-Flow

To intuitively understand the workflow of the "alignment-refinement" strategy in FIA-Flow, we visualize its intermediate output results in Figure 14. The first row displays the VAE features corresponding to the ground truth image.

The second row shows the features after structural alignment via LFSAM. It can be seen that the structurally aligned features  $z_s$  exhibit a clearer and more defined spatial layout compared to the original VAE features  $z_x$ . While still somewhat coarse, the boundaries of the individuals and the overall scene structure become more discernible after the

Table 11. The performance comparison of FIA-Flow with different upsample methods on L4-2 of ResNet-50.

Method	PS	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Acc $\uparrow$	LVLM-C $\uparrow$	LVLM-PL $\uparrow$	NIQE $\downarrow$	MANIQA $\uparrow$
FIA-Align	$\times$	20.19	0.601	0.652	4.0	5.9	0.877	11.471	0.2972
FIA-Align	$\checkmark$	<b>20.36</b>	<b>0.603</b>	0.643	4.4	6.3	0.878	11.309	0.2969
FIA-Flow	$\times$	20.31	0.588	0.425	35.5	13.8	0.899	5.275	0.5433
FIA-Flow	$\checkmark$	20.31	0.584	<b>0.397</b>	<b>36.8</b>	<b>18.0</b>	<b>0.902</b>	<b>5.098</b>	<b>0.5628</b>

Table 12. The performance comparison with different LVLM base models for LVLM-C and LVLM-PL. Bold indicates the best result of all methods.

Model	Layer	Method	GPT-4o-mini		Qwen2-VL-72B-Instruct	
			LVLM-C	LVLM-PL	LVLM-C	LVLM-PL
AlexNet	F-10	M&V	1.2	0.860	2.4	0.859
		DIP	10.6	0.880	35.1	0.885
		AR	4.8	0.880	24.2	0.881
		SG-DIP	3.6	0.865	8.1	0.866
		FIA-Align	9.3	0.883	43.4	0.890
		FIA-Flow	<b>16.6</b>	<b>0.900</b>	<b>63.0</b>	<b>0.905</b>
ResNet-50	L1-2	M&V	17.5	0.903	48.0	0.897
		DIP	39.9	0.905	97.3	0.930
		SG-DIP	65.3	0.922	94.0	0.912
		FIA-Align	70.0	0.923	97.5	0.932
		FIA-Flow	<b>70.1</b>	<b>0.929</b>	<b>97.5</b>	<b>0.936</b>
		ResNet-50	L4-2	M&V	3.0	0.860
DIP	9.4			0.881	45.7	0.889
SG-DIP	5.0			0.872	16.8	0.872
FIA-Align	6.3			0.878	35.0	0.886
FIA-Flow	<b>18.0</b>			<b>0.902</b>	<b>66.2</b>	<b>0.906</b>
Swin-B	F3-2			M&V	38.4	0.913
		DIP	54.5	0.920	97.8	0.928
		SG-DIP	62.3	0.913	94.2	0.912
		FIA-Align	51.5	0.919	92.9	0.923
		FIA-Flow	<b>63.2</b>	<b>0.925</b>	<b>98.2</b>	<b>0.932</b>
		YOLO11n	M-8	M&V	1.2	0.863
DIP	18.2			0.897	45.1	0.902
SG-DIP	18.1			<b>0.899</b>	49.7	0.891
FIA-Align	7.1			0.880	34.1	0.894
FIA-Flow	<b>23.9</b>			<b>0.899</b>	<b>53.0</b>	<b>0.910</b>
DINOv2-B	B-11			M&V	0.7	0.855
		DIP	4.9	0.868	10.4	0.873
		SG-DIP	28.3	0.905	58.3	0.921
		FIA-Align	16.7	0.890	47.9	0.901
		FIA-Flow	<b>30.4</b>	<b>0.909</b>	<b>76.2</b>	<b>0.925</b>

alignment process. This indicates that LFSAM effectively transfers the structural information from the reference features, leading to a better-organized representation. The third row presents the features after semantic refinement using DIFM. It can be seen that the semantically refined features  $\hat{z}_x$  demonstrate significantly improved semantic coherence and visual fidelity compared to both the original and the structurally aligned features. The individual figures are now much more recognizable, with richer details and more accurate object boundaries. This suggests that DIFM effectively injects semantic information, resulting in a representation

that is both structurally sound and semantically meaningful. This demonstrates that our proposed alignment-refinement strategy effectively disentangles structural and semantic information and progressively enhances the feature representation, leading to the high-quality visualizations achieved by FIA-Flow.

