arXiv:2411.10023v3 [cs.LG] 16 Aug 2026

Model Inversion Attacks: A Survey of Approaches and Countermeasures

ZHANKE ZHOU*, Hong Kong Baptist University, Hong Kong SAR
JIANING ZHU*, Hong Kong Baptist University, Hong Kong SAR
FENGFEI YU*, Hong Kong Baptist University, Hong Kong SAR
XUAN LI, Hong Kong Baptist University, Hong Kong SAR
XIONG PENG, Hong Kong Baptist University, Hong Kong SAR
TONGLIANG LIU, The University of Sydney, Australia
BO HAN, Hong Kong Baptist University, Hong Kong SAR

Deep neural networks have enabled numerous studies and applications on both Euclidean data, such as images and text, and non-Euclidean data, such as graphs. Because these networks may process private data, their deployment raises concerns about privacy leakage. Model inversion attacks (MIAs) exploit access to a trained model to reconstruct training examples or infer privacy-sensitive characteristics represented by the model. The effectiveness of MIAs has been demonstrated in various domains, including images, text, and graphs. These attacks highlight the vulnerability of neural networks and raise awareness about the risk of privacy leakage within the research community. This survey provides a threat-model- and assumption-aware synthesis of attacks and defenses. We compare exposed interfaces, attacker knowledge, reconstruction priors and targets, failure modes, deployment constraints, and privacy–utility trade-offs, while highlighting modeling principles, optimization challenges, and future directions. We also maintain an evolving repository of relevant research at https://github.com/AndrewZhou924/Awesome-model-inversion-attack.

CCS Concepts: • **Security and privacy → Privacy protections**.

Additional Key Words and Phrases: Model Inversion Attacks, Privacy Attacks and Defenses, Training Data Reconstruction

1 Introduction

Machine learning has driven advances in pattern recognition for Euclidean data, *e.g.,* images and text, and non-Euclidean data, commonly represented as graphs. Convolutional neural networks (CNNs) [36, 73, 74, 88, 224], language models (LMs) [30, 41, 136, 229, 233], and graph neural networks (GNNs) [68, 105, 106, 234, 255] have been developed for these domains. More recently, foundation models [1, 16, 30, 60, 230] have attracted growing interest because of their broad capabilities in face recognition [39, 134, 196, 219, 223], dialogue [3, 117, 229, 290], recommendation [23, 75, 217], and drug discovery [120, 122, 141, 249, 291].

Nevertheless, these applications process sensitive data, including private facial images, contact information, and personalized social networks, raising concerns about privacy leakage [4, 42, 85, 147, 212, 279] and motivating privacy-preserving methods [31, 49, 65, 86, 201, 256]. Model inversion attacks (MIAs) [55] exploit black- or white-box access to recover training examples or sensitive attributes across data domains (Fig. 1).

MIAs exploit model outputs or internals to infer private training information in settings such as:

- *Facial recognition* [7, 24, 55, 70, 97, 162, 163, 214, 238, 265, 275, 289], where an adversary updates candidate images using model responses to recover identity-revealing facial features or representative private images.
- *Medical diagnostics* [24, 238, 275, 289], where clinical-image models may expose sensitive class or patient information.

*These authors contributed equally to this research.

Authors' Contact Information: Zhanke Zhou, Hong Kong Baptist University, Hong Kong SAR; Jianing Zhu, Hong Kong Baptist University, Hong Kong SAR; Fengfei Yu, Hong Kong Baptist University, Hong Kong SAR; Xuan Li, Hong Kong Baptist University, Hong Kong SAR; Xiong Peng, Hong Kong Baptist University, Hong Kong SAR; Tongliang Liu, The University of Sydney, Australia; Bo Han, Hong Kong Baptist University, Hong Kong SAR.

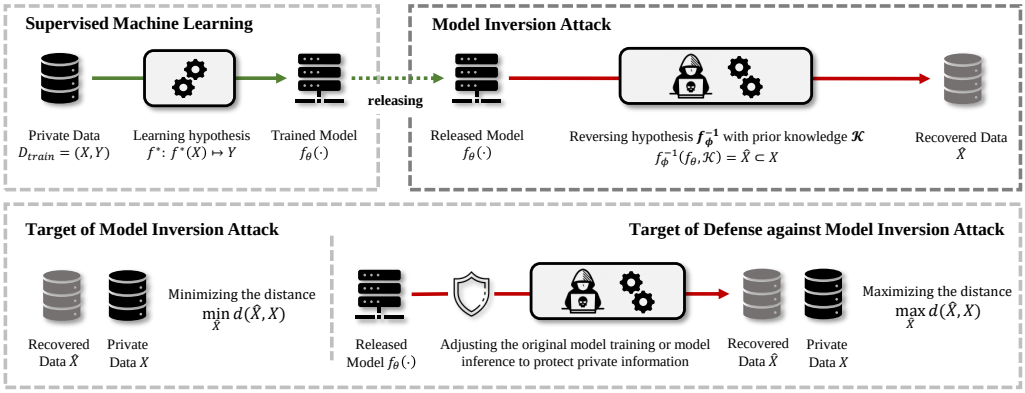


Fig. 1. Pipeline of an MIA in supervised learning: the attacker uses a released model and prior knowledge to recover private training information, while the defender limits recovery without sacrificing the target model's utility on legitimate inputs.

- *Social and recommendation systems* [20, 204], where graph-link reconstruction can expose private relationships.

Depending on the threat model, MIAs combine target access with nonsensitive auxiliary knowledge to recover images [214, 289], private text [283], or graph structure [76, 293]; label-only black-box recovery is also feasible [97].

Many attack approaches have been proposed across image, text, and graph data [55, 76, 97, 153, 283, 289, 293]; representative defenses include output restriction, representation regularization, training-time protection, and graph perturbation [80, 84, 129, 179, 240, 264]. Recent surveys by Fang et al. [54] and Yang et al. [262] already cover multiple data domains, attack taxonomies, defenses, datasets, and metrics. Our complementary focus is to compare methods through a common set of threat-model and methodological dimensions—exposed interface, attacker knowledge, auxiliary data and priors, reconstruction target, failure mode, deployment constraint, and privacy–utility boundary.

This work provides a threat-model- and assumption-aware survey of model inversion attacks. Our contributions are:

- We give a threat-model-centric account of MIAs, separate them from gradient inversion, membership inference, and model stealing, and position this survey against the closest prior ones (Sec. 2, Tab. 1).
- We compare every attack along one frame—exposed interface, assumed knowledge and prior, recovered target, failure conditions—across the image, text, graph, collaborative/split and vision-language settings (Secs. 3–5).
- We quantify what defenses cost, from the privacy–utility evidence their own papers report (Tab. 6) to the deployment and query budgets an attacker faces (Tab. 8), and summarize datasets, metrics and open problems (Secs. 6–8).

2 Overview

2.1 The Problem Definition of Model Inversion Attacks

Definition 2.1 (Supervised machine learning). *Given training and test sets $D_{train} = (X_{train}, Y_{train})$ and $D_{test} = (X_{test}, Y_{test})$, supervised learning estimates $f_\theta \in \mathcal{H}$ to map inputs X to labels or targets Y. Its parameters are commonly optimized on D_{train} by stochastic gradient descent, and its generalization is evaluated on D_{test} before deployment.*

Definition 2.2 (Model inversion attacks). *Given a trained model f_θ and prior knowledge $\mathcal{K}$, a model inversion attack seeks a reverse hypothesis f_ϕ^{-1} that recovers training data X_{train} or privacy-sensitive characteristics represented by f_θ. For sample-reconstruction attacks, $f_\phi^{-1}(f_\theta, \mathcal{K}) = \hat{X}_{train}$, where the recovered data $\hat{X}_{train}$ is a set of data samples that are expected to approximate those samples in X_{train}.*

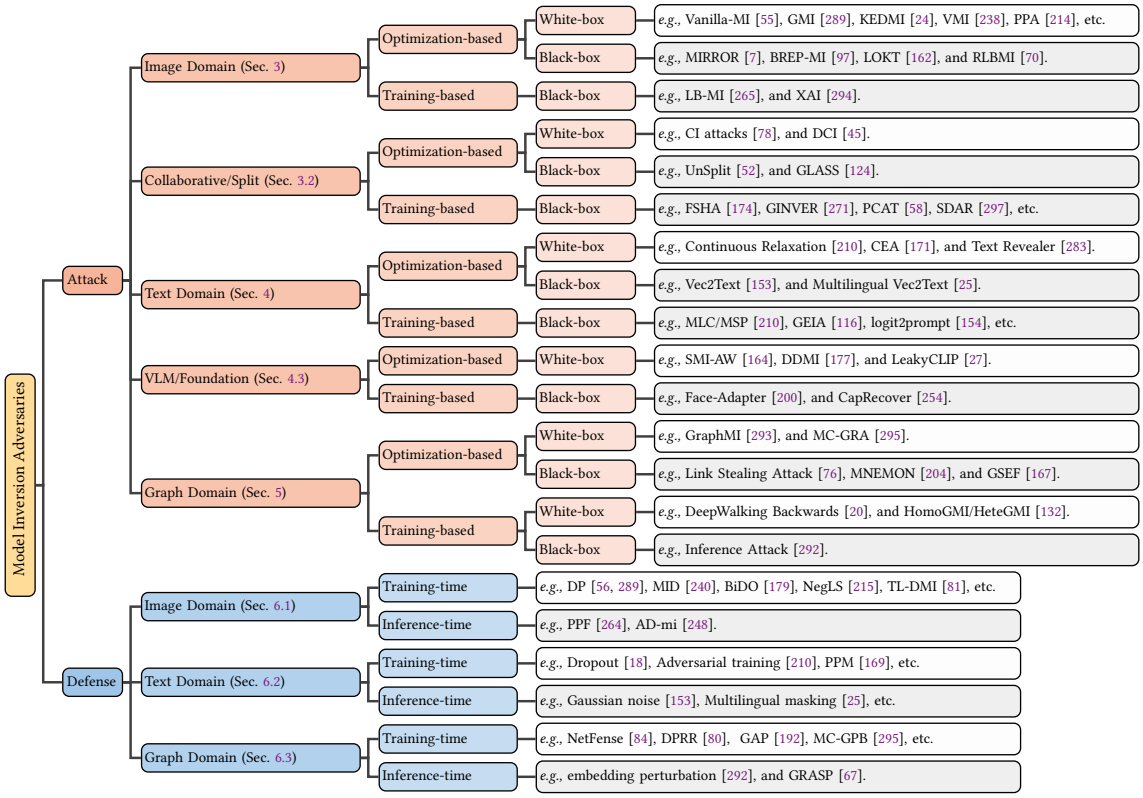


Fig. 2. Taxonomy of MIAs and defenses over the image, text and graph domains and over the collaborative/split and vision-language/foundation-model settings, with detailed coverage in Sections 3–6.

Remark 2.1 (Strict model inversion attacks). *When no prior knowledge is given, namely, $\mathcal{K} = \emptyset$, then an MIA reduces to finding the reverse hypothesis f_{ϕ}^{-1} that $f_{\phi}^{-1}(f_{\theta}) = \hat{X}_{train}$. This is a limiting case: pioneering studies typically used auxiliary information such as known nonsensitive attributes, a target identity or class, or public-data priors [55, 56].*

Remark 2.2 (The extent of inversion). *MIAs usually recover representative examples, attributes, graph relations, or a subset of X_{train}, rather than the entire training set. Evaluation therefore uses domain-appropriate measures of fidelity, attack success, distributional similarity, or recall. Full-set reconstruction is obtained only when $\hat{X}_{train}$ covers all samples in X_{train}.*

Definition 2.2 has three critical factors:

- **Factor 1: The domain-specific MIA methods to obtain f_{ϕ}^{-1}.** Fig. 2 organizes MIA designs by data domain and by the collaborative/split and vision-language settings that cut across them, because reconstruction difficulty depends on data structure, semantics, and the interface a deployment exposes.
- **Factor 2: The information contained in $\mathcal{K}$.** Many MIA settings assume domain knowledge and public auxiliary data but no private training records [163, 214, 238]; for example, public facial images can provide a generative prior. Stronger threat models assume information related to the private distribution or target records [289], such as partial or corrupted observations whose nonsensitive content constrains and narrows recovery in practice.

- **Factor 3: The white-box or black-box access of f_θ.** In a *white-box* attack, the adversary knows the target model's architecture and parameters and can compute gradients through it. In a *black-box* attack, the adversary does not know the architecture or parameters and cannot compute gradients through the target model. Restricted-access MIAs can be grouped by the exposed interface: (1) *full confidence scores*, from which the adversary can also compute a target label's prediction loss (*e.g.,* cross-entropy loss); (2) *hard label only*, where only the predicted label is returned and the adversary is given the most limited information; (3) *hidden representation*: the adversary reconstructs input information from a feature or embedding deliberately exposed by the target-model interface.

2.2 Principles for Enhancing or Defending against Model Inversion Attacks

General design principles transfer across methods; we therefore state the adversary's and defender's objectives:

Definition 2.3 (Target of MIAs)**.** *Given a trained model f_θ and prior knowledge $\mathcal{K}$, the target of conducting MIAs is to find the reverse hypothesis f_ϕ^{-1} that minimizes distance$(\hat{X}_{train}, X_{train})$, where $\hat{X}_{train} = f_\phi^{-1}(f_\theta, \mathcal{K})$. Notably, the ground truth data X_{train} is not accessible for conducting MIAs and is only used in the final evaluation of distance$(\hat{X}_{train}, X_{train})$.*

Definition 2.4 (Target of defending against MIAs)**.** *Given a dataset $D = \{D_{train}, D_{test}\}$ and an attack method to get the reverse hypothesis f_ϕ^{-1}, the defender trains a model f_θ on $D_{train} = (X_{train}, Y_{train})$ to (1) increase the reconstruction error distance$(\hat{X}_{train}, X_{train})$ under the post-hoc MIA and (2) preserve predictive performance on $D_{test} = (X_{test}, Y_{test})$. In practice, these objectives define a privacy-utility trade-off rather than two quantities that can always be maximized simultaneously.*

Next, we summarize principles for conducting MIAs toward the objective in Definition 2.3.

Principle 1. *Improve query strategies. Adapt queries to earlier responses so that the available budget more efficiently narrows the candidate space, estimates a decision boundary, or trains a surrogate under restricted access.*

Principle 2. *Leverage model internals. Exploit parameters and intermediate representations under white-box access to guide reconstruction more directly and reliably than black-box outputs permit during iterative inversion.*

Principle 3. *Exploit output probabilities. Use output probabilities to define reconstruction losses, rank candidates, or guide candidate updates with richer information than hard labels alone provide during reconstruction.*

Principle 4. *More prior knowledge. Use statistical and domain-specific priors to constrain the search and make better-informed guesses about the structure and distribution of private data before and during optimization.*

Principle 5. *Extra generative models. Use pre-trained GAN or diffusion priors and search their learned space rather than raw inputs; when matched to private data, these priors yield realistic, semantically meaningful reconstructions.*

Principle 6. *Transfer pre-trained models. Adapt models pre-trained on public data as inversion priors, optionally fine-tuning them with public data or target-model responses before subsequent reconstruction for the target task.*

Defenses follow complementary principles:

Principle 7. *Data-centric processing. Sanitize or obfuscate sensitive training information, perturb individual records, or replace them with synthetic examples that preserve utility without directly exposing real records.*

Principle 8. *Regularized representation learning. Penalize dependence between learned representations and private inputs so the features retain less record-specific information while supporting the intended prediction task.*

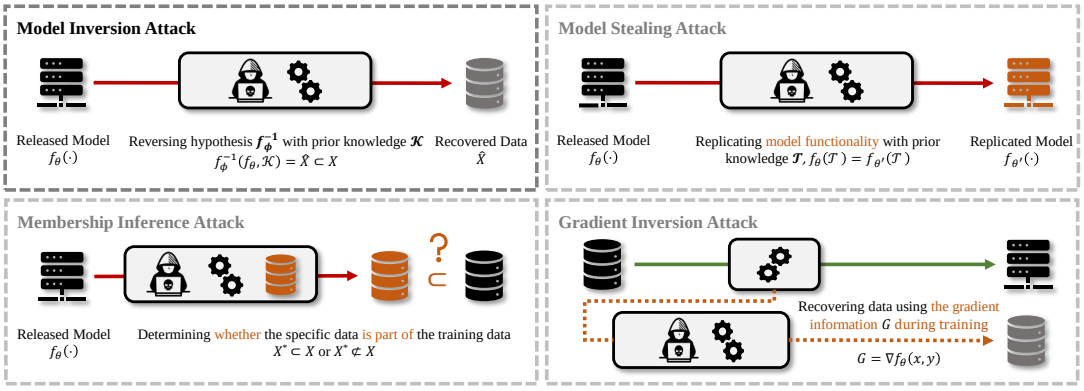


Fig. 3. Comparison of privacy attacks by objective and the information they may expose.

Principle 9. *Feature masking and dropout. Mask sensitive features or apply dropout to reduce reliance on individual inputs or hidden units and discourage brittle, record-specific encoding; neither operation guarantees privacy by itself.*

Principle 10. *Modify model output. Perturb, coarsen, or randomize released outputs to obscure exploitable input-output relationships; calibrated privacy mechanisms must explicitly balance leakage reduction against downstream utility.*

2.3 Relationship and Differences with Other Privacy Attacks

Fig. 3 compares MIAs and related privacy attacks by objective and exposed information.

- **Model stealing attacks** [22, 43, 48, 52, 87, 94, 98, 165, 168, 205, 244, 245, 259, 260, 276, 298]. Model stealing replicates target functionality, often by training a surrogate on query-response pairs. It threatens model functionality and intellectual property rather than reconstructing private training examples themselves.
- **Membership inference attacks** [13, 29, 72, 77, 102, 130, 137, 144, 145, 166, 187, 203, 207, 213, 232, 243, 266, 277, 281]. Membership inference attacks determine whether a given record belonged to a model's training set. Depending on the access setting, they exploit losses, labels, confidence scores, gradients, or trained attack models that distinguish members from nonmembers; revealing membership can itself disclose sensitive participation.
- **Gradient inversion attacks** [12, 40, 53, 59, 61, 71, 90, 95, 181, 251, 269, 286, 296]. Gradient inversion attacks reconstruct inputs from shared gradients in federated, distributed, or split learning. They optimize candidates to match observed gradients, sometimes using training-based inversion models or generative priors. The training gradient is their defining observable, unlike output- or parameter-based MIAs in our taxonomy.

2.4 Positioning with Respect to Existing Surveys

Prior surveys approach MIAs at different scopes. Broad security surveys treat MIAs alongside membership inference, model stealing, or adversarial attacks [5, 190], while others focus on machine-learning-as-a-service (MLaaS), attack tools, or gradient inversion [123, 206, 211, 212, 282]. Most relevantly, Fang et al. [54] cover image, text, and graph MIAs and release an attack toolbox, while Yang et al. [262] survey deep-learning attacks, defenses, datasets, metrics, and challenges. Our distinctive contribution is the analytical framework summarized in Tab. 1, rather than cross-domain breadth by itself.

