Anchoring Adversarial Trajectories to Data Manifolds: A Bilevel Transfer Optimization Framework

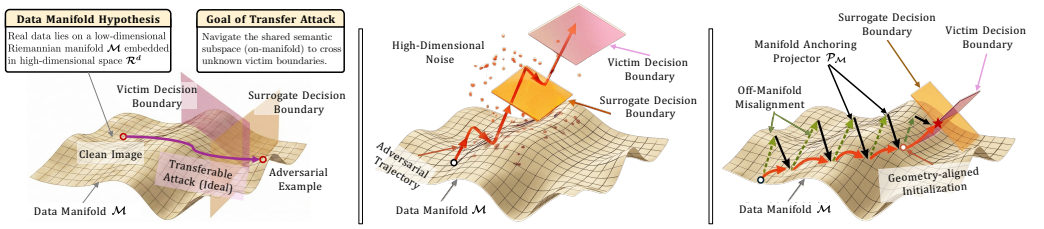
Yaohua Liu[†]
The University of Hong Kong
Hong Kong, China
liuyaohua.918@gmail.com

Yifan Guo[†]
International School of Information Science & Engineering
Dalian University of Technology, Dalian, China
friscoguo@gmail.com

Jiaxin Gao[*]
The Hong Kong Polytechnic University
Hong Kong, China
jiaxinn.gao@outlook.com

Abstract

A key bottleneck in adversarial transfer is a *trajectory-level geometric disconnect*: ambient gradients often drift away from the intrinsic data manifold, causing surrogate-specific overfitting. To rectify this, we propose *Manifold Anchored Bilevel Transfer (MABT)*, a unified framework that anchors adversarial trajectories to the shared semantic subspace. MABT introduces a relaxed manifold-anchoring operator as a semantic rectifier to suppress off-manifold noise. With this constraint, we cast transfer attack generation as a distributional bilevel optimization problem that learns a geometry-aligned initialization by maximizing expected adversarial loss across surrogate variants. We further develop a Hessian-free solver with linear-time complexity to handle the resulting hierarchy. Experiments demonstrate improved transferability for **10** baseline attackers across **28** attack configurations, diverse victim architectures, and defense mechanisms. The official repository is available at `https://github.com/callous-youth/MABT`.



(a) Data Manifold Perspective on Transfer Attacks (b) Off-Manifold Adversarial Trajectory (Standard) (c) Manifold-Anchored Adversarial Trajectory (Ours)

Figure 1: **Motivation.** (a) **Manifold hypothesis**: Natural images concentrate near a low-dimensional manifold $\mathcal{M}$ capturing intrinsic semantics, where decision boundaries of diverse models tend to partially align. (b) **Off-manifold divergence**: Standard iterative attacks inherently drift into high-dimensional noise away from $\mathcal{M}$; in this region, boundaries are less consistent, so the trajectory may cross the **surrogate boundary** but miss the **victim boundary**. (c) **Manifold anchoring**: MABT uses a manifold-anchoring projector with a geometry-aligned initialization to suppress off-manifold deviations while preserving task-relevant semantics to cross *shared boundaries*.

[*]Corresponding author: Jiaxin Gao. [†] Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Adversarial examples (AEs), crafted by adding imperceptible perturbations to clean inputs, expose the fragility of deep learning models [4, 47, 65]. In real-world scenarios where model internals are often inaccessible, *transfer-based attacks* pose a critical security threat [48, 1]. By generating AEs on a white-box surrogate that can mislead unknown black-box victims, these attacks exploit the transferability of adversarial features across different architectures [11, 59].

To boost transferability, extensive research has been devoted to improving attack generalization from multiple perspectives. Broadly, existing methods can be categorized into three primary streams: (i) *Gradient Optimization*, which stabilizes update directions via momentum or variance reduction [11, 32, 58]; (ii) *Input Transformations*, which augment inputs to mitigate overfitting to the surrogate's decision boundary [63, 12, 46]; and (iii) *Model-Refinement and Ensemble Strategies*, which diversify surrogates by modifying architectures [30, 36] or aggregating multi-model gradients [29, 27]. In addition, recent studies have explored feature-level objectives [37, 70] and generative priors [60], though these strategies often involve heuristic data augmentations or model priors. Despite their success, these advances mainly refine isolated factors in the attack pipeline, leaving the underlying mechanism of transferability less explicitly characterized.