Table 13. The performance comparison of FIA-Flow with different numbers of training samples on L4-2 of ResNet-50.

Number	Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Acc $\uparrow$	LVLM-C $\uparrow$	LVLM-PL $\uparrow$	NIQE $\downarrow$	MANIQA $\uparrow$
4096(0.32%)	FIA-Align	20.36	0.603	0.643	4.4	6.3	0.878	11.309	0.2969
	FIA-Flow	20.31	0.584	0.397	36.8	18.0	0.902	5.098	0.5628
1024(0.08%)	FIA-Align	19.94	0.600	0.651	4.4	5.6	0.877	11.243	0.2819
	FIA-Flow	20.04	0.583	0.432	27.7	14.5	0.898	5.362	0.5326
256(0.02%)	FIA-Align	19.66	0.596	0.650	4.7	6.2	0.876	11.138	0.2894
	FIA-Flow	19.45	0.566	0.411	31.1	12.8	0.900	4.768	0.5653
128(0.01%)	FIA-Align	19.32	0.592	0.661	4.1	4.3	0.875	11.204	0.2929
	FIA-Flow	19.01	0.554	0.442	27.7	12.5	0.898	5.234	0.5321

Table 14. The performance comparison of FIA-Flow with different victim layers of ResNet-50.

Layer	Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Acc $\uparrow$	LVLM-C $\uparrow$	LVLM-PL $\uparrow$	NIQE $\downarrow$	MANIQA $\uparrow$
L1-2	FIA-Align	29.86	0.810	0.157	64.3	70.0	0.923	5.136	0.5622
	FIA-Flow	30.01	0.8141	0.1002	71.3	70.1	0.929	4.408	0.6131
L2-2	FIA-Align	29.59	0.8083	0.1731	63.8	61.4	0.924	5.158	0.5406
	FIA-Flow	29.65	0.8098	0.1053	71.0	69.8	0.928	4.414	0.6056
L3-2	FIA-Align	26.36	0.7241	0.3494	32.0	45.0	0.913	8.021	0.4410
	FIA-Flow	26.29	0.7237	0.1607	69.8	63.4	0.913	4.658	0.6032
L4-2	FIA-Align	20.36	0.603	0.643	4.4	6.3	0.878	11.309	0.2969
	FIA-Flow	20.31	0.584	0.397	36.8	18.0	0.902	5.098	0.5628

Table 15. The performance comparison of FIA-Flow with different sample methods and steps on L4-2 of ResNet-50.

Diffusion	Steps	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	Acc $\uparrow$	LVLM-C $\uparrow$	LVLM-PL $\uparrow$	NIQE $\downarrow$	MANIQA $\uparrow$
DDPM	10	20.09	<b>0.599</b>	0.649	4.1	5.8	0.878	11.283	0.2962
	50	19.97	0.596	0.653	4.9	4.2	0.876	11.233	0.2875
	200	19.95	0.596	0.655	4.5	4.8	0.877	11.182	0.2861
FM	1	<b>20.31</b>	0.584	<b>0.397</b>	<b>36.8</b>	<b>18.0</b>	0.902	<b>5.098</b>	<b>0.5628</b>
	5	19.61	0.556	0.404	38.2	37.3	<b>0.914</b>	5.564	0.5052
	10	19.21	0.540	0.419	36.9	38.3	<b>0.914</b>	5.674	0.4968

## 11. Ablation Studies

### 11.1. Results with Different Upsampling Methods

Table 11 analyzes the impact of the PixelShuffle (PS) [44] upsampling in LFSAM. Incorporating PS yields only modest gains for the FIA-Align baseline, but leads to substantially larger improvements for FIA-Flow, especially on perceptual and semantic metrics. Specifically, for FIA-Align, the use of PS leads to a slight increase in PSNR and a minor improvement in perceptual metrics. In contrast, for FIA-Flow, the addition of PS yields substantial benefits, particularly in metrics related to perceptual and semantic quality. The LPIPS improves significantly from 0.425 to 0.397, while the LVLM-C metric achieves a boost from 13.8 to 18.0.