Table 1. Positioning relative to the most relevant MIA surveys. Entries describe the organizing emphasis actually provided by each work rather than a binary claim of topic absence.

Survey	Scope	Primary analytical organization	Emerging interfaces	Deployment, defense, and resources
Fang et al. [54]	ML/DNN MIAs across modalities and tasks	Attack/defense taxonomy and method features	Coverage through its 2024 cutoff	Attack toolbox and open issues
Yang et al. [262]	Deep-learning attacks, defenses, data, applications, and evaluation	Technique- and defense-oriented taxonomy	Generative and gradient-reconstruction lines through 2025	Challenges, directions, and resource repository
Gradient-inversion surveys [206, 282]	Federated/distributed gradient reconstruction	Threat models, matching procedures, and defenses	Shared-gradient interface	Federated assumptions and defenses
This survey	Image, text, and graph MIAs; bounded adjacent settings	Interface, knowledge, prior, target, mechanism, and failure mode	VLM responses, shared embeddings, LLM signals, diffusion priors, and split features	Deployment and privacy–utility analysis; datasets, metrics, and literature index

Our complementary emphasis is a *threat-model- and assumption-aware synthesis*: we compare attacks by exposed interface, auxiliary knowledge and priors, reconstruction space and procedure, and query, transfer, and prior-mismatch limitations. These dimensions connect optimization-based and training-based mechanisms to concrete deployment scenarios and clarify modality-specific risks: image MIAs rely strongly on visual priors, text MIAs bridge continuous signals and discrete sequences, and graph MIAs exploit relational structure. Relating these assumptions to defenses and evaluation explains why methods with similar recovery goals can have different practical implications.

3 Model Inversion Attacks on Image Data

Sec. 3.1 introduces the general methodology of model inversion [47, 143, 161, 208] in the image domain. It seeks inputs that either activate a network feature [47, 270, 272] or produce a high class response [62, 208, 239], giving optimization-based [47, 62, 161, 176, 178, 208, 239, 270, 272] and training-based approaches [46, 47, 158] (Fig. 4). Some precede privacy attacks but provide their methodological basis [55]; Sec. 3.2 applies them across threat settings.

3.1 A Summary of Model Inversion Approaches on Image Data

To make threat-model assumptions and methodological differences explicit, Tab. 2 compares attacks on image classifiers and face-embedding models by attacker interface or capability, reconstruction space and prior, auxiliary resources, distinctive mechanism, and granularity of recovered information; these complementary dimensions avoid conflating model access, generative priors, and solution procedures. Because collaborative/split inference and split learning expose intermediate representations under distinct server roles, Tab. 3 compares them separately.

Let $\mathbf{x}$ be a training input and $y = f_\theta(\mathbf{x})$ its prediction or representation. Inversion approximates the posterior $p(\mathbf{x} \mid y)$ either by per-target optimization or by a learned inverse mapping, the two paradigms detailed below.

Optimization-based Approaches. Optimization-based methods seek an image $\hat{\mathbf{x}}$ whose output matches y, either in pixel space [143, 270, 272] or in a generator's latent space. Because this problem is ill-posed and non-convex, an image prior $\mathcal{R}$ regularizes the search and favors plausible natural images during reconstruction [143, 208]:

$$\hat{\mathbf{x}}^* = \arg\min_{\hat{\mathbf{x}} \in \mathcal{X}} \mathcal{L}(f_\theta(\hat{\mathbf{x}}), y) + \lambda \mathcal{R}(\hat{\mathbf{x}}), \tag{1}$$

where $\mathcal{L}$ measures output discrepancy and $\mathcal{R}$ may encode an α-norm [208], total variation [143], or feature statistics [62, 272]. Beyond hand-crafted image priors, the image prior can also be learned from data through generative models, such as GANs [24, 214, 289] and diffusion models [118, 237]. Instead of directly optimizing in the raw pixel space, the adversary can search in a learned latent or generative space, which constrains reconstructed samples to be more realistic

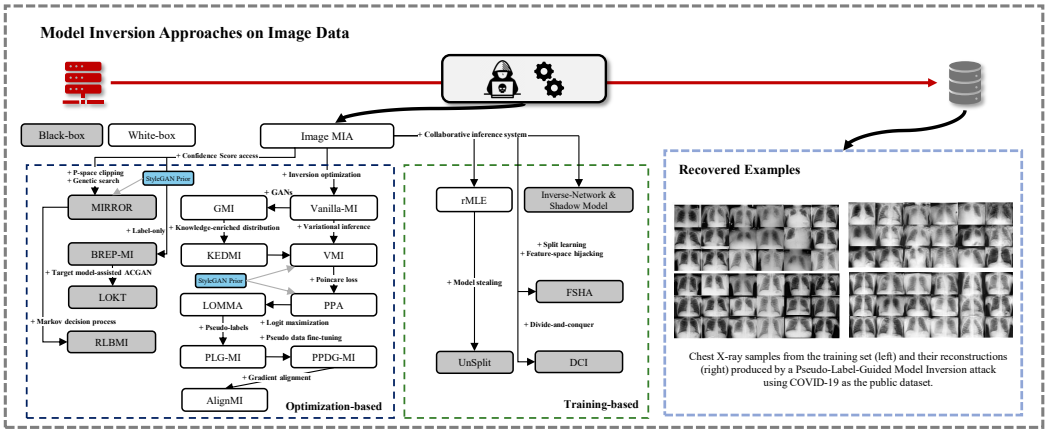


Fig. 4. Evolution of image-domain MIAs with representative recovered examples.

and semantically meaningful. Such learned priors provide stronger regularization for the ill-posed inversion problem, while their effectiveness depends on how well the learned distribution matches the private data distribution.

Training-based Approaches. Training-based methods instead learn a decoder $f^{-1}(\cdot;\boldsymbol{\phi})$ from auxiliary pairs $\{(\mathbf{x}_i, y_i) : y_i = f_{\theta}(\mathbf{x}_i)\}$ [46, 47]. For example, a pixel-space decoder can minimize

$$\hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}} \sum_i \|f^{-1}(y_i;\boldsymbol{\phi}) - \mathbf{x}_i\|_2^2. \tag{2}$$

Unlike optimization-based methods, which iteratively update the input or latent code per target, training-based methods learn a direct mapping from predictions or features to images, reconstructing in one forward pass once trained.

3.2 Applying Image MIAs to Different Scenarios

We organize image-domain MIAs into attacks on *standard classifiers* and on *collaborative/split inference*.

MIAs on Standard Classification. (Principle 1: [97]; Principle 3: [55, 214, 289]; Principle 2: [163]; Principle 4: [24, 265, 294]; Principle 5: [7, 70, 118, 131, 163, 214, 237, 238, 289]) Here an adversary infers private-data representatives from a trained classifier (Fig. 5). We compare attacks along three dimensions: the *attacker interface* (white-box gradients or logits, confidence scores, hard labels, or explanations); the *reconstruction space and prior* (pixels or latent spaces induced by GANs, StyleGANs, or diffusion models); and the *solution procedure* (per-target optimization, derivative-free search, boundary estimation, transfer, or amortized inference). Weaker interfaces require extra queries, priors, or surrogates.

White-box optimization attacks. These attacks use target gradients or internal activations to optimize pixels or a latent representation, under hand-crafted regularization, GAN/StyleGAN priors, or diffusion priors; the generative prior is thus a reconstruction mechanism rather than a separate threat model. Their development pursues two related goals: confining reconstructions to the manifold of realistic images, and obtaining an optimization signal that remains stable and discriminative of the target identity throughout the search process for each target.

► **The first MIA algorithm.** Fredrikson et al. [55] optimize pixels under denoising and sharpening priors (Eq. 1) to invert a face classifier; such priors encode little semantic structure, so on deep models input-space search degenerates into adversarial samples [222] that score highly under the target yet carry no semantic meaning.

Table 2. Fine-grained comparison of representative model inversion attacks against image classifiers and face-embedding models. "Interface" denotes target-model access; "Auxiliary resource" denotes data or models required beyond the target.

Method	Interface / capability	Reconstruction space / prior	Auxiliary resource	Distinctive mechanism	Recovered information
Vanilla-MI [55] (2015)	Model confidence / gradients	Pixels + hand-crafted priors	None	Direct input optimization with denoising and sharpening	Class-representative image
LB-MI [265] (2019)	Prediction vectors	Learned inverse model	Generic public query images	Align background knowledge and learn output-to-image mapping	Input-like image
GMI [289] (2020)	White-box gradients	GAN latent space	Public images	Joint discriminator-prior and identity-guided latent search	Class / identity prototype
KEDMI [24] (2021)	White-box predictions / gradients	GAN + latent distribution	Public images + target soft labels	Knowledge-enriched GAN and distributional inversion	Class distribution / identity
VMI [238] (2021)	White-box predictions / gradients	StyleGAN + flow distribution	Pretrained generator	Variational posterior instead of a point estimate	Diverse class distribution
XAI-MI [294] (2021)	Predictions + explanations	Learned inverse model	Auxiliary query images	Fuse saliency or other explanation signals into inversion	Input instance
PPA [214] (2022)	White-box logits / gradients	Pretrained StyleGAN	Pretrained generator	Poincare loss, transformations, and sample selection	High-resolution identity
MIRROR [7] (2022)	Confidence-score black-box	Pretrained StyleGAN	Pretrained generator	P-space clipping and genetic-algorithm latent search	High-resolution identity
BREP-MI [97] (2022)	Hard labels	Generator latent space	Pretrained generator	Estimate direction by repelling samples from the decision boundary	Class / identity prototype
LOMMA [163] (2023)	Internal logits, features, and gradients	GAN / StyleGAN latent space	Public images + augmented models	Logit maximization and model augmentation	Identity prototype
PLG-MI [275] (2023)	White-box predictions / gradients	Conditional GAN	Public images + pseudo-labels	Class-decoupled prior and max-margin inversion	Class / identity prototype
RLBMI [70] (2023)	Confidence scores	Generator latent space	Pretrained generator	Soft Actor-Critic policy for sequential latent search	Identity prototype
LOKT [162] (2023)	Hard labels	T-ACGAN + surrogate models	Public unlabeled images	Transfer target behavior and then perform white-box inversion	Identity prototype
PPDG-MI [176] (2024)	White-box gradients	Adapted GAN prior	Public + pseudo-private images	Iteratively select reconstructions and tune the generator	Identity prototype
Diff-MI [118] (2024)	White-box target knowledge	Target-specific conditional diffusion	Public images + pseudo-labels	Pretrain/fine-tune diffusion prior and top-k margin guidance	Class / identity prototype
Unstoppable [131] (2024)	Hard labels	Conditional diffusion	Task-related auxiliary images	Train diffusion with target-predicted labels as conditions	Representative class samples
AlignMI [178] (2025)	Gradients or confidence scores	GAN manifold	Public images + pre-trained generator	Align updates with the generator tangent space	Identity prototype
SMILE [125] (2025)	Confidence scores	Pretrained generator	Public images + surrogate model	Long-tailed surrogate initialization and NGOpt search	High-resolution identity
DiffMI [237] (2026)	Embeddings / similarity signal	Fixed unconditional diffusion	Pretrained diffusion; no target-specific training	Robust initialization and ranked adversarial latent refinement	Identity-consistent face

► **The first GAN-based generative MIA.** GMI [289] introduces a GAN prior [66, 101, 185] for attacking deep classifiers. It first trains a Wasserstein GAN [8] on public facial data, then searches its latent space for realistic samples that the target assigns to the desired identity, recovering training data with far higher fidelity than the earlier pixel-space attack. Its objective is $\mathbf{z}^* = \arg\min_{\mathbf{z}} -\mathrm{D}(\mathrm{G}(\mathbf{z})) - \lambda \log f_{\boldsymbol{\theta}}(\mathrm{G}(\mathbf{z}))$, where a lower prior loss demands more realistic images and a lower identity loss encourages a higher prediction score under $f_{\boldsymbol{\theta}}$ for the target class.

► **Distill class information from the target model.** Instead of only discriminating real from generated samples, KEDMI [24] uses target-model *soft labels* on public data to provide class-aware supervision during GAN training. It also replaces the usual point estimate with latents drawn from a learnable Gaussian $p_{\text{gen}} = \mathcal{N}(\mu, \sigma^2)$, recovering diverse samples that capture class-level semantics.

► **MIAs as variational inference.** Earlier generative MIAs performed well but rested on an unclear theoretical basis. VMI [238] approximates $p_{\text{TAR}}(\mathbf{x} \mid y)$ with a flow-based variational distribution and a StyleGAN generator [101],

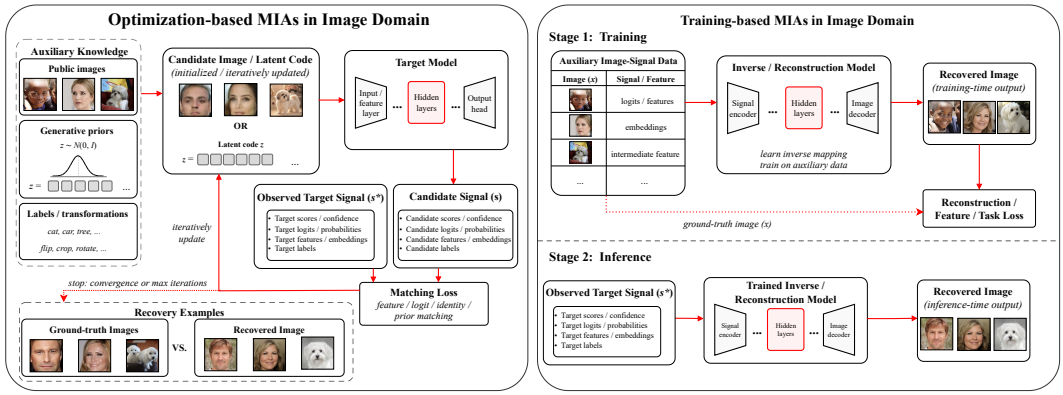


Fig. 5. Illustrations of model inversion attacks on standard image classifiers.

minimizing $\mathbb{E}_{q(\mathbf{x})}[-\log p_{\text{TAR}}(y \mid \mathbf{x})] + D_{\text{KL}}(q(\mathbf{x}) \| p_{\text{TAR}}(\mathbf{x}))$; recovering a distribution rather than a single optimum improves sample diversity, represents multiple plausible modes, and avoids a single point estimate.

► **Practical high-resolution MIAs.** Struppek et al. [214] identify four degradation factors: entanglement between image priors and target models, distributional shift between priors and private data, vanishing gradients of the cross-entropy objective, and non-robust targets. PPA [214] combines a pre-trained StyleGAN prior with cropping and resizing to reduce the *distributional shift* between prior and private data; a Poincaré loss mitigates vanishing gradients, random transformations counter non-robust targets, and high-resolution reconstructions are finally selected.

► **Logit maximization and model augmentation.** LOMMA [163] replaces GMI and KEDMI's cross-entropy identity loss [24, 289] with *logit maximization* (LOM), which maximizes the target logit $h^\top w_k$ from the penultimate-layer activations h and last-layer weights w_k. They also first define *MI overfitting*, whereby one target model lets random variation bias the reconstruction, countered by an ensemble of augmented targets during inference.

► **Pseudo-labels and max-margin loss.** PLG-MI [275] addresses the class-coupled latent space, the indirect constraints on the generator, and the gradient vanishing of earlier GAN-based MIAs. It uses *target pseudo-labels* to train a conditional GAN, separating class-specific latent spaces more directly than earlier GAN-based MIAs [24, 289]; during inversion, a max-margin objective replaces cross-entropy to reduce gradient vanishing directly.

► **Iterative generator adaptation.** Conventional MIAs rely on a static prior learned from public data. Pseudo-Private Data Guided Model Inversion (PPDG-MI) [176] reduces this gap by keeping reconstructions with stable predictions as pseudo-private data and iteratively fine-tuning the generator to increase their density, narrowing the prior-private gap.

► **Enhance MIAs via gradient-manifold alignment.** AlignMI [178] revisits generative inversion under the manifold hypothesis: backpropagation through a generator implicitly filters loss gradients along the generator manifold's tangent space, while off-manifold components hinder inversion. AlignMI does not explicitly compute this tangent-space projection during an attack. Instead, it samples perturbed or semantically transformed variants x' around the current synthetic input x and averages their loss gradients as $\widetilde{\nabla}\mathcal{L}(x) = \mathbb{E}_{x' \sim p(\cdot|x)}[\nabla\mathcal{L}(x')]$. The resulting training-free smoothing improves inversion across multiple white-box generative MIAs and image resolutions.

► **Diffusion priors for model inversion.** A recent line of work replaces or complements GAN priors with diffusion priors for better realism and semantic consistency. Diff-MI [118] trains a target-specific conditional diffusion model for target-guided reconstruction, showing the strength of diffusion priors when model-specific generation is

affordable. DiffMI [237], despite the similar name, targets embedding-based face recognition and refines latent codes of a fixed pre-trained diffusion model, reconstructing identity-consistent faces without target-specific generator training. Unstoppable [131] extends diffusion-prior MIAs to the label-only setting by conditioning generation on predicted labels. Together, they span target-specific, training-free, and label-only inversion, with distinct attacker-knowledge, generation-cost, and prior-alignment trade-offs across access settings and deployment budgets.

Comparison of reconstruction mechanisms under white-box access. Moving from pixel optimization to learned generative priors improves realism without removing the dependence on attacker assumptions. The original pixel-space attack [55] needs no generator, yet readily produces high-confidence images without semantic meaning. GMI, KEDMI, and PLG-MI increasingly exploit public data and target-model knowledge for realism and class specificity, at the price of public-private distribution alignment and generator coverage. VMI recovers a distribution rather than a point estimate and so improves diversity; PPA and LOMMA instead address optimization stability, robustness, and overfitting, with LOMMA additionally assuming internal logits and features; PPDG-MI and AlignMI adapt the prior or the optimization geometry. Diffusion priors give a more expressive reconstruction space but are not white-box-specific, having also been used under label-only access, and their cost depends on whether target-specific training, distillation, or iterative generation is required. Recurring failure modes include prior mismatch, incomplete coverage of private identities, vanishing or off-manifold gradients, and realistic but training-unrepresentative reconstructions.

Score-based black-box attacks. When gradients and model internals are unavailable, confidence scores can still rank candidate reconstructions. These attacks replace backpropagation with derivative-free latent search or a training-based policy, trading weaker access assumptions for more queries during reconstruction.