This analysis reveals a critical insight: the primary performance gain is not a result of the upsampling mechanism alone, but rather a direct outcome of our FIA-Flow method-

ology. The FIA-Flow method, even without the PS layer, significantly outperforms the FIA-Align baseline (without PS), boosting the classification accuracy from 4.0% to 35.5% and demonstrating substantial improvements across all other metrics. Therefore, the PS module acts as a synergistic component that provides a final boost, allowing our core methodology to achieve its full potential and deliver state-of-the-art results.

### 11.2. Results with Different LVLM Base Models

To verify the applicability of LVLM-C and LVLM-PL metrics to different LVLM base models and to further evaluate the effectiveness of FIA-Flow, we additionally validate them using Qwen2-VL-72B-Instruct [52] (Qwen2-VL) as the LVLM base model, with results shown in Table 12. It can be seen that our proposed LVLM-C and LVLM-PL metrics are effective with the Qwen2-VL base model. Furthermore, FIA-Flow consistently achieves the best performance

on both LVLM-C and LVLM-PL metrics across all victim models (AlexNet, ResNet-50, Swin-B) and layers, which further validates the effectiveness of our method. Additionally, it is noteworthy that all methods generally achieve significantly higher LVLM-C values when using Qwen2-VL compared to GPT-4o-mini, which may suggest that Qwen2-VL has a stronger capability for recognizing fine-grained image details.

### 11.3. Results with Different Training Numbers

To evaluate the data efficiency of our framework, we conducted ablation studies on the ResNet-50 L4-2 layer, a challenging target due to its highly abstract features. We trained FIA-Flow using subsets of ImageNet-1K containing 4096 (0.32%), 1024 (0.08%), 256 (0.02%), and just 128 (0.01%) samples.

The results in Table 13 underscore FIA-Flow’s data efficiency. Even when trained with only 128 images, our method achieves an Acc of 27.7%. While the PSNR decreases slightly with fewer samples, the semantic metric LVLM-PL remains robust, scoring 0.898 with just 128 samples. These results indicate that the alignment-refinement strategy can be learned from very limited supervision, which makes FIA-Flow a realistic threat even when only a few image-feature pairs are available.

### 11.4. Results with Different Victim Layers Depth

We investigated the robustness of FIA-Flow across different network depths by evaluating its performance on four layers of ResNet-50: L1-2, L2-2, L3-2, and L4-2. The complete performance comparison against the FIA-Align baseline is detailed in Table 14.

As expected, the performance of all methods degrades as layers deepen and features become more abstract. However, FIA-Flow consistently outperforms the baseline across all layers. This significant performance advantage in deep layers is primarily attributed to our DIFM module. On the L4-2 layer, FIA-Flow achieves an accuracy of 36.8%, more than eight times higher than the 4.4% achieved by FIA-Align. This confirms that DIFM effectively refines the structurally-aligned features from LFSAM, correcting distributional mismatches and recovering semantically coherent images even from highly processed features. This capability reveals a significant privacy risk in deep layers previously considered more secure.

### 11.5. Results with Different Sampling Methods and Steps

We compared our DIFM approach with the conventional stochastic sampling used in DDPM. The experiments, conducted on the ResNet-50 L4-2 layer, were designed to highlight the fundamental suitability of our deterministic method for the high-fidelity reconstruction required by FIA.

The detailed results are presented in Table 15. The results clearly favor our deterministic approach. The iterative, stochastic nature of DDPM is ill-suited for reconstructing a specific input with precision. Even with 200 sampling steps, DDPM achieves a mere 4.5% accuracy and an LVLM-C score of 4.8, indicating poor semantic recovery.

In contrast, DIFM achieves high-fidelity inversion with high efficiency. With a single Euler step, FIA-Flow already attains 36.8% accuracy and an LVLM-C of 18.0, outperforming the 200-step DDPM baseline on all metrics. Increasing the number of DIFM steps from 1 to 10 slightly reduces PSNR but further improves semantic metrics (Acc, LVLM-C, LVLM-PL), indicating that the learned vector field is stable and that additional refinement steps can extract more semantic information from the features.

## References

- [1] Nilesh Ahuja, Parual Datta, Bhavya Kanzariya, V Srinivasa Somayazulu, and Omesh Tickoo. Neural rate estimator and unsupervised learning for efficient distributed image analytics in split-DNN models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2022–2030, 2023. 1
- [2] Dake Chen, Shiduo Li, Yuke Zhang, Chenghao Li, Souvik Kundu, and Peter A Beerel. DIA: Diffusion based inverse network attack on collaborative inference. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 124–130, 2024. 2
- [3] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024. 1
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255, 2009. 4
- [5] Ulyanov Dmitry, Andrea Vedaldi, and Lempitsky Victor. Deep image prior. *Int. J. Comput. Vis.*, 128(7):1867–1888, 2020. 2, 5
- [6] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 3
- [7] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4829–4837, 2016. 5
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 5
- [9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.*, pages 6840–6851, 2020. 8, 6
- [10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *Int. Conf. Learn. Represent.*, 2022. 5
- [11] Nguyen Hung-Quang, Yingjie Lao, Tung Pham, Kok-Seng Wong, and Khoa D Doan. Understanding the robustness of

- randomized feature defense against query-based adversarial attacks. In *Int. Conf. Learn. Represent.*, 2024. 7
- [12] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957. 6
- [13] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 5
- [14] Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15045–15053, 2022. 1, 5
- [15] Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *ACM SIGARCH Computer Architecture News*, 45(1):615–629, 2017. 1, 5
- [16] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1
- [17] Jyotirmoy Karjee, Praveen Naik, Kartik Anand, and Vanamala N Bhargav. Split computing: Dnn inference partition with load balancing in iot-edge platform for beyond 5g. *Measurement: Sensors*, 23:100409, 2022. 1
- [18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4401–4410, 2019. 1
- [19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8110–8119, 2020. 1
- [20] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 6
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 5, 7
- [22] Wa-Kin Lei, Jun-Cheng Chen, and Shang-Tse Chen. DRAG: Data reconstruction attack using guided diffusion. *arXiv preprint arXiv:2509.11724*, 2025. 2
- [23] Dawei Li, Xiaolong Wang, and Deguang Kong. Deeprebirth: Accelerating deep neural network execution on mobile devices. In *AAAI*, 2018. 1
- [24] Haoyang Li, Li Bai, Qingqing Ye, Haibo Hu, Yaxin Xiao, Huadi Zheng, and Jianliang Xu. A sample-level evaluation and generative framework for model inversion attacks. In *AAAI*, pages 18287–18295, 2025. 1
- [25] Ziang Li, Hongguang Zhang, Juan Wang, Meihui Chen, Hongxin Hu, Wenzhe Yi, Xiaoyang Xu, Mengda Yang, and Chenjun Ma. From head to tail: Efficient black-box model inversion attack via long-tailed learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 29288–29298, 2025. 1
- [26] Shijun Liang, Evan Bell, Qing Qu, Rongrong Wang, and Saiprasad Ravishankar. Analysis of deep image prior and exploiting self-guidance for image reconstruction. *IEEE Transactions on Computational Imaging*, 2025. 2, 5
- [27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014. 6
- [28] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2, 6
- [29] Hao Liu, Lijun He, Miao Zhang, and Fan Li. Vadiffusion: Compressed domain information guided conditional diffusion for video anomaly detection. *IEEE Trans. Circuit Syst. Video Technol.*, 34(9):8398–8411, 2024. 1
- [30] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2
- [31] Yufan Liu, Wanqian Zhang, Dayan Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. Prediction exposes your face: Black-box model inversion via prediction alignment. In *Eur. Conf. Comput. Vis.*, pages 288–306. Springer, 2024. 1, 5
- [32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, pages 10012–10022, 2021. 5
- [33] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5188–5196, 2015. 2, 5
- [34] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Sign. Process. Letters*, 20(3):209–212, 2012. 5
- [35] Waleed Hassan Mubark, Jagannath Guptha Kasula, and Md Yusuf Sarwar Uddin. Asap: Asynchronous split inference for accelerated dnn execution. In *Proceedings of the 25th International Conference on Distributed Computing and Networking*, pages 32–44, 2024. 1, 5
- [36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5
- [37] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021. 1
- [38] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. NuScenes-QA: A Multi-Modal Visual Question Answering Benchmark for Autonomous Driving Scenario. In *AAAI*, pages 4542–4550, 2024. 1
- [39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2016. 7, 6
- [40] Renan A Rojas-Gomez, Raymond A Yeh, Minh N Do, and Anh Nguyen. Inverting adversarially robust networks for image synthesis. In *ACCV*, pages 2221–2238, 2022. 2, 5
- [41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 5, 6