► **StyleGAN priors and black-box search for high-fidelity inversion.** An et al. [7] note that plain GAN priors as used in GMI [289] have limited generative capacity and entangled latent spaces, and propose MIRROR, which combines three components: (i) a pre-trained StyleGAN prior whose style-disentangled latent space enables higher-fidelity reconstruction; (ii) *latent clipping* that confines the search to StyleGAN's intermediate P-space, keeping candidates on the natural-image manifold and mitigating entanglement; and (iii) a genetic-algorithm search [14] optimizing the inversion objective where gradients are unavailable. The StyleGAN prior is not unique to MIRROR—VMI [238] instantiates its variational family with a StyleGAN generator and PPA [214] also builds on such priors—so MIRROR's distinctive contribution is the P-space clipping and genetic-algorithm black-box search built atop this shared prior.

► **MIAs as a Markov decision process.** RLBMI [70] replaces MIRROR's genetic search [7] with a Soft Actor-Critic policy: the latent vector is the state, and confidence-derived rewards guide actions through the search space. More recently, SMILE [125] combines long-tailed surrogate training with gradient-free optimization to obtain stronger latent-search initialization and high-resolution reconstructions with fewer queries in practice.

Comparison of score-based attacks. MIRROR, RLBMI, and SMILE all rely on confidence scores and a strong generative prior, but allocate computation differently: MIRROR avoids surrogate training at the cost of a query-hungry population search, RLBMI learns a sequential search policy that may improve exploration but adds reinforcement-learning overhead, and SMILE pays an upfront surrogate-training cost for stronger initializations and fewer subsequent queries. All degrade under score truncation, output perturbation, rate limits, or public-private shift; evaluations should report accuracy together with query count, prior-training cost, and exposed output information.

Label-only attacks. Hard-label interfaces remove confidence information, yielding a stricter, more deployment-realistic threat model. Attacks compensate by estimating local decision boundaries through repeated queries or transferring target-label behavior to a differentiable surrogate under restricted access in realistic deployments.

► **Gradient estimation in the label-only setting.** BREP-MI [97] is the first attack to require only hard labels. It probes N latent points on a sphere of radius R, expands the sphere while all generated samples retain the target label, and otherwise estimates a boundary-based search direction from the observed label changes.

► **Reformulate black-box inversion as white-box inversion.** LOKT [162] queries hard labels for synthetic samples from a target model-assisted ACGAN (T-ACGAN), trains its discriminator-classifier pair as differentiable surrogates, then applies white-box inversion. The label-only attack is thereby reformulated as a white-box inversion on the surrogates, and unlike BREP-MI's per-target boundary search [97] it amortizes target queries across attacks.

Comparison of label-only attacks. BREP-MI estimates a search direction around the decision boundary without assuming a differentiable surrogate, but its local probing is query-intensive and sensitive to boundary geometry. LOKT instead amortizes queries into surrogate training and then applies white-box inversion, trading online boundary search for synthetic-data generation, surrogate-training overhead, and transfer error. Both leak under minimal outputs, yet depend more than score-based attacks on query budget, initialization, and approximation fidelity.

► **Training-based MIAs with representation challenges.** For a face classifier the adversary need not know the private identities, only that the data are facial images, so a corpus crawled from the Internet suffices. Yang et al. [265] query a black-box target with generic same-domain images to construct $(\mathbf{x}, y)$ pairs and train the decoder in Eq. 2, without knowing private identities. XAI-aware inversion [294] additionally exploits spatial signals in saliency maps, feature visualizations, activations, or concept vectors, showing that interpretability and privacy can be in tension because such explanations may still expose information exploitable by an adversary under that interface.

Cross-paradigm trade-offs. Optimization-based attacks solve a fresh search problem per target, which suits small numbers of reconstructions, whereas training-based attacks pay auxiliary-data, query, and training costs upfront and then reconstruct in one forward pass. Explanation-aware inversion gains richer spatial signals but assumes an interface exposing more than ordinary predictions, so it is a distinct threat model. No family is uniformly strongest: comparison requires jointly reporting target access, auxiliary-data alignment, generator or simulator pretraining, query budget, per-target cost, and whether the output captures class attributes, identity, or a training instance.

MIAs on Collaborative/Split Inference and Split Learning. (Principle 2: [45, 78]; Principle 4: [58, 174, 257, 297]; Principle 5: [124, 271]) Collaborative inference partitions a DNN between a client and a server: the client computes an intermediate representation, the *smashed data*, and sends it to the server, which completes the prediction. This reduces client-side computation but exposes a representation that may retain enough information to reconstruct the input. We distinguish *split inference*, where the server observes forward activations at prediction time, from *split learning*, where activations and gradients are exchanged during training. *Split federated learning* further combines this partitioned training with aggregation across multiple clients. A *passive* (honest-but-curious) server follows the protocol and analyzes observed messages, whereas an *active* server manipulates gradients or training objectives. These distinctions materially change reconstruction feasibility and assumptions, as summarized in Tab. 3. Parameter-gradient inversion alone is an adjacent federated-learning setting rather than a post-training MIA, and is excluded here.

► **Foundational attacks under varying client knowledge.** He et al. [78] give the first systematic study of MIAs against collaborative inference: for an observed intermediate representation $v = f_{\theta_1}(\mathbf{x})$, their white-box attack optimizes a candidate input directly, their black-box attack trains an inverse network from input-representation pairs, and their query-free attack first approximates the client model. Reconstruction risk therefore depends not only on the cut layer but on the attacker's model access and auxiliary-data knowledge. Dong et al. [45] then propose data-free divide-and-conquer inversion (DCI), decomposing the client network into shallow inversion modules that use stored batch-normalization statistics. DCI avoids real auxiliary samples but assumes access to the client model in practice.

Table 3. Representative attacks in collaborative and split systems: stages, signals, assumptions, mechanisms, targets, and limits.

Attack	Stage / observed signal	Attacker knowledge/capability	Aux. data / prior	Core mechanism	Target	Main limitation
Foundational CI attacks [78]	Inference / smashed data	White-box, black-box, or query-free	Varies by setting	Input optimization, inverse decoder, or shadow client	Input	Access, auxiliary data, and cut-layer dependent
FSHA [174]	Training / activations and gradients	Active server controls returned gradients	Public same-domain data	Feature-space hijacking and inverse decoder	Training inputs	Protocol deviation is potentially detectable
DCI [45]	Inference / smashed data and BN statistics	Client model and BN statistics	None	Divide-and-conquer layer inversion	Input	Requires client-model and stored-BN access
UnSplit [52]	Training / activations and gradients	Passive server; client architecture	None	Alternating model stealing and input optimization	Input, model, label	Architecture assumption and iterative cost
GINVER [271]	Inference / smashed data	White-box model or black-box queries	None in black-box setting	Generative inverse mapping; NES black-box optimization	Input	Model access or repeated queries
PCAT [58]	Training / activations and server snapshots	Unknown client model	Small labeled same-task dataset	Pseudo-client training and inverse decoder	Input, model, label	Labeled auxiliary data and training history
GLASS [124]	Inference / protected smashed data	Passive server	Pre-trained StyleGAN prior	Latent search with feature matching	Input	Prior mismatch and per-target search
FORA [257]	Training / smashed data	No client architecture or weights	Public auxiliary data	Domain discriminator and MK-MMD alignment	Training inputs	Auxiliary-domain alignment
SDAR [297]	Training / smashed data	No client weights or queries; simulator architecture	Labeled same-domain data	Adversarially regularized simulator and decoder	Inputs, labels	Extra training and simulator/domain mismatch

► **Active attacks and joint model reconstruction in split learning.** Unlike passive inference-time attacks, the Feature-Space Hijacking Attack (FSHA) [174] assumes a malicious server that changes the learning objective through forged gradients, aligning private client representations with an attacker-controlled feature space so that an inverse decoder trained on public data recovers client inputs; it is powerful but requires active protocol deviation. UnSplit [52] instead follows the protocol, alternating between estimating the client model and optimizing the private input, and jointly enables model inversion, model stealing, and label inference. PCAT [58] handles unknown client architectures by using same-task data and server snapshots to approximate the client model and recover private inputs and labels.

► **Generative priors and passive feature reconstruction.** Recent attacks improve reconstruction without manipulating the protocol. GINVER [271] learns a generative inverse mapping from exposed intermediate features, its black-box variant optimizing the inverse model by Natural Evolution Strategies without auxiliary data. GLASS and GLASS++ [124] search a pre-trained StyleGAN's latent space to match protected smashed data, showing feature perturbation alone may not prevent visually faithful recovery. FORA [257] assumes no access to client architecture, parameters, or private data, exploiting representation preference in observed smashed data and aligning a surrogate client's features with a domain discriminator and multi-kernel maximum mean discrepancy. SDAR [297] likewise targets a passive server without client weights or query access, using adversarial regularization to learn a decodable client simulator from labeled auxiliary data and evaluating known and unknown client architectures.

► **Comparative observations.** The trade-off here is between attacker capability and prior knowledge. Active attacks such as FSHA deliberately increase leakage but are more detectable and need control of training messages, whereas passive attacks better match an honest-but-curious cloud server yet typically require client-model access, a distribution-matched auxiliary dataset, or a strong generative prior. Deeper cut layers discard more low-level detail, but their protection depends on architecture and task, so split depth alone is not a reliable guarantee. Benchmarking studies [209, 261] should jointly measure attack knowledge, cut position, defense strength, utility, and pixel fidelity.

4 Model Inversion Attacks on Text and Multimodal Data

In natural language processing (NLP), models such as BERT [41] support text embedding, classification, and language modeling, but their growing complexity and deployment intensify privacy concerns. Tab. 4 summarizes text-domain MIAs and Fig. 6 their chronology. Sec. 4.1 introduces optimization-based and training-based approaches; Sec. 4.2 then

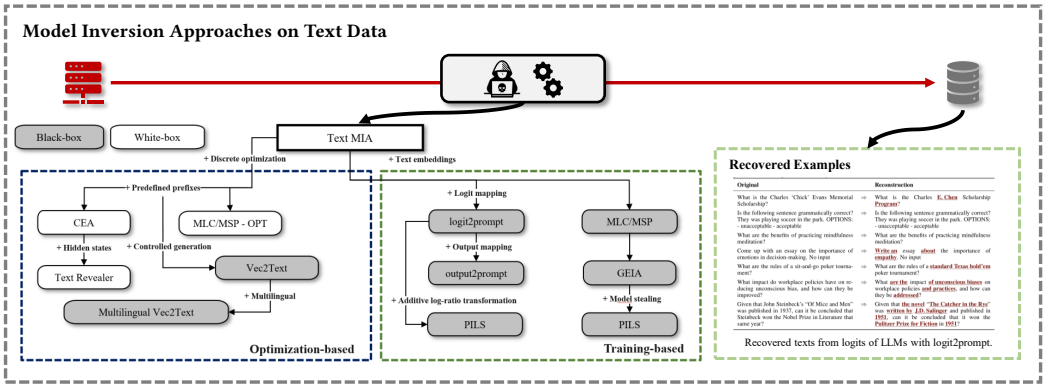


Fig. 6. Evolution of text-domain MIAs with representative recovered examples.

organizes attacks by embedding, text-classification and language-generation interfaces, and Sec. 4.3 covers vision-language and foundation models.

4.1 A Summary of Model Inversion Approaches on Text Data

Let $\mathbf{x}$ denote private text and $f_\theta(\mathbf{x})$ an observable model signal, such as a representation, prediction score, generated response, or activation. Attackers reconstruct $\mathbf{x}$ either by directly optimizing a candidate (or a continuous surrogate) or by training an inversion model on auxiliary signal-text pairs. Some methods, such as Vec2Text [153], are hybrid: they train a corrector and then iteratively refine each candidate through repeated target-encoder queries at inference.

Optimization-based Approaches. Optimization-based attacks update a candidate text representation so that an observable signal matches the target. Public data may initialize or constrain the candidate, but the solution need not be an element of an auxiliary corpus. Because text is discrete, white-box attacks often optimize a continuous relaxation and then decode tokens. Writing $\mathrm{Emb}(\mathbf{x})$ for an exposed target embedding and $\mathcal{X}_{\text{relax}}$ for the relaxed candidate space, a representative objective is $\hat{\mathbf{x}}^* = \arg\min_{\hat{\mathbf{x}} \in \mathcal{X}_{\text{relax}}} \mathcal{L}(\mathrm{Emb}(\hat{\mathbf{x}}), \mathrm{Emb}(\mathbf{x}))$. Candidate fidelity depends on the relaxation, prior, and exposed signal; iterative refinement can also use black-box encoder queries when a learned corrector replaces target-model gradients and supplies the optimization signal in realistic black-box deployments.

Training-based Approaches. Training-based inversion treats the observable mapping as an encoder and learns a decoder $f_{\boldsymbol{\phi}}^{-1}$ from auxiliary signal-text pairs. After querying f_θ on $\hat{\mathbf{x}}_i \in D_{\text{auxiliary}}$, the attacker minimizes $\mathcal{L}_{f^{-1}}(\boldsymbol{\phi}) = -\frac{1}{N}\sum_{i=1}^{N} \log p_{\boldsymbol{\phi}}(\hat{\mathbf{x}}_i \mid f_\theta(\hat{\mathbf{x}}_i))$ and reconstructs $\hat{\mathbf{x}}^* = f_{\boldsymbol{\phi}}^{-1}(f_\theta(\mathbf{x}))$. Once collected, the pairs allow black-box operation, though transfer depends on how well the auxiliary data and interface match the victim setting.

4.2 Applying Text MIAs to Different Scenarios

MIAs on Language Models. Text MIAs exploit task-specific interfaces, which we organize into the three task settings of embedding, text classification, and language generation (Fig. 7). For embeddings, Song and Raghunathan [210] optimize a continuous relaxation, whereas Huang et al. [91], Li et al. [116], Song and Raghunathan [210] train inversion models; Vec2Text variants [25, 153] are hybrid, training a corrector but recovering text by iterative black-box re-embedding. For classification, Parikh et al. [171], Zhang et al. [283] optimize relaxed tokens or generator states. For generation, Morris

Table 4. Summary of model inversion attacks on text and multimodal interfaces.

Approach	Method (venue)	Interface / capability	Auxiliary resource	Distinctive mechanism	Recovered target	Reported limitation
Optimization-based	Continuous Relaxation [210] (CCS'20)	White-box embeddings	Embedding matrix, vocab.	Continuous relaxation of the embedding matrix	Input text from embeddings	Fidelity depends on relaxation/prior
	CEA [171] (ACL'22)	White-box classifier	Prefix, label, embeddings	Temperature-scaled suffix logits vs. target label	Memorized canary suffix	Canary-level recovery only
	Text Revealer [283] (arXiv'22)	White-box classifier	Same-domain corpus, GPT-2	PPLM-style perturbation of a generator	Label-representative templates	Template-level, not exact inputs
	DDMI [177] (OpenRev.'25)	White-box CLIP	Distilled diffusion model	Diffusion prior replaces GAN; cheaper sampling	Private training images	Needs white-box CLIP embeddings
	LeakyCLIP [27] (arXiv'25)	White-box CLIP	Stable Diffusion prior	Adversarial fine-tuning + diffusion refinement	CLIP training images	Needs white-box CLIP embeddings
	SMI-AW [164] (CVPR'26)	White-box VLM	Response-level signals	Adaptive token-/sequence-level weighting	Private visual training images	Needs white-box VLM responses
	Vec2Text [153] (EMNLP'23)	Black-box embeddings	Text–embedding pairs	Trained corrector + iterative re-embedding	Full input text	Per-example query/decoding cost
	MultiVec2Text [25] (ACL'24)	Black-box multilingual	Cross-lingual pairs	Vec2Text correction, multilingual	Full text, many languages	Per-example query/decoding cost
Training-based	CapRecover [254] (ACM MM'25)	Split-inference features	Feature–semantics pairs	Inversion model on client-side features	Semantic labels and captions	Semantics only, no image
	MLC & MSP [210] (CCS'20)	Black-box embeddings	Sentence-embedding pairs	Multi-label or multi-set word prediction	Input word sets	Lexical only; word order lost
	GEIA [116] (ACL'23)	Black-box embeddings	Aux. pairs + pre-trained LM	MLP-aligned prefix, teacher forcing	Sentence-level input text	Domain and model shift
	logit2prompt [154] (ICLR'24)	Black-box logits	Signal–text pairs	MLPs map logprob blocks to embeddings	Hidden prompt	Needs logprob access; query limits
	output2prompt [278] (EMNLP'24)	Black-box responses	Multiple outputs/prefix	Sparse encoder, max-likelihood prefix	Hidden system prompt	Response access; query limits
	Transferable EI [91] (ACL'24)	Released embeddings	Small leak + surrogate	Surrogate stealing + adversarial alignment	Input text from embeddings	Assumes small training-data leak
	PILS [159] (NeurIPS'25)	Black-box logprobs	Signal–text pairs	Additive log-ratio compression	Hidden prompt	Needs logprob access; query limits
	AIA [35] (ACL'25)	Training activations	Shadow activation data	Attack model on shadow activation pairs	Private LLM training text	Needs collaborative training
	Face-Adapter [200] (CVPRW'25)	Black-box embeddings	Face foundation model	Adapter into a foundation-model space	Identity-consistent face	Identity-level recovery only
	Inv^{2}A [267] (AAAI'26)	Black-box responses	Target LM as decoder	Inverse encoder, denoised representation	Hidden prompt	Prompt recovery only

et al. [154], Nazir et al. [159], Ye et al. [267], Zhang et al. [278] recover prompts from output text or next-token signals, while Dai et al. [35] reconstruct training text from activations observed during decentralized LLM training.

MIAs on Text Embedding. (Principle 1: [25, 153]; Principle 2: [210]; Principle 4: [91]; Principle 6: [116]) For embedding tasks, embedding models and text encoders map words or sentences to dense vectors supporting similarity search, retrieval, and clustering. Classic word-level examples include Word2Vec [150] and GloVe [180]; modern sentence encoders provide the interfaces attacked by full-text reconstruction methods. Because these vectors retain semantic and sometimes lexical information, releasing them creates an inversion channel for private text.

► **Continuous relaxation for gradient-based inversion.** Enumerating an auxiliary corpus is slow and resource-intensive once that corpus is large. Song and Raghunathan [210] therefore replace enumeration with white-box optimization of one continuous variable z_i per token position, making the discrete sentence search differentiable via $\text{relaxed}(\mathbf{Z}, T) = \mathbf{V}^\top \text{softmax}(\mathbf{Z}/T)$ for embedding matrix $\mathbf{V}^\top$ and temperature T.

► **Train an embedding-to-text inverse network.** With only output access, Song and Raghunathan [210] train $f_{\boldsymbol{\phi}}^{-1}$ on auxiliary sentence-embedding pairs to maximize $\log p(\mathcal{W}(\hat{\mathbf{x}}) \mid f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}))$. Multi-label classification predicts an unordered vocabulary-wide word set, whereas multi-set prediction selects elements sequentially from the remaining multiset. Both recover lexical content but not word order, which limits these classification-style inversions.

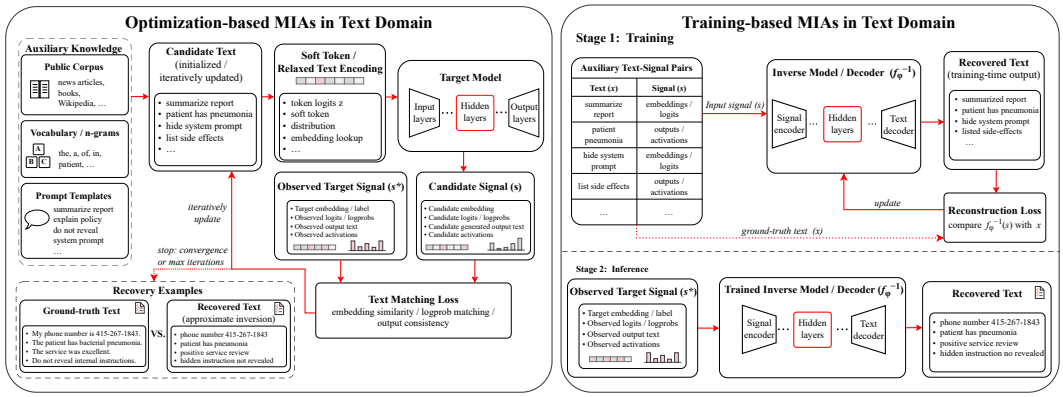


Fig. 7. Illustrations of model inversion attacks in the text domain.

► **MIA as a generation task.** GEIA [116] replaces the repetitive, semantically weak word sets of Song and Raghunathan [210] with autoregressive text. An MLP aligns the target embedding to the model's token representation, which is concatenated with the first word embedding; the pre-trained language model is then trained with teacher forcing [250] on auxiliary pairs, framing black-box inversion as *sequence generation*. At inference the decoder emits the first token from the aligned embedding and continues autoregressively on its own context.

► **Controlled generation and iterative refinement.** Vec2Text [153] generates, *re-embeds*, and iteratively corrects a text hypothesis, marginalizing over intermediate hypotheses after N iterations before beam search. Chen et al. [25] extend it to multilingual and cross-lingual inversion, finding multilingual models more susceptible.

► **Query-free embedding inversion.** Transferable EI [91] first uses a small assumed leak from D_{train} to steal a surrogate encoder, then treats that surrogate as a generator and aligns it adversarially with the target space through a discriminator, and trains a decoder to invert target embeddings *without any further queries*.

MIAs on Text Classification. (Principle 2: [171]; Principle 3: [283]) For text classification tasks, language models are fine-tuned to predict predefined classes or labels for applications such as sentiment analysis, intent classification (IC), and named-entity recognition (NER). Models such as BERT [41] and XLNet [263] commonly use a task-specific output layer that exposes class scores or probabilities, which provide the inversion signal in this setting.

► **Text completion on natural-language-understanding (NLU) models.** In white-box canary extraction (CEA), Parikh et al. [171] recover an unknown suffix given a known prefix, such as "My phone number is…". They assign each suffix position a temperature-scaled logit vector, map its softmax through the embedding matrix, and minimize the classification loss on the concatenation of prefix and relaxed suffix embeddings against the target label; the highest-logit tokens form the recovered canary.

► **Transformer-based text classification.** Text Revealer [283], the first MIA specifically targeting transformer-based text classifiers, extracts common templates by n-gram analysis of same-domain public data, fine-tunes GPT-2 as a domain generator, and then follows the PPLM idea [37] by optimizing the generator's *hidden-state perturbation* against the white-box target classifier for a chosen label, *e.g.,* a 4-star review. Thus GPT-2 supplies the text prior, while victim-classifier feedback steers reconstruction toward fluent text representative of the target label.

MIAs on Language Generation. (Principle 2: [35]; Principle 3: [154, 159]; Principle 6: [267, 278]) For language generation tasks, generative language models learn from large text corpora to predict the next-token distribution over

a vocabulary $\mathcal{V}$ and produce coherent sequences. Transformer architectures [233], including GPT-2 [186], expose generated text, next-token probabilities, and, during training, intermediate activations as potential inversion signals.

▶ **Reconstruct hidden prompts from logit vectors.** Observing that logit outputs carry rich information about the preceding text, Logit2prompt [154] maps the next-token probability vector into a sequence of pseudo-embeddings, using MLPs on blocks of log probabilities to bridge the vocabulary and encoder dimensions. A conditional inversion model then decodes these pseudo-embeddings into the hidden prompt from observed outputs.

▶ **Reconstruct hidden prompts from text outputs.** Output2prompt [278] needs only generated text: for a hidden prefix and varying user prompts, it sparsely encodes several distinct but semantically related outputs into one aggregated representation, then reconstructs the prefix by maximum-likelihood training at lower cost than a self-attention encoder, recovering prompts that closely match the true prefixes under this output-only interface in practice.

▶ **Reconstruct hidden prompts from compact next-token distribution representation.** Based on Morris et al. [154], Nazir et al. [159] propose Prompt Inversion from Logprob Sequences (PILS), a logprob-based attack that reconstructs hidden prompts from next-token probability sequences. PILS observes that $\mathbf{p} = \text{softmax}(W\mathbf{h})$ lies in a low-dimensional subspace determined by the hidden state $\mathbf{h}$. Its additive log-ratio transformation gives $\text{alr}(\mathbf{p})_D = A_D\mathbf{h}$, enabling lossless compression of each probability vector. An encoder-decoder inverter then maps the compact sequence back to the hidden prompt, reducing dimensionality without requiring internal activations. Recently, Inv^{2}A [267] studies prompt inversion from a complementary latent-space perspective. It formalizes an invariant latent space hypothesis, treating the language model as an invariant decoder and learning a lightweight inverse encoder that maps model outputs to a denoised pseudo-representation for prompt recovery. This shifts prompt inversion from compressing next-token distributions to exploiting semantic and cyclic invariance across model inputs and outputs.

▶ **Reconstruct training data from activations in decentralized LLM training.** Beyond hidden prompts, recent work targets the training corpus itself under decentralized training. Dai et al. [35] propose the activation inversion attack (AIA), which builds a shadow dataset of text-activation pairs from public data, trains an attack model, and reconstructs private training text from victim activations exchanged by decentralized LLMs.

4.3 MIAs on Vision-Language and Foundation Models

(Principle 2: [164, 254]; Principle 5: [27, 177]; Principle 6: [200]) Recent MIAs either target vision-language and multimodal models or use foundation models as reconstruction priors. We group them by the observable signal or prior: generated VLM responses, shared image-text embeddings, biometric embeddings decoded with foundation-model priors, and intermediate visual features in split VLM inference. Chen et al. [27], Nguyen et al. [164], Peng et al. [177] iteratively invert CLIP-style embeddings or VLM response signals, whereas Shahreza et al. [200], Xiu and Zhang [254] train adapters or inversion models for biometric embeddings and split-inference features. Under the corresponding threat model, the exposed interface determines whether recovery yields images, identities, labels, captions, or other textual semantics in practice. Across these interfaces, prompt, embedding, activation and generated-response inversion all reconstruct an input that the model processed, and therefore remain classical MIAs under Definition 2.2; training-data extraction differs in kind, because it elicits memorized strings by sampling the model rather than by inverting an observed signal, so we treat it as an adjacent problem rather than a form of model inversion.

▶ **Generated-response interfaces for visual reconstruction.** For generative VLMs exposing language responses rather than a class prediction, the response sequence itself supplies inversion supervision. SMI-AW [164] replaces fixed-logit maximization with token- and sequence-level signals reflecting visually grounded concepts, adaptively

weighting informative response tokens to better supervise reconstruction of private visual training images. VLM language interfaces thus leak visual training information even when no image classifier is exposed.

▶ **Shared image-text embedding interfaces.** The shared image-text space of CLIP-style models is itself an inversion interface. DDMI [177] replaces GAN priors with a distilled single-step diffusion generator, cutting the sampling cost of diffusion-prior inversion. LeakyCLIP [27] combines adversarial fine-tuning, embedding alignment, and Stable Diffusion refinement to extract CLIP training data, showing multimodal rather than text-only leakage.

▶ **Foundation-prior interfaces for biometric embeddings.** Face-recognition services often expose compact biometric embeddings rather than multimodal outputs, and a foundation model can supply the prior space needed for reconstruction. Face-Adapter [200] learns an adapter translating embeddings from a black-box face-recognition model into a face foundation model's embedding space, enabling identity-consistent face reconstruction. Foundation models thus serve not only as attack targets but as priors that make compressed biometric representations more invertible.

▶ **Split-VLM feature interfaces for semantic recovery.** In split-VLM inference, intermediate visual features travel from the client-side visual encoder to the cloud-side language component and can disclose high-level semantics. CapRecover [254] recovers semantic text such as labels and captions directly from these features, targeting semantic disclosure without reconstructing the input image. This complements prompt- and activation-based inversion by showing that textual semantics learned by multimodal foundation models are exposed at the representation level.

▶ **Comparative observations.** The exposed interface governs text-reconstruction feasibility: embeddings preserve semantics but may omit word order; logprob sequences and responses reveal prompts under visibility and query limits; activations and split features are richer but need training or collaborative access. Training-based decoders amortize recovery yet suffer domain and model shift, whereas iterative correction pays per-example query and decoding cost. Foundation models may be the target, the exposed interface, or the reconstruction prior—distinct roles.

5 Model Inversion Attacks on Graph Data

Graphs represent arbitrary relationships through non-grid structures, such as social or molecular interactions; Fig. 8 traces this line of work. Graph neural networks (GNNs) [63, 105, 106, 280] exploit these structures in applications such as recommendation [216] and drug discovery [92], but models trained on private graphs may leak sensitive information. Graph MIAs primarily target topology (edge connectivity), and link inference can expose private social ties or proprietary structures [76].

Graph MIAs face two challenges: (1) adjacency is discrete and combinatorial and graph size varies, so gradient optimization needs relaxation or projection; and (2) structural priors differ across molecular, social, and citation graphs.

5.1 A Summary of Model Inversion Approaches on Graph Data

Let $\boldsymbol{G} = (V, \boldsymbol{A})$ have node set V and adjacency $\boldsymbol{A} \in \{0,1\}^{N\times N}$, where $A_{uv} = 1$ iff u and v are connected. Define $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$, $\tilde{D}_{vv} = \sum_u \tilde{A}_{vu}$, and $\hat{\boldsymbol{A}} = \tilde{\boldsymbol{D}}^{-1/2}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-1/2}$. Let $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ be node features, $\boldsymbol{Y} \in \{0,1\}^{N\times C}$ one-hot labels, and $\boldsymbol{h}_v^{(\ell)}$ the layer-ℓ representation, with $\boldsymbol{h}_v^{(0)} = \boldsymbol{x}_v$. The adversary may know any subset $\mathcal{K} \subseteq \{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{h}_v^{(\ell)}\}$. Tab. 5 summarizes graph MIAs by paradigm, interface, auxiliary resource, mechanism, recovered target, and limitation.

We retain the original taxonomy's two families: optimization-based approaches [132, 167, 204, 293, 295] and training-based approaches [20, 76, 292].

Optimization-based Approaches. These attacks solve for an adjacency candidate $\hat{A}$ using an observable signal O—task loss, predictions, explanations, or embedding consistency—and graph priors such as sparsity or smoothness:

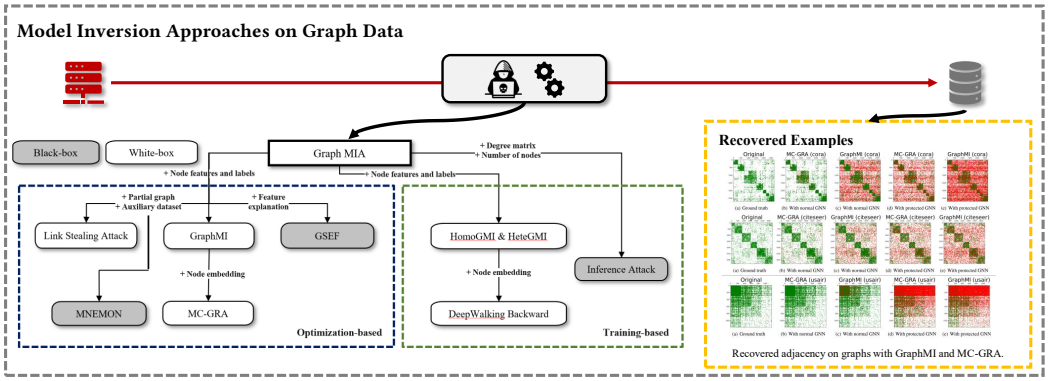


Fig. 8. Evolution of graph MIAs with representative recovered examples.

$\hat{A}^* = \arg\min_{\hat{A}} \mathcal{L}_{\text{obs}}\Big(\Phi(\hat{A}; f_{\theta^*}, \mathcal{K}), O\Big) + \lambda \mathcal{R}(\hat{A})$. Analytic inversion instead solves the embedding relation directly without training an auxiliary attack model.

Training-based Approaches. These methods train a classifier or decoder on auxiliary observation-target pairs, such as GNN outputs labeled by link existence or public graphs paired with embeddings, and apply it to victim outputs [76, 292]. They need representative attack-training data but neither target architecture nor parameters.

5.2 Applying Graph MIAs to Different Scenarios

Optimization via Graph-related Knowledge. (Principle 4: [76, 167]) The pioneering Link Stealing Attack [76] infers whether each pair of nodes in a GNN's training graph is connected. Let $s(\boldsymbol{\phi}(u), \boldsymbol{\phi}(v))$ denote an edge score derived from target outputs or learned by an attack model; the attack predicts $\hat{A}_{uv} = 1$ when $s \geq T$. A raw distance can instead be converted to a similarity or compared with the reverse inequality. Its prior knowledge $\mathcal{K}$ may include node features X_{node}, a partial target graph A_{sub}, or an auxiliary graph $D_{\text{auxiliary}}$. Across eight threat scenarios, the method constructs scores from prediction or entropy-vector distances (cosine, Euclidean, or correlation-based variants) and, when labeled links from a partial or auxiliary graph are available, trains an MLP link classifier.

► **Information leakage from feature explanation.** Olatunji et al. [167] quantify how much training-graph structure leaks when GNN feature explanations are released. They propose an explanation-only attack, which reconstructs the graph from the similarity of explanation vectors, and an explanation-augmentation attack, which additionally exploits node features and labels. Gradient-based explanations leak the most structure despite low utility, whereas for perturbation- and surrogate-based explanations leakage grows with explanation utility. As a defense, perturbing explanation bits with randomized responses reduces the attack to random guessing in their reported experiments.

Optimization via Model Intermediate Information. (Principle 2: [132, 295]; Principle 3: [132, 293, 295]; Principle 4: [204]) A trained GNN may leak private topology through several correlated variables. Markov Chain-based Graph Reconstruction Attack (MC-GRA) [295] *models the layered GNN computation as a chain* and reconstructs adjacency by aligning combinations of known features, labels, hidden representations, and outputs between the target and attack chains. Its information-theoretic analysis relates recovery to the information captured by these variables.

► **Node features and partial labels are available.** GraphMI [293], shown in Fig. 9, considers the white-box setting with $\mathcal{K} \subseteq \{X_{\text{node}}, Y_{\text{sub}}\}$. It uses projected gradient descent to handle discrete edges while regularizing topological

Table 5. Summary of graph model inversion attacks, grouped by paradigm and exposed interface.

Approach	Method (venue)	Interface / capability	Auxiliary resource	Distinctive mechanism	Recovered target	Reported limitation
Optimization-based	GraphMI [293] (IJCAI'21)	White-box gradients	Node features, labels	Projected gradient on relaxed adjacency + autoencoder	Full adjacency	Needs parameters + features and labels
	MC-GRA [295] (ICML'23)	White-box internals	Node features, labels, hidden representations	Layered GNN as a Markov chain; chain alignment	Full adjacency	Needs intermediate representations
	MNEMON [204] (CCS'22)	Released embeddings only	Node embeddings	Graph metric + self-supervised structure learning	Graph structure	Structure only; no features or model access
	GSEF [167] (PETS'23)	Released explanations	Node features, labels, feature explanations	Explanation-vector similarity (+ features, labels)	Edge existence	Varies by explanation type; randomized response defeats it
	HomoGMI / HeteGMI [132] (SecureComm'23)	White-box gradients	Node features, labels	Classification loss + 1st-/2nd-order proximity	Topology of homo-/heterogeneous graphs	Needs parameters; type schema for HeteGMI
Training-based	Link Stealing [76] (USENIX'21)	Black-box posteriors	Node features, partial target graph, auxiliary dataset	Link predictor on posterior pairs; similarity scoring	Edge existence	Needs representative attack-training data
	DeepWalking Backwards [20] (ICML'21)	Released embeddings	Degree matrix, node count	Linear solve or PPMI-error minimization	Adjacency, community structure	Individual edges, triangle density may be lost
	Inference Attack [292] (USENIX'22)	Black-box graph embeddings	Node features, labels, node embeddings	Decoders on public graph-embedding pairs	Graph size, density, subgraph, approximate topology	Graph-level only; defenses shown graphically

sparsity and feature smoothness: $\hat{A}^* = \arg\min_{\hat{A}} \mathcal{L}_{\text{attack}} = \mathcal{L}_{\text{cls}}(f_\theta, Y_{\text{sub}}) + \alpha\mathcal{L}_{\text{smo}} + \beta\mathcal{L}_{\text{spa}}$, where $\mathcal{L}_{\text{cls}}$ is the classification loss, $\mathcal{L}_{\text{smo}} = {}^1\!/_2 \sum_{i,j=1}^{N} \hat{A}_{i,j} \left\| \frac{\boldsymbol{x}_i}{\sqrt{d_i}} - \frac{\boldsymbol{x}_j}{\sqrt{d_j}} \right\|_2^2$ is the feature-smoothness loss, and $\mathcal{L}_{\text{spa}} = \|\hat{A}\|_F$ is the sparsity loss.

► **Graph MIAs via model-agnostic information.** MNEMON [204] introduces a *graph recovery attack that relies solely on released node embeddings*, combining graph metric learning, which learns a data-specific distance, with self-supervised structure learning through a graph autoencoder. Reconstruction requires only released embeddings, not the embedding model, node features, a partial target graph, or a separately collected auxiliary graph.

Whereas a homogeneous graph has one node and edge type, a heterogeneous graph has multiple types; this distinction is not the same as homophily, the tendency of adjacent nodes to share labels. Liu et al. [132] propose HomoGMI and HeteGMI for these two graph types. Both are gradient-based optimization attacks combining the target GNN's classification loss with first- and second-order proximity objectives for topology recovery.