- [42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6
- [43] Aqsa Shabbir, Halil İbrahim Kanpak, Alptekin Küpçü, and Sinem Sav. A taxonomy of attacks and defenses in split learning, 2025. 7
- [44] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1874–1883, 2016. 11
- [45] Muhammad Shiraz, Abdullah Gani, Rashid Hafeez Khokhar, and Rajkumar Buyya. A review on distributed application processing frameworks in smart mobile devices for mobile cloud computing. *IEEE Communications Surveys & Tutorials*, 15(3):1294–1313, 2012. 1
- [46] Abhishek Singh, Ayush Chopra, Ethan Garza, Emily Zhang, Praneeth Vepakomma, Vivek Sharma, and Ramesh Raskar. Disco: Dynamic and invariant sensitive channel obfuscation for deep neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12125–12135, 2021. 6, 7
- [47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Int. Conf. Learn. Represent.*, pages 2256–2265, 2015. 6
- [48] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *Int. Conf. Pattern Recog.*, pages 2464–2469. IEEE, 2016. 1, 5
- [49] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Distributed deep neural networks over the cloud, the edge and end devices. In *2017 IEEE 37th international conference on distributed computing systems (ICDCS)*, pages 328–339. IEEE, 2017. 1
- [50] Tom Titcombe, Adam J Hall, Pavlos Papadopoulos, and Daniele Romanini. Practical defences against model inversion attacks for split neural networks. *arXiv preprint arXiv:2104.05743*, 2021. 6, 7
- [51] Praneeth Vepakomma, Abhishek Singh, Otkrist Gupta, and Ramesh Raskar. Nopeek: Information leakage reduction to share activations in distributed deep learning. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 933–942. IEEE, 2020. 6
- [52] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution, 2024. 11
- [53] Ruofeng Yang, Bo Jiang, Cheng Chen, Ruinan Jin, Baoxiang Wang, and Shuai Li. Few-shot diffusion models escape the curse of dimensionality. In *Adv. Neural Inform. Process. Syst.*, pages 68528–68558. Curran Associates, Inc., 2024. 3
- [54] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1191–1200, 2022. 5
- [55] Zipeng Ye, Wenjian Luo, Muhammad Luqman Naseem, Xiangkai Yang, Yuhui Shi, and Yan Jia. C2fmi: Corse-to-fine black-box model inversion attack. *IEEE Transactions on Dependable and Secure Computing*, 21(3):1437–1450, 2023. 1, 5
- [56] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Adv. Neural Inform. Process. Syst.*, 27, 2014. 3
- [57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018. 4, 5
- [58] Sai Qian Zhang, Ziyun Li, Chuan Guo, Saeed Mahloujifar, Deeksha Dangwal, Edward Suh, Barbara De Salvo, and Chiao Liu. Unlocking visual secrets: Inverting features with diffusion priors for image reconstruction. *arXiv preprint arXiv:2412.10448*, 2024. 2, 5
- [59] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 5, 2, 3
- [60] Zhanke Zhou, Jianing Zhu, Fengfei Yu, Xuan Li, Xiong Peng, Tongliang Liu, and Bo Han. Model inversion attacks: A survey of approaches and countermeasures. *arXiv preprint arXiv:2411.10023*, 2024. 1, 5
- [61] Yixuan Zhu, Wenliang Zhao, Ao Li, Yansong Tang, Jie Zhou, and Jiwen Lu. Flowie: Efficient image enhancement via rectified flow. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13–22, 2024. 2