Inverting Released Node or Graph Embeddings (Principle 4: [20, 292]). DeepWalking Backwards [20] reconstructs a graph from DeepWalk embeddings by *solving a linear system* or minimizing positive pointwise mutual-information matrix error. It can preserve community structure and downstream utility, although individual edges and properties such as triangle density may be lost. For pooled graph embeddings, Zhang et al. [292] infer graph size, density, subgraph containment, and approximate topology, showing that released representations can disclose global and fine-grained structure even when embeddings are released solely to support downstream classification tasks.

► **Comparative observations.** Signals imply different graph targets: prediction/loss attacks optimize topology under feature or label knowledge, explanations expose attribution patterns, and released embeddings can reveal edges or global properties without model access. Success depends on matching sparsity, homophily, types, and size; discrete projection and quadratic edge search limit scalability, while pooled embeddings may preserve global properties but lose individual edges. Federated-gradient reconstruction is instead adjacent signal sharing, not released-artifact inversion.

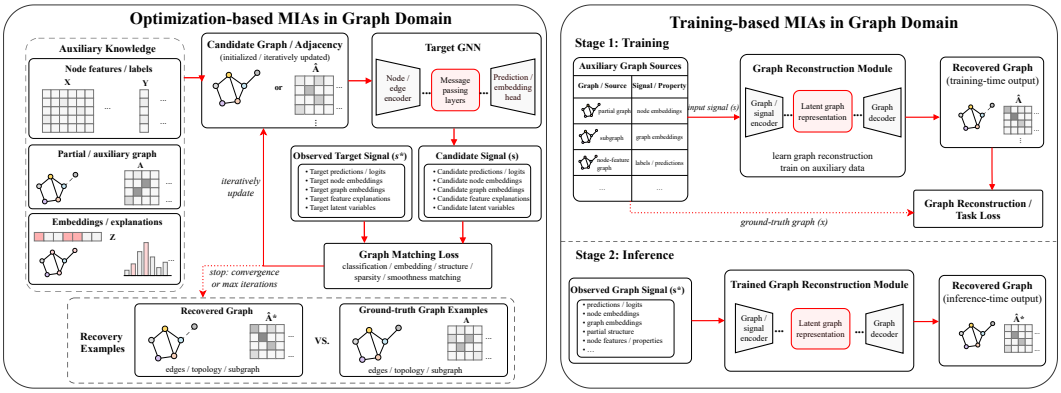


Fig. 9. Illustrations of model inversion attacks in the graph domain.

6 On Defending against Model Inversion Attacks

MIA countermeasures act during training or inference (Fig. 10). Training-time methods weaken input-output dependence by restricting information flow or perturbing training, whereas inference-time methods perturb released outputs and mainly address restricted-access attacks. We organize both classes by domain below for clarity.

6.1 Defense Techniques in Image Domain

Image defenses act during training through differential privacy, deception, label smoothing, information restriction, transfer learning, or data-centric processing, and during inference through output perturbation (Fig. 11).

Training-time Defense Approaches. (Principle 7: [56, 289]; Principle 8: [64, 81, 179, 215, 240]) *Differential privacy* (DP) adds calibrated noise to protect record participation [56] but does not conceal the overall training distribution, and empirically may not stop distribution-level inversion without substantial utility loss [56, 289]. Wang et al. [240] formalize this record-versus-distribution limitation through an indistinguishability game.

► **Deception-based defenses.** GAN-ID [64] trains the target with GAN-generated confounding samples so that inversion yields misleading reconstructions. Its weighted objective combines a private-sample loss that confuses inversion, a public-sample loss that associates fakes with the protected label, and a continual-learning loss that preserves the primary task. GAN-ID needs public data and shadow attacks, protects one class, and risks unintended behavior.

► **Exploring label smoothing with MI robustness.** Negative Label Smoothing (NegLS) [215] uses a negative smoothing factor α in a label-smoothed cross-entropy objective. Positive smoothing can increase leakage when training data are scarce, whereas negative smoothing maintains high confidence away from decision boundaries and steers inversion optima away from private samples. Nevertheless, attacks such as PLG-MI still challenge these defenses [215].

► **Architecture-level and feature-compression defenses.** Beyond the loss or the confidence scores, a recent line of defenses targets the model architecture and its intermediate features. In the architectures evaluated by Koh et al. [107], skip connections reinforce MI attacks; the authors also propose MI-resilient designs that outperform prior defenses in robustness. LoFt [274] instead constrains the rank of the intermediate features, on the finding that higher-rank features leak more, shrinking the information an attacker can recover while preserving utility. Together they show architectural and representation-level design to be an effective but underexplored axis of MI defense.

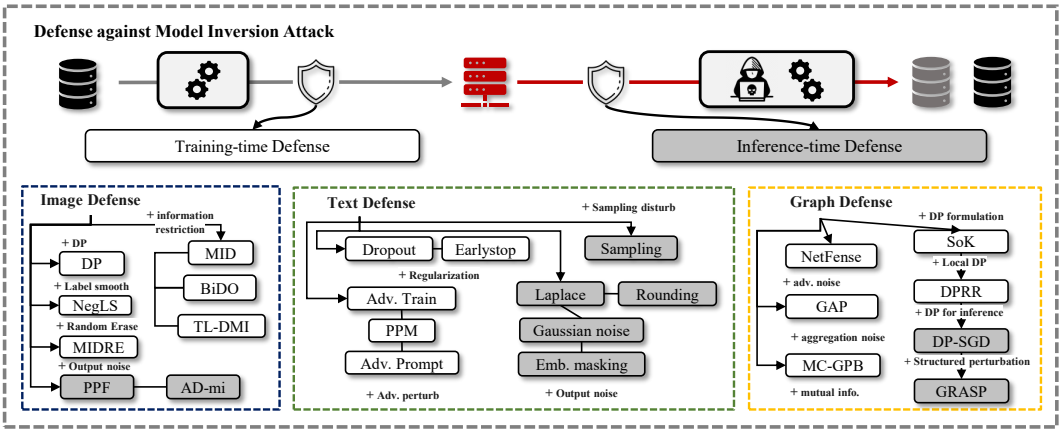


Fig. 10. A taxonomy of defense methods against model inversion attacks across different domains.

► **Information restriction-based defenses.** Dependency regularization limits information flow into learned representations. MID [240] penalizes *mutual information* between inputs and outputs, which can conflict with classification. BiDO [179] instead minimizes input-feature dependence while maximizing feature-label dependence, using the Hilbert–Schmidt independence criterion (HSIC) or constrained covariance measures at the cost of additional hyperparameter tuning.

► **Transfer learning-based defense.** TL-DMI [81] uses Fisher information to identify layers that encode private data; it pre-trains on public data, *freezes early layers*, and fine-tunes only later layers for the private task.

► **Data-centric defenses via random erasing.** MIDRE [231] *randomly erases image regions* during training, reducing reconstruction quality and attack accuracy while retaining natural accuracy. It complements model-centric defenses such as BiDO [179], NegLS [215], and TL-DMI [81], whose combination lowers attack accuracy further.

Inference-time Defense Approaches. (Principle 9: [264]; Principle 10: [248]) The prediction purification framework (PPF) [264] uses an autoencoder to reduce the dispersion of predictions before release, making confidence vectors less input-sensitive, and trains the purifier adversarially against a surrogate inversion attack for inference-time use.

► **Adversarial output noise.** Because PPF may reduce utility and retain facial cues, Wen et al. [248] instead use inversion-model gradients to perturb predictions, maximizing reconstruction error while preserving target accuracy. A *label modifier* retains the original predicted label, yielding zero accuracy loss in their experiments.

6.2 Defense Techniques in Text Domain

In the text domain, training-time defenses use overfitting prevention, representation changes, and adversarial learning, whereas inference-time defenses use output obfuscation, embedding masking, or modified decoding and sampling (Fig. 12).

Training-time Defense Approaches. (Principle 7: [169, 210]; Principle 8: [18, 171]) Since memorization drives text leakage [18, 19, 171, 210], training-time defenses combine regularization, early stopping, and adversarial learning.

► **Overfitting-prevention and representation defenses.** Dropout suppresses activations during training, and early stopping halts training once validation performance plateaus [9, 171]. However, dropout did not significantly reduce unintended memorization in the generative sequence-model experiments of Carlini et al. [18]. Parikh et al. [171]

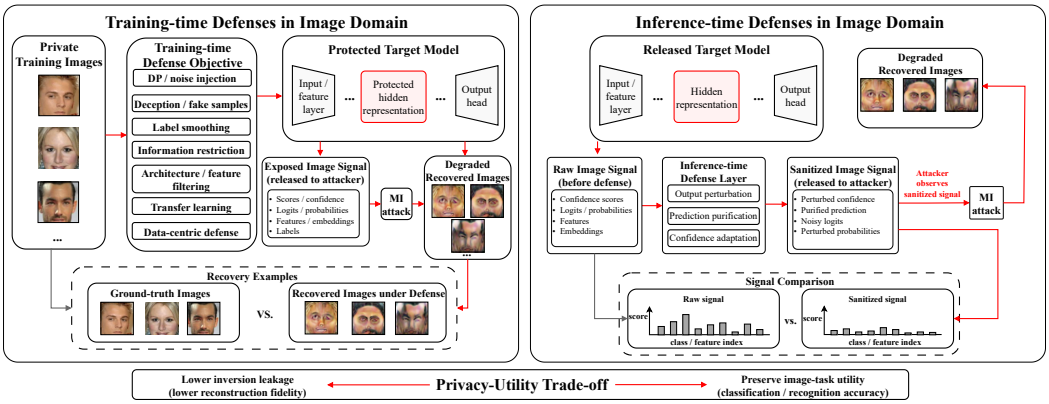


Fig. 11. Illustrations of image-domain model inversion defenses.

additionally concatenate character-CNN representations with token embeddings, changing the NLU model's input representation to make canary reconstruction harder only in their evaluated NLU setting.

► **Adversarial training for privacy-preserving representation learning.** Adversarial representation learning jointly trains an encoder against a word or attribute predictor [32, 50, 51, 121, 210, 253]. A privacy-preserving mapping can similarly distort embeddings within a bounded range [169, 193]. For hidden prompts, explicit non-disclosure instructions [278, 288] can reduce empirical leakage but offer a behavioral rather than formal privacy guarantee.

Inference-time Defense Approaches. (Principle 9: [25]; Principle 10: [25, 153, 154, 169]) Inference-time defenses obscure released outputs without retraining the target, using data obfuscation, embedding masking, or changes to decoding and sampling procedures at the interface exposed to attackers during black-box inference.

► **Data obfuscation-based defense.** *Output obfuscation* adds Laplace [49, 169] or Gaussian noise [25, 153], or rounds embedding coordinates [169]. The multilingual masking defense instead replaces the first embedding coordinate with a language identifier [25]. Each mechanism removes or redistributes fine-grained information at a distinct utility cost. Eguard [133] maps LLM embeddings through a transformer projector trained to reduce text mutual information, thereby limiting recovery while preserving semantic utility for downstream tasks under its reported protocol.

► **Sampling-based defenses.** Morris et al. [154] reduce prompt-inversion signals by *increasing sampling temperature* or restricting decoding to top-p or top-K candidates. These transformations flatten or truncate the next-token distribution used for generation, trading output quality and diversity for lower inversion performance.

6.3 Defense Techniques in Graph Domain

Defenses against graph MIAs mainly protect *private link information* rather than node features. Training-time methods perturb graph processing or restrict topology-dependent information, whereas inference-time methods perturb embeddings or explanations exposed to users (Fig. 13). Khosla [104] notes that transparency may expose sensitive structure.

Training-time Defense Approaches. (Principle 8: [80, 84, 156, 192, 295]) NetFense [84] is a graph-preprocessing defense rather than a differential-privacy mechanism. It selects and optimizes inconspicuous edge perturbations that preserve graph and model utility while reducing an attacker's confidence in private-label inference.

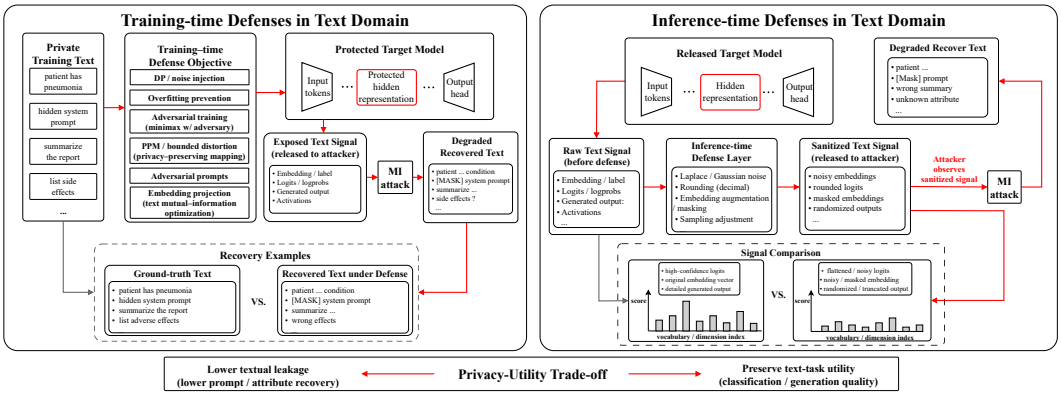


Fig. 12. Illustrations of text-domain model inversion defenses.

► **Employing differential privacy.** Graph DP can protect edges, nodes, or whole graphs; the SoK by Mueller et al. [157] systematizes these neighboring-dataset definitions but is a survey, not itself a defense algorithm. For multi-graph classification, Mueller et al. [156] apply DP-SGD during training by clipping per-example gradients and adding Gaussian noise before parameter updates; their GNNExplainer analysis compares private and non-private representations rather than adding noise to final predictions. Degree-Preserving Randomized Response (DPRR) [80] proposes a *local DP* algorithm designed to protect edges in GNNs. DPRR combines randomized response [247] with *strategic edge sampling* to preserve degree information under edge DP. For graph classification over unattributed social graphs, DPRR provides edge DP while approximately preserving degree information. It outperforms randomized response, local Laplacian noise, and non-private-user-only baselines, especially when some users are non-private in its reported experiments.

► **Aggregation perturbation.** GAP [192] injects *independent Gaussian noise* into node-level aggregation and accounts for its privacy loss, yielding edge- or node-level DP for multi-hop GNNs. This is stochastic perturbation, not adversarial training. GAP outperforms node-classification baselines while quantifying its privacy-utility trade-off.

► **Employing information restriction during training.** This line of work aims to minimize the model's reliance on private links during training. Markov Chain-based Graph Privacy Bottleneck (MC-GPB) [295] prevents leakage by minimizing mutual information between the *node embedding matrix* and the graph's *adjacency matrix*. Its Graph Information Plane tracks leakage and suggests that GNNs first memorize and later forget private information.

Inference-time Defense Approaches. (Principle 10: [67, 292]) When graph embeddings or intermediate aggregations must be exposed, perturbing that interface can reduce reconstruction. Zhang et al. [292] evaluate *output perturbation* against property, subgraph, and topology inference from graph embeddings. GRASP [67] uses a Bernoulli selector to apply shared or independent Gaussian noise per node embedding. This shifts pairwise similarities in both directions, disrupting reconstruction rankings while retaining edge-level DP during training and inference.

6.4 Privacy–Utility Trade-off of Defenses

Every defense trades exposed information against accuracy, fidelity, computation, or applicability. The controlling variable may be a privacy budget or noise scale, a regularization weight, a smoothing or erasing factor, frozen layers, or a sampling parameter; costs can also include public data, adversarial training, tuning, or extra inference. Since studies use different attacks, datasets, metrics, and privacy notions, universal Low/Medium/High ratings would be misleading.

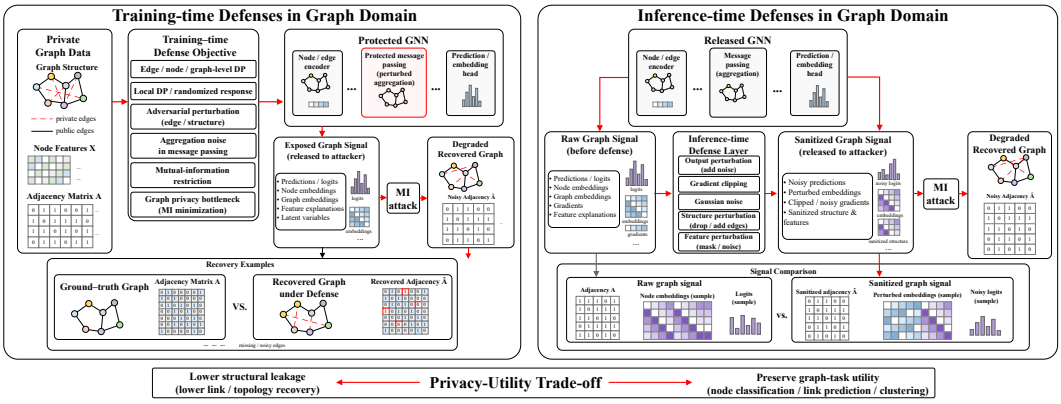


Fig. 13. Illustrations of graph-domain model inversion defenses.

Tab. 6 therefore reports paper-specific privacy and downstream-utility evidence where the original studies provide it, together with each method's control and applicability boundary in its original setting and threat model.

The trade-off is interface- and domain-dependent: output perturbation does not protect gradients or manipulated split-learning representations; text defenses may preserve classification while degrading embedding or generation quality; and graph defenses must identify the protected object. The reported values above are illustrative rather than a ranking because datasets, attacks, evaluators, and privacy strengths differ. Robustness should therefore be reported with adaptive attacks, task utility, resource cost, and the exact privacy notion under the stated protocol.

7 Datasets and Evaluation Metrics

MIBench [184] provides a standardized, modular benchmark and toolbox integrating representative attacks and defenses under common protocols. It standardizes the data-preparation, attack, defense, and evaluation stages so that reported numbers share matched target models and auxiliary data, which removes much of the setup variance that otherwise makes published numbers incomparable. Its scope, however, is the image domain rather than the three modalities covered here, so cross-domain comparison still rests on the per-domain metrics below.

7.1 Datasets

Tab. 7 lists datasets across the three domains with their scale, structure, and typical role in MI studies, supporting reproducibility and controlled comparisons among inversion attacks under common threat models.

7.2 Evaluation Metrics

We group metrics by fidelity, identity/semantics, distributional realism, privacy risk/attack success, and utility, as reported by the studies surveyed in Secs. 3–5 and behind the privacy–utility evidence in Tab. 6. Because no single metric establishes inversion, reports should combine target-level fidelity or semantics with attack success and utility.

General evaluation metrics. For true positive (TP), true negative (TN), false positive (FP), and false negative (FN), *accuracy (Acc.)* is $(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$, *recall (R)* is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, *precision (P)* is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, and *F1 score (F1)* is their harmonic mean. Reports must name the evaluator and positive event, e.g. identity, private-token, or private-link recovery.

Table 6. Privacy–utility evidence for representative defenses. Reported values are paper-specific and not directly comparable; Train and Inf. denote training- and inference-time defenses; Arch. and Repr. refine the training-time class.

Defense	Dom./ stage	Control	Reported privacy / utility evidence	Cost / boundary
GAN-ID; MID/BiDO [64, 179, 240]	Image/Train	loss/dependence weights	BiDO-HSIC, CelebA/KEDMI: attack acc. > 29% lower at ∼ 6% classifier acc.	public/shadow data; tuning cost
NegLS; MIDRE [215, 231]	Image/Train	smoothing/erasing	NegLS ($\alpha = -0.05$), FaceScrub/PPA: top-1 attack acc. 80.0% lower at ∼ 3.4% test acc.	calibration or input modification; gains shrink at low resolution
TL-DMI [81]	Image/Train	frozen layers	CelebA/KEDMI: top-1 attack acc. 39% lower at ∼ 5.6% classifier acc.	public pre-training; classification only
PPF / adv. noise [248, 264]	Image/Inf.	contraction/noise	FaceScrub inversion error: purification 0.0114 → 0.0448 at < 0.4% acc. cost	attack-dependent; the noise gain collapses to 0.0271 under an adaptive attacker
LoFt [274]	Image/Repr.	feature-rank constraint	FaceScrub/IR-152: IF top-1 attack acc. 82.2 → 13.0% at test acc. 98.2 → 97.1%	dimension-dependent: attack acc. back to 77.8% at dimension 512; image classification only
Skip-connection design [107]	Image/Arch.	last-stage skip connections	FaceScrub/PPA, ResNet-101: removing stage-4 skips 83.0 → 6.8%; deployable RoLSS 83.0 → 58.7% at acc. 94.9 → 92.4%	requires retraining; architecture-specific; evaluated on face recognition
Regularization [18, 210]	Text/Train	dropout/stopping/loss	dropout/weight decay/early stopping leave canary exposure unchanged; DP-SGD: 31.0 → 2.8 for +8% test loss	no formal guarantee without DP; adversarial-training utility cost only graphical
Eguard [133]	Text/Inf.	projection/MI loss	SST2/NLI/retrieval/summarization: inversion F1/recall ∼ 4%, > 98% of accuracy/ROUGE retained	extra projector; 1.6–3.4× one-time training overhead
Obfuscation/sampling [25, 154, 169]	Text/Inf.	noise/masking/sampling	language-ID masking, BEIR: BLEU 61.6 → 4.7 with NDCG@10 unchanged	far weaker on multilingual encoders; sampling defenses report no utility metric
NetFense [84]	Graph/Train	edge budget	CiteSeer: private-label acc. 84.0 → 68.5% at target-label acc. 98.0 → 84.5%	measured on the attacked node subset, not the full test set
DPRR / GAP [80, 192]	Graph/Train	privacy budget/noise	GAP-NDP, Facebook: node-MI AUC 50.2–52.7% at $\epsilon = 1$–16; node acc. 63.2% at $\epsilon = 8$	formal node DP; privacy strength and dataset affect utility
MC-GPB [295]	Graph/Train	privacy weight	Cora: non-learnable GRA AUC 76.6 → 70.6%; node acc. down 3.0% rel.	estimation cost; on USA the same setting costs 16.8% accuracy
Embedding perturbation / GRASP [67, 292]	Graph/Inf.	noise/selector	GRASP (GCN, $\sigma = 1.0$), Cora: GRA AUC 95.8 → 74.8% at node acc. 87.8 → 87.5%	GRASP reports σ, not ϵ; Zhang et al. report defense curves only graphically

Image-specific evaluation metrics. Pixel, feature, and identity fidelity. *Mean squared error (MSE)* is squared pixel error; *peak signal-to-noise ratio (PSNR)* is its logarithmic ratio; and *pixelwise similarity* is 1 − MSE after $[0, 1]$ normalization. *Feature distance* is L_2 distance between pixel vectors or, more commonly, evaluation-model embeddings; *K-nearest-neighbor distance* averages distances to the K nearest references. Identity-evaluator Acc. measures recognizability, not pixel fidelity. *Fréchet Inception Distance (FID)* [79] compares feature distributions, not target recovery. *Learned Perceptual Image Patch Similarity (LPIPS)* [284] is deep-feature distance; *Structural Similarity Index Measure (SSIM)* [246] measures image structure; *Natural Image Quality Evaluator (NIQE)* [151] measures no-reference naturalness; and the *Contrastive Language–Image Pre-training score (CLIP Score)* is embedding cosine similarity. Distance metrics are minimized, similarity metrics maximized.

Text-specific evaluation metrics. Lexical overlap and likelihood. *Bilingual Evaluation Understudy (BLEU)* [170] combines modified n-gram precision and a brevity penalty; *Recall-Oriented Understudy for Gisting Evaluation (ROUGE)* [126], including ROUGE-N, measures reference-content recall. *Perplexity (PPL)* is exponentiated negative autoregressive log-likelihood, and *pseudo-log-likelihood (PLL)* is its masked-language-model analogue; both measure fluency, not privacy recovery. *Named Entity Recovery Ratio (NERR), Stop Word Ratio* [116], and *Recovery Rate (RR)* [283] quantify recovered content. *Hamming distance per token* [69] measures token mismatch, *cosine similarity (CS)* compares representations, and *large-language-model evaluation (LLM-Eval)* [128] assesses appropriateness, content, grammar, and relevance. Since paraphrases can have low overlap, lexical metrics should be paired with content-recovery and semantic measures.

Table 7. Cross-domain datasets used in model inversion studies, grouped by data role, with scale ranges for canonical releases and family shading for Image, Text, and Graph. Individual studies may use subsets or preprocessing variants.

Family	Datasets (sources)	Scale / structure	Typical MIA use
face / biometric	CelebA [138]; CelebA-HQ [99]; FaceScrub [160]; LFW [89]; FFHQ [101]; VGGFace2 [17]; VGGFace [172]; CASIA-WebFace [268]; MOBIO [146]; AgeDB [155]; AT&T Face [194]; MetFaces [100]; PubFig83 [182]; iCV MEFED [140]	400–3.31M images; up to 10,575 identities	Identity, attribute, and facial-expression reconstruction; face-embedding inversion
general classification	MNIST [113]; QMNIST [258]; Fashion-MNIST [252]; CIFAR-10 [109]; ImageNet1K [191]; ImageNet [38]; TinyImageNet [38]; Omniglot [112]	32.5K–14.2M images; 10–21,841 classes	Class-level reconstruction under white-box, score-based, or label-only access
fine-grained recognition	Stanford Dogs [103]; AFHQ Dogs [28]; Oxford IIIT Pet [173]; Amur Tigers [119]; Stanford Cars [108]; CUB-200-2011 [235]	7.3K–20.6K images; 37–200 classes	Fine-grained class or identity reconstruction with learned image priors
medical	ChestX-ray8 [242]; ChestX-ray14 [242]; CheXpert [93]	108K–224K images; 8/14 multilabel findings	Reconstruction of sensitive diagnostic images and attributes
vision-language	LAION-400M [198]; LAION-2B [197]; Furniture Object [139]; Flickr30K [273]; COCO [127]; Flickr8K [82]	166 images to 2.32B image-text pairs	CLIP/VLM image reconstruction and semantic recovery
intent / classification	ATIS [34]; NLU-Evaluation [135]; Snips [33]; Emotion Dataset [195]; Yelp [287]; IMDB [142]; AG News [287]	4,978–700K utterances or documents	Intent, label, sentence, and embedding reconstruction
language modeling	WikiText [148]; BookCorpus [299]; Wikipedia [210]; MTG [26]; ArXiv [35]; OpenWebText [35]; The Pile [35]; ArXiv Math [267]	Large corpora spanning 100M tokens to millions of documents	Training-text memorization and activation inversion
QA / retrieval	QNLI [236]; Natural Questions [111]; MSMARCO [11]; BEIR [228]; CLIRMatrix [218]	18-dataset benchmark; up to 8.8M passages	Embedding inversion and recovery of queries, passages, or answers
dialogue / preference	PersonaChat [285]; Helpfulness & Harmlessness [10]; ShareGPT [227]	8,939–433,723 conversations or preference pairs	Dialogue, response, and preference-data reconstruction
prompts / instructions	Unnatural Instructions [83]; Instructions-2M [154]; Alpaca [226]; Synthetic GPTs [278]; Real GPTs [278]; Awesome-ChatGPT-Prompts [6]; Alpaca Code [21]; Instruction corpora [267]	153–2.33M prompts or instructions	Hidden-prompt, system-prompt, and instruction-data inversion
clinical	MIMIC-III [96]	2.08M clinical notes	Recovery of sensitive medical text
citation networks	Cora [105]; CoraML [199]; CiteSeer [105]; PubMed [105]	One graph; 2.5K–19.7K nodes and 5.3K–44.3K edges	Adjacency, feature, and link reconstruction
social / interaction	Polblogs [2]; USA and Brazil [188]; Actor [175]; Facebook [115]; BlogCatalog [183]; E-mail and YouTube [114]; Bitcoin [110]	One graph; 131–1.13M nodes and 2.1K–2.99M edges	Topology recovery, link inference, and graph-embedding inversion
heterogeneous	ACM [241]; DBLP [225]; IMDB [57]	One graph; 9K–18.4K nodes and 13K–67.9K edges	Recovery of typed relations and private links
molecular / biological	AIDS [189]; DD [44]; ENZYMES and PROTEINS_full [15]; PPI [183]; NCI1, OVCAR-8H, PC3, and MOLT-4H [152]; COX2 and DHFR [221]	One graph or 467–4,110 graphs; approximately 16–284 nodes per graph	Graph-level property, structure, and attribute reconstruction
web / coauthor	Wiki-CS [149]; Amazon Computers and Amazon Photos [202]; Coauthor CS and Coauthor Physics [202]	One graph; 7.5K–34.5K nodes and 81.9K–248K edges	Large-scale link, feature, and neighborhood reconstruction

Graph-specific evaluation metrics. Link and structure metrics. *Area under the receiver operating characteristic curve (AUC)* ranks recovered links across thresholds; *average precision (AP)* summarizes the precision–recall curve and is more informative under class imbalance, with an edge as the positive event. *Joint degree distribution* gives endpoint degrees, and *relative Frobenius error* $\|A-\hat{A}\|_F/\|A\|_F$ compares adjacency or *positive pointwise mutual information (PPMI)* matrices. Agreement between true and reconstructed graphs is also reported as relative graph-property error, cosine

Table 8. Deployment interfaces, operational constraints, and residual reconstruction risks. TEE denotes a trusted execution environment.

Environment	Interface	Attacks	Constraint	Controls	Residual risk
Cloud/ MLaaS [212]	labels, scores, text, embeddings	output/embedding inversion	rate/API limits, truncation, monitoring	restrict/perturb outputs; monitor abuse	adaptive low-query attacks; utility loss
Edge/on-device [220]	parameters, embeddings, runtime	white-box/representation inversion	device compute; protected execution	TEE/model protection; minimize representations	extracted parameters; exposed boundary
Collaborative/split [78]	smashed data, activations, gradients	passive/active reconstruction	cut, auxiliary pairs, deviation, detection	protect representations; check active servers	architecture/task dependence; priors
Federated/ decentralized [86]	per-round gradients/activations	gradient/activation reconstruction	rounds, batch, aggregation, participation	DP; secure aggregation; validation	utility loss; malicious server; raw messages

similarity, Wasserstein or Jensen–Shannon distance, or the Weisfeiler–Lehman (WL) kernel. Aggregate similarity does not prove exact sensitive-edge recovery, so link-level and structural metrics should be paired.

Comparability and evaluator bias. Evaluator false positives, miscalibration, and domain shift can overstate recovery. Results are not comparable across evaluators, preprocessing, targets, attacks, splits, or privacy budgets. A defensible protocol pairs identity Acc. with target-level and distributional image quality; RR/NERR with overlap, semantics, and fluency for text; or AUC/AP with structural similarity for graphs, while reporting utility and cost.

8 Further Discussions on Future Directions

Problem Settings and Advanced Techniques. Open problems include imperfect side information, generative, reinforcement, and federated learning, and fine-grained attribute recovery. Optimization and generative priors require evaluation under domain knowledge, query limits, efficiency, and legitimate utility in deployment.

Deployment Environments. Cloud, edge, split, and federated systems expose outputs, parameters/state, features, and gradients/activations. Tab. 8 maps them to constraints, controls, and residual risk: MLaaS output perturbation does not protect split features, and secure aggregation does not protect released embeddings in practice.

Cost and Query-Budget Constraints. Optimization-based attacks pay per target, whereas training-based attacks amortize auxiliary-data and training costs. Neither is uniformly costlier. Reports should include target count, iterations/samples, surrogate training, hardware/API cost, rate limits, wall-clock time, GPU-hours, queries, reconstruction count, and quality for reproducibility (*e.g.,* SMILE [125], BREP-MI [97], and LOKT [162]). Detectability is a further constraint. High-volume query campaigns are visible at the API boundary as rate and input-distribution anomalies; an active split-inference server deviates from the protocol in auditable ways; and assembling an auxiliary corpus may itself be costly or traceable. Whether a defender can flag inversion queries without hurting legitimate users remains open.

9 Conclusion

This survey has organized model inversion attacks and defenses around the four questions that decide whether reconstruction succeeds: which interface the adversary observes, what knowledge and prior it may assume, what it can recover, and under what conditions it fails. Across the image, text, and graph domains, reconstruction quality follows the strength of the generative prior and the richness of the exposed interface, and the interfaces worth defending now reach beyond released classifiers to split features, shared embeddings, and vision-language responses. Defenses have not kept pace, and each trades exposed information against accuracy, fidelity, or computation. Progress will depend on improving that trade-off, and on evaluation that reflects the deployment and query budgets an attacker actually faces.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. 2023. Gpt-4 technical report. arXiv:2303.08774. URL

[2] L. Adamic and N. Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In LinkKDD. URL

[3] D. Adiwardana, M. Luong, D. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, et al. 2020. Towards a human-like open-domain chatbot. arXiv:2001.09977. URL

[4] C. Akcora, B. Carminati, and E. Ferrari. 2012. Privacy in social networks: How risky is your social graph?. In ICDE. URL

[5] N. Akhtar, A. Mian, N. Kardan, and M. Shah. 2021. Advances in adversarial attacks and defenses in computer vision: A survey. IEEE Access (2021). URL

[6] F. Akın. 2022. Awesome-ChatGPT-Prompts. GitHub repository. URL

[7] S. An, G. Tao, Q. Xu, Y. Liu, G. Shen, Y. Yao, J. Xu, and X. Zhang. 2022. Mirror: Model inversion for deep learning network with high fidelity. In NDSS. URL

[8] M. Arjovsky, S. Chintala, and L. Bottou. 2017. Wasserstein generative adversarial networks. In ICML. URL

[9] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al. 2017. A closer look at memorization in deep networks. In ICML. URL

[10] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862. URL

[11] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. arXiv:1611.09268. URL

[12] M. Balunovic, D. Dimitrov, N. Jovanović, and M. Vechev. 2022. Lamp: Extracting text from gradients with language model priors. In NeurIPS. URL

[13] M. Bertran, S. Tang, A. Roth, M. Kearns, J. Morgenstern, and S. Wu. 2023. Scalable membership inference attacks via quantile regression. In NeurIPS. URL

[14] D. Bhandari, C. Murthy, and S. Pal. 1996. Genetic algorithm with elitist model and its convergence. IJPRAI (1996). URL

[15] K. Borgwardt, C. Ong, S. Schönauer, S. Vishwanathan, A. Smola, and H. Kriegel. 2005. Protein function prediction via graph kernels. Bioinformatics (2005). URL

[16] T. Brown, B. Mann, N. Ryder, et al. 2020. Language models are few-shot learners. In NeurIPS. URL

[17] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In IEEE FG. URL

[18] N. Carlini, C. Liu, U. Erlingsson, J. Kos, and D. Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In USENIX Security. URL

[19] N. Carlini, F. Tramer, E. Wallace, et al. 2021. Extracting training data from large language models. In USENIX Security. URL

[20] S. Chanpuriya, C. Musco, K. Sotiropoulos, and C. Tsourakakis. 2021. Deepwalking backwards: from embeddings back to graphs. In ICML. URL

[21] S. Chaudhary. 2023. Code Alpaca: An Instruction-following LLaMA model for code generation. GitHub repository. URL

[22] K. Chen, S. Guo, T. Zhang, X. Xie, and Y. Liu. 2021. Stealing deep reinforcement learning models for fun and profit. In ASIA-CCS. URL

[23] L. Chen, L. Wu, R. Hong, K. Zhang, and M. Wang. 2020. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. In AAAI. URL

[24] S. Chen, M. Kahla, R. Jia, and G. Qi. 2021. Knowledge-enriched distributional model inversion attacks. In ICCV. URL

[25] Y. Chen, H. Lent, and J. Bjerva. 2024. Text Embedding Inversion Security for Multilingual Language Models. In ACL. URL

[26] Y. Chen, Z. Song, X. Wu, D. Wang, J. Xu, J. Chen, H. Zhou, and L. Li. 2022. MTG: A benchmark suite for multilingual text generation. In NAACL. URL

[27] Y. Chen, S. Wang, X. Wang, R. He, X. Ma, and Y. Jiang. 2025. LeakyCLIP: Extracting Training Data from CLIP. arXiv:2508.00756. URL

[28] Y. Choi, Y. Uh, J. Yoo, and J. Ha. 2020. Stargan v2: Diverse image synthesis for multiple domains. In CVPR. URL

[29] C. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot. 2021. Label-only membership inference attacks. In ICML. URL

[30] A. Chowdhery, S. Narang, J. Devlin, et al. 2023. Palm: Scaling language modeling with pathways. JMLR (2023). URL

[31] C. Clifton, M. Kantarcıoğlu, J. Vaidya, X. Lin, and M. Zhu. 2002. Tools for privacy preserving distributed data mining. SIGKDD Explor. (2002). URL

[32] M. Coavoux, S. Narayan, and S. Cohen. 2018. Privacy-preserving neural representations of text. In ACL. URL

[33] A. Coucke, A. Saade, A. Ball, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. arXiv:1805.10190. URL

[34] D. Dahl, M. Bates, M. Brown, et al. 1994. Expanding the scope of the ATIS task: The ATIS-3 corpus. In HLT. URL

[35] C. Dai, L. Lu, and P. Zhou. 2025. Stealing Training Data from Large Language Models in Decentralized Training through Activation Inversion Attack. In ACL. URL

[36] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. 2017. Deformable convolutional networks. In ICCV. URL

[37] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu. 2019. Plug and play language models: A simple approach to controlled text generation. In ICLR. URL

[38] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. 2009. Imagenet: A large-scale hierarchical image database. In CVPR. URL

[39] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In CVPR. URL

[40] J. Deng, Y. Wang, J. Li, S. Wang, C. Shang, H. Liu, S. Rajasekaran, and C. Ding. 2021. Tag: Gradient attack on transformer-based language models. In EMNLP. URL

[41] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL. URL

[42] S. Dibbo. 2023. Sok: Model inversion attack landscape: Taxonomy, challenges, and future roadmap. In IEEE CSF. URL

[43] P. Dissanayake and S. Dutta. 2024. Model reconstruction using counterfactual explanations: A perspective from polytope theory. In NeurIPS. URL

[44] P. Dobson and A. Doig. 2003. Distinguishing enzyme structures from non-enzymes without alignments. J. Mol. Biol. (2003). URL

[45] X. Dong, H. Yin, J. Alvarez, J. Kautz, P. Molchanov, and H. Kung. 2022. Privacy Vulnerability of Split Computing to Data-Free Model Inversion Attacks. In BMVC. URL

[46] A. Dosovitskiy and T. Brox. 2016. Generating Images with Perceptual Similarity Metrics based on Deep Networks. In NeurIPS. URL

[47] A. Dosovitskiy and T. Brox. 2016. Inverting Visual Representations With Convolutional Networks. In CVPR. URL

[48] V. Duddu, D. Samanta, D. Rao, and V. Balas. 2018. Stealing neural networks via timing side channels. arXiv:1812.11720. URL

[49] C. Dwork, A. Roth, et al. 2014. The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. (2014). URL

[50] H. Edwards and A. Storkey. 2016. Censoring representations with an adversary. In ICLR. URL

[51] Y. Elazar and Y. Goldberg. 2018. Adversarial removal of demographic attributes from text data. In ACL. URL

[52] E. Erdoğan, A. Küpçü, and A. Çiçek. 2022. UnSplit: Data-Oblivious Model Inversion, Model Stealing, and Label Inference Attacks against Split Learning. In WPES. URL

[53] H. Fang, B. Chen, X. Wang, Z. Wang, and S. Xia. 2023. Gifd: A generative gradient inversion method with feature domain optimization. In ICCV. URL

[54] H. Fang, Y. Qiu, H. Yu, W. Yu, J. Kong, B. Chong, B. Chen, X. Wang, S. Xia, and K. Xu. 2024. Privacy leakage on dnns: A survey of model inversion attacks and defenses. arXiv:2402.04013. URL

[55] M. Fredrikson, S. Jha, and T. Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In CCS. URL

[56] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. 2014. Privacy in pharmacogenetics: An End-to-End case study of personalized warfarin dosing. In USENIX Security. URL

[57] X. Fu, J. Zhang, Z. Meng, and I. King. 2020. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In WWW. URL

[58] X. Gao and L. Zhang. 2023. PCAT: Functionality and Data Stealing from Split Learning by Pseudo-Client Attack. In USENIX Security. URL

[59] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller. 2020. Inverting gradients-how easy is it to break privacy in federated learning?. In NeurIPS. URL

[60] Gemini Team et al. 2023. Gemini: a family of highly capable multimodal models. arXiv:2312.11805. URL

[61] J. Geng, Y. Mou, Q. Li, F. Li, O. Beyan, S. Decker, and C. Rong. 2024. Improved gradient inversion attacks and defenses in federated learning. IEEE Trans. Big Data (2024). URL

[62] A. Ghiasi, H. Kazemi, S. Reich, C. Zhu, M. Goldblum, and T. Goldstein. 2022. Plug-in inversion: Model-agnostic inversion for vision with data augmentations. In ICML (Proceedings of Machine Learning Research). PMLR. URL

[63] J. Gilmer, S. Schoenholz, P. Riley, O. Vinyals, and G. Dahl. 2017. Neural Message Passing for Quantum Chemistry. In ICML. URL

[64] X. Gong, Z. Wang, S. Li, Y. Chen, and Q. Wang. 2023. A gan-based defense framework against model inversion attacks. IEEE TIFS (2023). URL

[65] Y. Gong, Y. Fang, and Y. Guo. 2015. Privacy-preserving collaborative learning for mobile health monitoring. In GLOBECOM. URL

[66] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. In NeurIPS. URL

[67] Z. Guo, Y. Liu, X. Ao, and Q. He. 2025. Grasp: Differentially private graph reconstruction defense with structured perturbation. In SIGKDD. URL

[68] W. Hamilton, Z. Ying, and J. Leskovec. 2017. Inductive representation learning on large graphs. In NeurIPS. URL

[69] R. Hamming. 1950. Error detecting and error correcting codes. Bell Syst. Tech. J. (1950). URL

[70] G. Han, J. Choi, H. Lee, and J. Kim. 2023. Reinforcement learning-based black-box model inversion attacks. In CVPR. URL

[71] A. Hatamizadeh, H. Yin, H. Roth, W. Li, J. Kautz, D. Xu, and P. Molchanov. 2022. Gradvit: Gradient inversion of vision transformers. In CVPR. URL

[72] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro. 2019. Logan: Membership inference attacks against generative models. In PoPETs. URL

[73] K. He, G. Gkioxari, P. Dollár, and R. Girshick. 2017. Mask r-cnn. In ICCV. URL

[74] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In CVPR. URL

[75] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In SIGIR. URL

[76] X. He, J. Jia, M. Backes, N. Gong, and Y. Zhang. 2021. Stealing links from graph neural networks. In USENIX Security. URL

[77] X. He, R. Wen, Y. Wu, M. Backes, Y. Shen, and Y. Zhang. 2021. Node-level membership inference attacks against graph neural networks. arXiv:2102.05429. URL

[78] Z. He, T. Zhang, and R. Lee. 2019. Model Inversion Attacks against Collaborative Inference. In ACSAC. URL

[79] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In NeurIPS. URL

[80] S. Hidano and T. Murakami. 2022. Degree-Preserving Randomized Response for Graph Neural Networks under Local Differential Privacy. arXiv:2202.10209. URL

[81] S. Ho, J. Koh, K. Chandrasegaran, N. Nguyen, and N. Cheung. 2024. Model Inversion Robustness: Can Transfer Learning Help?. In CVPR. URL

[82] M. Hodosh, P. Young, and J. Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. JAIR (2013). URL

[83] O. Honovich, T. Scialom, O. Levy, and T. Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In ACL. URL

[84] I. Hsieh and C. Li. 2023. Netfense: Adversarial defenses against privacy attacks on neural networks for graph data. IEEE TKDE (2023). URL

[85] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. Yu, and X. Zhang. 2022. Membership inference attacks on machine learning: A survey. ACM CSUR (2022). URL

[86] R. Hu, Y. Guo, H. Li, Q. Pei, and Y. Gong. 2020. Personalized federated learning with differential privacy. IEEE IoT J. (2020). URL

[87] X. Hu, L. Liang, S. Li, L. Deng, P. Zuo, Y. Ji, X. Xie, Y. Ding, C. Liu, T. Sherwood, et al. 2020. Deepsniffer: A dnn model extraction framework based on learning architectural hints. In ASPLOS. URL

[88] G. Huang, Z. Liu, L. Van Der Maaten, and K. Weinberger. 2017. Densely connected convolutional networks. In CVPR. URL

[89] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. 2007. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Tech. Rep., University of Massachusetts Amherst (2007). URL

[90] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora. 2021. Evaluating gradient inversion attacks and defenses in federated learning. In NeurIPS. URL

[91] Y. Huang, Y. Tsai, H. Hsiao, H. Lin, and S. Lin. 2024. Transferable Embedding Inversion Attack: Uncovering Privacy Risks in Text Embeddings without Model Queries. In ACL. Association for Computational Linguistics. URL

[92] V. Ioannidis, D. Zheng, and G. Karypis. 2020. Few-shot link prediction via graph neural networks for covid-19 drug-repurposing. In ICML Workshop. URL

[93] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In AAAI. URL

[94] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot. 2020. High accuracy and high fidelity extraction of neural networks. In USENIX Security. URL

[95] J. Jeon, J. Kim, K. Lee, S. Oh, and J. Ok. 2021. Gradient inversion with generative image prior. In NeurIPS. URL

[96] A. Johnson, T. Pollard, L. Shen, L. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. Mark. 2016. MIMIC-III, a freely accessible critical care database. Sci. Data (2016). URL

[97] M. Kahla, S. Chen, H. Just, and R. Jia. 2022. Label-only model inversion attacks via boundary repulsion. In CVPR. URL

[98] S. Kariyappa, A. Prakash, and M. Qureshi. 2021. Maze: Data-free model stealing attack using zeroth-order gradient estimation. In CVPR. URL

[99] T. Karras, T. Aila, S. Laine, and J. Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In ICLR. URL

[100] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. 2020. Training generative adversarial networks with limited data. In NeurIPS. URL

[101] T. Karras, S. Laine, and T. Aila. 2019. A style-based generator architecture for generative adversarial networks. In CVPR. URL

[102] Y. Kaya and T. Dumitras. 2021. When does data augmentation help with membership inference attacks?. In ICML. URL

[103] A. Khosla, N. Jayadevaprakash, B. Yao, and F. Li. 2011. Novel Dataset for Fine-Grained Image Categorization: Stanford Dogs. In CVPR Workshops. URL

[104] M. Khosla. 2022. Privacy and transparency in graph machine learning: A unified perspective. arXiv:2207.10896. URL

[105] T. Kipf and M. Welling. 2016. Semi-supervised classification with graph convolutional networks. In ICLR. URL

[106] T. Kipf and M. Welling. 2016. Variational graph auto-encoders. In NeurIPS. URL

[107] J. Koh, S. Ho, N. Nguyen, and N. Cheung. 2024. On the Vulnerability of Skip Connections to Model Inversion Attacks. In ECCV. URL

[108] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 2013. 3D Object Representations for Fine-Grained Categorization. In ICCV Workshops. URL

[109] A. Krizhevsky. 2009. Learning multiple layers of features from tiny images. Technical Report. University of Toronto. URL

[110] S. Kumar, F. Spezzano, V. Subrahmanian, and C. Faloutsos. 2016. Edge weight prediction in weighted signed networks. In ICDM. URL

[111] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. 2019. Natural questions: a benchmark for question answering research. TACL (2019). URL

[112] B. Lake, R. Salakhutdinov, and J. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. Science (2015). URL

[113] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. Proc. IEEE (1998). URL

[114] J. Leskovec and A. Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. Online dataset collection. URL

[115] J. Leskovec and J. Mcauley. 2012. Learning to discover social circles in ego networks. In NeurIPS. URL

[116] H. Li, M. Xu, and Y. Song. 2023. Sentence Embedding Leaks More Information than You Expect: Generative Embedding Inversion Attack to Recover the Whole Sentence. In ACL. URL

[117] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan. 2016. A persona-based neural conversation model. In ACL. URL

[118] O. Li, Y. Hao, Z. Wang, B. Zhu, S. Wang, Z. Zhang, and F. Feng. 2024. Model Inversion Attacks Through Target-Specific Conditional Diffusion Models. arXiv:2407.11424. URL

[119] S. Li, J. Li, H. Tang, R. Qian, and W. Lin. 2020. ATRW: a benchmark for Amur tiger re-identification in the wild. In ACM Multimedia. URL

[120] X. Li, Z. Zhou, J. Yao, Y. Rong, L. Zhang, and B. Han. 2024. Neural atoms: propagating long-range interaction in molecular graphs through efficient communication channel. In ICLR. URL

[121] Y. Li, T. Baldwin, and T. Cohn. 2018. Towards robust and privacy-preserving text representations. In ACL. URL

[122] Z. Li, M. Jiang, S. Wang, and S. Zhang. 2022. Deep learning methods for molecular representation and property prediction. Drug Discov. Today (2022). URL

[123] Z. Li, L. Wang, G. Chen, M. Shafq, et al. 2022. A survey of image gradient inversion against federated learning. TechRxiv preprint (2022). URL

[124] Z. Li, M. Yang, Y. Liu, J. Wang, H. Hu, W. Yi, and X. Xu. 2023. GAN You See Me? Enhanced Data Reconstruction Attacks against Split Inference. In NeurIPS. URL

[125] Z. Li, H. Zhang, J. Wang, M. Chen, H. Hu, W. Yi, X. Xu, M. Yang, and C. Ma. 2025. From Head to Tail: Efficient Black-box Model Inversion Attack via Long-tailed Learning. In CVPR. URL

[126] C. Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out. URL

[127] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. 2014. Microsoft COCO: Common Objects in Context. In ECCV. URL

[128] Y. Lin and Y. Chen. 2023. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. In NLP4ConvAI. URL

[129] D. Liu, C. Xu, J. Li, A. Nassereldine, and J. Xiong. 2025. Ensembler: Protect Collaborative Inference Privacy from Model Inversion Attack via Selective Ensemble. In DAC. URL

[130] L. Liu, Y. Wang, G. Liu, K. Peng, and C. Wang. 2023. Membership inference attacks against machine learning models via prediction sensitivity. IEEE TDSC (2023). URL

[131] R. Liu, D. Wang, Y. Ren, Z. Wang, K. Guo, Q. Qin, and X. Liu. 2024. Unstoppable Attack: Label-Only Model Inversion via Conditional Diffusion Model. IEEE TIFS (2024). URL

[132] R. Liu, W. Zhou, J. Zhang, X. Liu, P. Si, and H. Li. 2024. Model Inversion Attacks on Homogeneous and Heterogeneous Graph Neural Networks. In SecureComm. URL

[133] T. Liu, H. Yao, F. Lin, T. Wu, Z. Qin, and K. Ren. 2026. Eguard: Defending LLM Embeddings Against Inversion Attacks via Text Mutual Information Optimization. In AAAI. URL

[134] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In CVPR. URL

[135] X. Liu, A. Eshghi, P. Swietojanski, and V. Rieser. 2021. Benchmarking natural language understanding services for building conversational agents. In IWSDS. URL

[136] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692. URL

[137] Y. Liu, Z. Zhao, M. Backes, and Y. Zhang. 2022. Membership inference attacks by exploiting loss trajectory. In CCS. URL

[138] Z. Liu, P. Luo, X. Wang, and X. Tang. 2015. Deep learning face attributes in the wild. In ICCV. URL

[139] A. Lohia. 2024. sample_furniture_object. Hugging Face dataset. URL

[140] C. Loob, P. Rasti, I. Lüsi, J. Jacques, X. Baró, S. Escalera, T. Sapinski, D. Kaminska, and G. Anbarjafari. 2017. Dominant and complementary multi-emotional facial expression recognition using c-support vector classification. In FG. URL

[141] H. Ma, Y. Bian, Y. Rong, W. Huang, T. Xu, W. Xie, G. Ye, and J. Huang. 2022. Cross-dependent graph neural networks for molecular property prediction. Bioinformatics (2022). URL

[142] A. Maas, R. Daly, P. Pham, D. Huang, A. Ng, and C. Potts. 2011. Learning Word Vectors for Sentiment Analysis. In ACL. URL

[143] A. Mahendran and A. Vedaldi. 2015. Understanding Deep Image Representations by Inverting Them. In CVPR. URL

[144] T. Matsumoto, T. Miura, and N. Yanai. 2023. Membership inference attacks against diffusion models. In SPW. URL

[145] J. Mattern, F. Mireshghallah, Z. Jin, B. Schölkopf, M. Sachan, and T. Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In ACL. URL

[146] C. McCool, R. Wallace, M. McLaren, L. El Shafey, and S. Marcel. 2013. Session Variability Modelling for Face Authentication. IET Biometrics (2013). URL

[147] S. Mehnaz, S. Dibbo, E. Kabir, N. Li, and E. Bertino. 2022. Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In USENIX Security. URL

[148] S. Merity, C. Xiong, J. Bradbury, and R. Socher. 2017. Pointer sentinel mixture models. In ICLR. URL

[149] P. Mernyei and C. Cangea. 2020. Wiki-cs: A wikipedia-based benchmark for graph neural networks. In ICML Workshop. URL

[150] T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In ICLR. URL

[151] A. Mittal, R. Soundararajan, and A. Bovik. 2013. Making a Completely Blind Image Quality Analyzer. IEEE SPL (2013). URL

[152] C. Morris, N. Kriege, F. Bause, K. Kersting, P. Mutzel, and M. Neumann. 2020. Tudataset: A collection of benchmark datasets for learning with graphs. In ICML. URL

[153] J. Morris, V. Kuleshov, V. Shmatikov, and A. Rush. 2023. Text embeddings reveal (almost) as much as text. In EMNLP. URL

[154] J. Morris, W. Zhao, J. Chiu, V. Shmatikov, and A. Rush. 2024. Language Model Inversion. In ICLR. URL

[155] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. 2017. AgeDB: The First Manually Collected, In-the-Wild Age Database. In CVPRW. URL

[156] T. Mueller, J. Paetzold, C. Prabhakar, D. Usynin, D. Rueckert, and G. Kaissis. 2023. Differentially Private Graph Neural Networks for Whole-Graph Classification. IEEE TPAMI (2023). URL
[157] T. Mueller, D. Usynin, J. Paetzold, D. Rueckert, and G. Kaissis. 2022. SoK: Differential privacy on graph-structured data. arXiv:2203.09205. URL
[158] C. Nash, N. Kushman, and C. Williams. 2019. Inverting Supervised Representations with Autoregressive Neural Density Models. In AISTATS. URL
[159] M. Nazir, M. Finlayson, J. Morris, X. Ren, and S. Swayamdipta. 2025. Better Language Model Inversion by Compactly Representing Next-Token Distributions. In NeurIPS. URL
[160] H. Ng and S. Winkler. 2014. A data-driven approach to cleaning large face datasets. In ICIP. URL
[161] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In NeurIPS. URL
[162] B. Nguyen, K. Chandrasegaran, M. Abdollahzadeh, and N. Cheung. 2023. Label-Only Model Inversion Attacks via Knowledge Transfer. In NeurIPS. URL
[163] N. Nguyen, K. Chandrasegaran, M. Abdollahzadeh, and N. Cheung. 2023. Re-thinking model inversion attacks against deep neural networks. In CVPR. URL
[164] N. Nguyen, S. Ho, J. Koh, and N. Cheung. 2026. Do Vision-Language Models Leak What They Learn? Adaptive Token-Weighted Model Inversion Attacks. In CVPR. URL
[165] S. Oh, B. Schiele, and M. Fritz. 2019. Towards Reverse-Engineering Black-Box Neural Networks. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer. URL
[166] I. Olatunji, W. Nejdl, and M. Khosla. 2021. Membership inference attack on graph neural networks. In TPS-ISA. URL
[167] I. Olatunji, M. Rathee, T. Funke, and M. Khosla. 2023. Private graph extraction via feature explanations. PETS (2023). URL
[168] T. Orekondy, B. Schiele, and M. Fritz. 2019. Knockoff nets: Stealing functionality of black-box models. In CVPR. URL
[169] X. Pan, M. Zhang, S. Ji, and M. Yang. 2020. Privacy risks of general-purpose language models. In SP. URL
[170] K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In ACL. URL
[171] R. Parikh, C. Dupuy, and R. Gupta. 2022. Canary extraction in natural language understanding models. In ACL. URL
[172] O. Parkhi, A. Vedaldi, and A. Zisserman. 2015. Deep face recognition. In BMVC. URL
[173] O. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar. 2012. Cats and dogs. In CVPR. URL
[174] D. Pasquini, G. Ateniese, and M. Bernaschi. 2021. Unleashing the Tiger: Inference Attacks on Split Learning. In CCS. URL
[175] H. Pei, B. Wei, K. Chang, Y. Lei, and B. Yang. 2020. Geom-gcn: Geometric graph convolutional networks. In ICLR. URL
[176] X. Peng, B. Han, F. Liu, T. Liu, and M. Zhou. 2024. Pseudo-private data guided model inversion attacks. In NeurIPS. URL
[177] X. Peng, B. Han, F. Yu, F. Liu, T. Liu, and M. Zhou. 2025. Single-Step Diffusion Model-Based Generative Model Inversion Attacks. OpenReview preprint (withdrawn ICLR 2025 submission). URL
[178] X. Peng, B. Han, F. Yu, T. Liu, F. Liu, and M. Zhou. 2025. Generative Model Inversion Through the Lens of the Manifold Hypothesis. In NeurIPS. URL
[179] X. Peng, F. Liu, J. Zhang, L. Lan, J. Ye, T. Liu, and B. Han. 2022. Bilateral dependency optimization: Defending against model-inversion attacks. In KDD. URL
[180] J. Pennington, R. Socher, and C. Manning. 2014. Glove: Global vectors for word representation. In EMNLP. URL
[181] I. Petrov, D. Dimitrov, M. Baader, M. Müller, and M. Vechev. 2024. DAGER: Exact Gradient Inversion for Large Language Models. In NeurIPS. URL
[182] N. Pinto, Z. Stone, T. Zickler, and D. Cox. 2011. Scaling Up Biologically-Inspired Computer Vision: A Case Study in Unconstrained Face Recognition on Facebook. In CVPR Workshops. URL
[183] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang. 2018. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In WSDM. URL
[184] Y. Qiu, H. Yu, H. Fang, T. Zhuang, W. Yu, B. Chen, X. Wang, S. Xia, and K. Xu. 2024. MIBench: A Comprehensive Framework for Benchmarking Model Inversion Attack and Defense. arXiv:2410.05159. URL
[185] A. Radford, L. Metz, and S. Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In ICLR. URL
[186] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI blog (2019). URL
[187] M. Rahman, T. Rahman, R. Laganière, N. Mohammed, and Y. Wang. 2018. Membership inference attack against differentially private deep learning model. Trans. Data Priv. (2018). URL
[188] L. Ribeiro, P. Saverese, and D. Figueiredo. 2017. struc2vec: Learning node representations from structural identity. In KDD. URL
[189] K. Riesen and H. Bunke. 2008. IAM graph database repository for graph based pattern recognition and machine learning. In S+SSPR. URL
[190] M. Rigaki and S. Garcia. 2023. A survey of privacy attacks in machine learning. ACM CSUR (2023). URL
[191] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. IJCV (2015). URL
[192] S. Sajadmanesh, A. Shamsabadi, A. Bellet, and D. Gatica-Perez. 2023. Gap: Differentially private graph neural networks with aggregation perturbation. In USENIX Security. URL

[193] S. Salamatian, A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft. 2015. Managing your private and public data: Bringing down inference attacks against your privacy. IEEE JSTSP (2015). URL

[194] F. Samaria and A. Harter. 1994. Parameterisation of a stochastic model for human face identification. In WACV. URL

[195] E. Saravia, H. Liu, Y. Huang, J. Wu, and Y. Chen. 2018. Carer: Contextualized affect representations for emotion recognition. In EMNLP. URL

[196] F. Schroff, D. Kalenichenko, and J. Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In CVPR. URL

[197] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, et al. 2022. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. In NeurIPS D&B. URL

[198] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In NeurIPS DCAI Workshop. URL

[199] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. 2008. Collective classification in network data. AI Mag. (2008). URL

[200] H. Shahreza, A. George, and S. Marcel. 2025. Face Reconstruction from Face Embeddings using Adapter to a Face Foundation Model. In CVPR Workshops. URL

[201] Z. Shan, K. Ren, M. Blanton, and C. Wang. 2018. Practical secure computation outsourcing: A survey. ACM CSUR (2018). URL

[202] O. Shchur, M. Mumme, A. Bojchevski, and S. Günnemann. 2018. Pitfalls of graph neural network evaluation. In NeurIPS Workshop. URL

[203] V. Shejwalkar, H. Inan, A. Houmansadr, and R. Sim. 2021. Membership inference attacks against nlp classification models. In NeurIPS PriML Workshop. URL

[204] Y. Shen, Y. Han, Z. Zhang, M. Chen, T. Yu, M. Backes, Y. Zhang, and G. Stringhini. 2022. Finding mnemon: Reviving memories of node embeddings. In CCS. URL

[205] Y. Shen, X. He, Y. Han, and Y. Zhang. 2022. Model stealing attacks against inductive graph neural networks. In SP. URL

[206] Y. Shi, O. Kotevska, V. Reshniak, A. Singh, and R. Raskar. 2024. Dealing Doubt: Unveiling Threat Models in Gradient Inversion Attacks under Federated Learning, A Survey and Taxonomy. arXiv:2405.10376. URL

[207] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. 2017. Membership inference attacks against machine learning models. In SP. URL

[208] K. Simonyan, A. Vedaldi, and A. Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In ICLR Workshop. URL

[209] A. Singh, V. Sharma, R. Sukumaran, J. Mose, J. Chiu, J. Yu, and R. Raskar. 2024. SIMBA: Split Inference – Mechanisms, Benchmarks and Attacks. In ECCV. URL

[210] C. Song and A. Raghunathan. 2020. Information leakage in embedding models. In CCS. URL

[211] J. Song and D. Namiot. 2023. A Survey of Model Inversion Attacks and Countermeasures. Proc. Inst. Syst. Anal. RAS (2023). URL

[212] J. Song and D. Namiot. 2023. A survey of the implementations of model inversion attacks. In DCCN. URL

[213] L. Song, R. Shokri, and P. Mittal. 2019. Membership inference attacks against adversarially robust deep learning models. In SPW. URL

[214] L. Struppek, D. Hintersdorf, A. De Almeida Correia, A. Adler, and K. Kersting. 2022. Plug and Play Attacks: Towards Robust and Flexible Model Inversion Attacks. In ICML. URL

[215] L. Struppek, D. Hintersdorf, and K. Kersting. 2024. Be Careful What You Smooth For: Label Smoothing Can Be a Privacy Shield but Also a Catalyst for Model Inversion Attacks. In ICLR. URL

[216] J. Sun, Y. Zhang, W. Guo, H. Guo, R. Tang, X. He, C. Ma, and M. Coates. 2020. Neighbor interaction aware graph convolution networks for recommendation. In SIGIR. URL

[217] J. Sun, Y. Zhang, C. Ma, M. Coates, H. Guo, R. Tang, and X. He. 2019. Multi-graph convolution collaborative filtering. In ICDM. URL

[218] S. Sun and K. Duh. 2020. CLIRMatrix: A Massively Large Collection of Bilingual and Multilingual Datasets for Cross-Lingual Information Retrieval. In EMNLP. URL

[219] Y. Sun, D. Liang, X. Wang, and X. Tang. 2015. Deepid3: Face recognition with very deep neural networks. arXiv:1502.00873. URL

[220] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou. 2020. MobileBERT: A Compact Task-Agnostic BERT for Resource-Limited Devices. In ACL. URL

[221] J. Sutherland, L. O'Brien, and D. Weaver. 2003. Spline-fitting with a genetic algorithm: A method for developing classification structure- activity relationships. J. Chem. Inf. Comput. Sci. (2003). URL

[222] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. 2014. Intriguing properties of neural networks. In ICLR. URL

[223] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In CVPR. URL

[224] M. Tan and Q. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In ICML. URL

[225] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. 2008. Arnetminer: extraction and mining of academic social networks. In KDD. URL

[226] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. GitHub repository. URL

[227] S. Tey and D. Eccleston. 2024. ShareGPT. Online dataset. URL

[228] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In NeurIPS D&B. URL

[229] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. 2022. Lamda: Language models for dialog applications. arXiv:2201.08239. URL

[230] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288. URL
[231] V. Tran, N. Nguyen, S. Mai, H. Vandierendonck, I. Assent, A. Kot, and N. Cheung. 2025. Random Erasing vs. Model Inversion: A Promising Defense or a False Hope? TMLR (2025). URL
[232] B. van Breugel, H. Sun, Z. Qian, and M. van der Schaar. 2023. Membership inference attacks against synthetic data through overfitting detection. In AISTATS. URL
[233] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In NeurIPS. URL
[234] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. 2018. Graph attention networks. In ICLR. URL
[235] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep., California Institute of Technology CNS-TR-2011-001 (2011). URL
[236] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In ACL. URL
[237] H. Wang, S. Wang, C. Lu, and I. Echizen. 2026. DiffMI: Breaking Face Recognition Privacy via Diffusion-Driven Training-Free Model Inversion. IEEE TIFS (2026). URL
[238] K. Wang, Y. Fu, K. Li, A. Khisti, R. Zemel, and A. Makhzani. 2021. Variational model inversion attacks. In NeurIPS. URL
[239] P. Wang, Y. Li, K. Singh, J. Lu, and N. Vasconcelos. 2021. IMAGINE: Image Synthesis by Image-Guided Model Inversion. In CVPR. URL
[240] T. Wang, Y. Zhang, and R. Jia. 2021. Improving robustness to model inversion attacks via mutual information regularization. In AAAI. URL
[241] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. Yu. 2019. Heterogeneous graph attention network. In WWW. URL
[242] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In CVPR. URL
[243] Y. Wang, L. Huang, P. Yu, and L. Sun. 2021. Membership inference attacks on knowledge graphs. arXiv:2104.08273. URL
[244] Y. Wang, J. Li, H. Liu, Y. Wang, Y. Wu, F. Huang, and R. Ji. 2022. Black-box dissector: Towards erasing-based hard-label model stealing attack. In ECCV. URL
[245] Y. Wang and X. Lin. 2022. Enhance model stealing attack via label refining. In ICSP. URL
[246] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. IEEE TIP (2004). URL
[247] S. Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. JASA (1965). URL
[248] J. Wen, S. Yiu, and L. Hui. 2021. Defending against model inversion attack by adversarial examples. In CSR. URL
[249] O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel, and T. Langer. 2020. A compact review of molecular property prediction with graph neural networks. Drug Discov. Today Technol. (2020). URL
[250] R. Williams and D. Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. Neural Comput. (1989). URL
[251] R. Wu, X. Chen, C. Guo, and K. Weinberger. 2023. Learning to invert: Simple adaptive attacks for gradient inversion in federated learning. In UAI. URL
[252] H. Xiao, K. Rasul, and R. Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747. URL
[253] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig. 2017. Controllable invariance through adversarial feature learning. In NeurIPS. URL
[254] K. Xiu and S. Zhang. 2025. CapRecover: A Cross-Modality Feature Inversion Attack Framework on Vision-Language Models. In ACM MM. URL
[255] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. 2019. How powerful are graph neural networks?. In ICLR. URL
[256] K. Xu, H. Yue, L. Guo, Y. Guo, and Y. Fang. 2015. Privacy-preserving machine learning algorithms for big data systems. In ICDCS. URL
[257] X. Xu, M. Yang, W. Yi, Z. Li, J. Wang, H. Hu, Y. Zhuang, and Y. Liu. 2024. A Stealthy Wrongdoer: Feature-Oriented Reconstruction Attack against Split Learning. In CVPR. URL
[258] C. Yadav and L. Bottou. 2019. Cold case: The lost mnist digits. In NeurIPS. URL
[259] A. Yan, R. Hou, X. Liu, H. Yan, T. Huang, and X. Wang. 2022. Towards explainable model extraction attacks. Int. J. Intell. Syst. (2022). URL
[260] M. Yan, C. Fletcher, and J. Torrellas. 2020. Cache telepathy: Leveraging shared resource attacks to learn DNN architectures. In USENIX Security. URL
[261] M. Yang, Z. Li, J. Wang, H. Hu, A. Ren, X. Xu, and W. Yi. 2022. Measuring Data Reconstruction Defenses in Collaborative Inference Systems. In NeurIPS. URL
[262] W. Yang, S. Wang, D. Wu, T. Cai, Y. Zhu, S. Wei, Y. Zhang, X. Yang, Z. Tang, and Y. Li. 2025. Deep Learning Model Inversion Attacks and Defenses: A Comprehensive Survey. Artif. Intell. Rev. (2025). URL
[263] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In NeurIPS. URL
[264] Z. Yang, B. Shao, B. Xuan, E. Chang, and F. Zhang. 2020. Defending model inversion and membership inference attacks via prediction purification. arXiv:2005.03915. URL
[265] Z. Yang, J. Zhang, E. Chang, and Z. Liang. 2019. Neural network inversion in adversarial setting via background knowledge alignment. In CCS. URL
[266] J. Ye, A. Maddi, S. Murakonda, V. Bindschaedler, and R. Shokri. 2022. Enhanced membership inference attacks against machine learning models. In CCS. URL

[267] W. Ye, J. Hu, H. Wang, X. Ti, Z. Xiao, H. Chen, L. Li, L. Feng, S. Wu, and J. Zhao. 2026. An Invariant Latent Space Perspective on Language Model Inversion. In AAAI. URL
[268] D. Yi, Z. Lei, S. Liao, and S. Li. 2014. Learning Face Representation from Scratch. arXiv:1411.7923. URL
[269] H. Yin, A. Mallya, A. Vahdat, J. Alvarez, J. Kautz, and P. Molchanov. 2021. See through gradients: Image batch recovery via gradinversion. In CVPR. URL
[270] H. Yin, P. Molchanov, J. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. Jha, and J. Kautz. 2020. Dreaming to Distill: Data-Free Knowledge Transfer via DeepInversion. In CVPR. URL
[271] Y. Yin, X. Zhang, H. Zhang, F. Li, Y. Yu, X. Cheng, and P. Hu. 2023. Ginver: Generative Model Inversion Attacks Against Collaborative Inference. In WWW. URL
[272] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. 2015. Understanding Neural Networks Through Deep Visualization. In ICML. URL
[273] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. TACL (2014). URL
[274] H. Yu, Y. Qiu, H. Fang, T. Zhuang, B. Chen, S. Yu, B. Wang, S. Xia, and K. Xu. 2026. Rank Matters: Understanding and Defending Model Inversion Attacks via Low-Rank Feature Filtering. In KDD. URL
[275] X. Yuan, K. Chen, J. Zhang, W. Zhang, N. Yu, and Y. Zhang. 2023. Pseudo label-guided model inversion attack via conditional generative adversarial network. In AAAI. URL
[276] X. Yuan, L. Ding, L. Zhang, X. Li, and D. Wu. 2022. Es attack: Model stealing against deep neural networks without data hurdles. IEEE TETCI (2022). URL
[277] X. Yuan and L. Zhang. 2022. Membership inference attacks and defenses in neural network pruning. In USENIX Security. URL
[278] C. Zhang, J. Morris, and V. Shmatikov. 2024. Extracting Prompts by Inverting LLM Outputs. In EMNLP. URL
[279] C. Zhang, J. Sun, X. Zhu, and Y. Fang. 2010. Privacy and security for online social networks: challenges and opportunities. IEEE Netw. (2010). URL
[280] M. Zhang and Y. Chen. 2018. Link prediction based on graph neural networks. In NeurIPS. URL
[281] M. Zhang, Z. Ren, Z. Wang, P. Ren, Z. Chen, P. Hu, and Y. Zhang. 2021. Membership inference attacks against recommender systems. In CCS. URL
[282] R. Zhang, S. Guo, J. Wang, X. Xie, and D. Tao. 2022. A survey on gradient inversion: Attacks, defenses and future directions. In IJCAI. URL
[283] R. Zhang, S. Hidano, and F. Koushanfar. 2022. Text Revealer: Private Text Reconstruction via Model Inversion Attacks against Transformers. arXiv:2209.10505. URL
[284] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR. URL
[285] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too?. In ACL. URL
[286] X. Zhang, W. Guan, J. Lu, Z. Qiu, J. Cheng, X. Wu, and Y. Zheng. 2024. GraphLeak: Patient Record Leakage through Gradients with Knowledge Graph. In WWW. URL
[287] X. Zhang, J. Zhao, and Y. LeCun. 2015. Character-level convolutional networks for text classification. In NeurIPS. URL
[288] Y. Zhang, N. Carlini, and D. Ippolito. 2024. Effective Prompt Extraction from Language Models. In COLM. URL
[289] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song. 2020. The secret revealer: Generative model-inversion attacks against deep neural networks. In CVPR. URL
[290] Y. Zhang, S. Sun, M. Galley, Y. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In ACL. URL
[291] Z. Zhang, L. Chen, F. Zhong, D. Wang, J. Jiang, S. Zhang, H. Jiang, M. Zheng, and X. Li. 2022. Graph neural network approaches for drug-target interactions. Curr. Opin. Struct. Biol. (2022). URL
[292] Z. Zhang, M. Chen, M. Backes, Y. Shen, and Y. Zhang. 2022. Inference attacks against graph neural networks. In USENIX Security. URL
[293] Z. Zhang, Q. Liu, Z. Huang, H. Wang, C. Lu, C. Liu, and E. Chen. 2021. Graphmi: Extracting private graph data from graph neural networks. In IJCAI. URL
[294] X. Zhao, W. Zhang, X. Xiao, and B. Lim. 2021. Exploiting explanations for model inversion attacks. In ICCV. URL
[295] Z. Zhou, C. Zhou, X. Li, J. Yao, Q. Yao, and B. Han. 2023. On Strengthening and Defending Graph Reconstruction Attack with Markov Chain Approximation. In ICML. URL
[296] L. Zhu, Z. Liu, and S. Han. 2019. Deep leakage from gradients. In NeurIPS. URL
[297] X. Zhu, X. Luo, Y. Wu, Y. Jiang, X. Xiao, and B. Ooi. 2025. Passive Inference Attacks on Split Learning via Adversarial Regularization. In NDSS. URL
[298] Y. Zhu, Y. Cheng, H. Zhou, and Y. Lu. 2021. Hermes attack: Steal DNN models with lossless inference accuracy. In USENIX Security. URL
[299] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In ICCV. URL