We revisit the motivation (Fig. 1) from the geometry of the optimization process, and attribute the transferability bottleneck to a *geometric misalignment* that unfolds in three stages. *(i) Manifold premise.* Prior work in vision [28] and classical manifold theory [5] suggest that natural images concentrate near a low-dimensional manifold embedded in the high-dimensional pixel space; within this semantically structured region, decision boundaries of different models are more likely to exhibit partial alignment. *(ii) Off-manifold drift.* However, standard iterative attacks are driven by ambient-space gradients and, even under ℓ_p constraints, tend to accumulate perturbations dominated by components orthogonal to the manifold. Consequently, the adversarial trajectory gradually departs from the shared semantic subspace and enters off-manifold regions where the natural-image prior provides little geometric guidance. *(iii) Misalignment.* Once this drift occurs, optimization increasingly overfits surrogate-specific artifacts; in these ambient regions, decision boundaries can differ substantially across architectures, which in turn weakens transferability.

1.1 Our Contributions

This geometric perspective suggests that improving transferability requires more than strengthening a single attack component. Instead, the attack trajectory should be explicitly regularized so that its evolution remains aligned with semantically meaningful directions that are more likely to be shared across models. As illustrated in Fig. 2, once such geometric consistency is better preserved, the induced feature shift on victim models grows more steadily rather than saturating early due to surrogate overfitting.

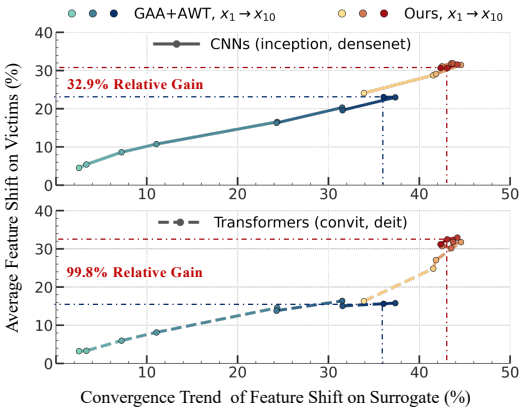


Figure 2: **Feature space divergence.** We quantify victim-side feature shift during the attack process as $1 - \cos(f_{clean}, f_{adv})$, where $\mathbf{x}_1 \to \mathbf{x}_{10}$ denotes the adversarial trajectory over 10 attack iterations. Compared with GAA [16]+AWT [6], MABT induces substantially larger semantic shifts on victim models, with relative gains of **99.8%** on Transformers and **32.9%** on CNNs.

To bridge this geometric disconnect, we propose ***M**anifold-**A**nchored **B**ilevel **T**ransfer (MABT)*, a unified framework that explicitly anchors adversarial trajectories to the intrinsic data geometry (Fig. 1c). First, we construct a relaxed manifold-anchoring operator that actively suppresses off-manifold components while preserving task-relevant semantics. Building on this geometric foundation, we formulate transfer attack as a distributional bilevel optimization problem, which learns a geometry-aligned initialization by minimizing the *expected transfer risk* over a surrogate uncertainty distribution. To handle the resulting hierarchical structure efficiently, we further develop a Hessian-free solver based on

the regularized gap function. Overall, by enforcing geometric consistency throughout optimization, MABT maintains a more stable positive correlation between surrogate-side and victim-side feature shifts, thereby inducing more universal adversarial features and consistently improving transferability across disparate architectures. We summarize the main contributions as follows:

- We revisit transfer attacks through a *manifold anchoring* lens, showing that off-manifold drift is a critical bottleneck hindering transferability and motivating a unified framework to correct the resulting geometric misalignment.
- We introduce the Manifold Anchoring Projector (MAP), a relaxed anchoring operator that filters surrogate-specific off-manifold noise, preserves task-relevant semantics, and supports flexible geometric instantiations.
- We formulate transfer attacks as *Distributional Bilevel Optimization (DBO)* to learn a geometry-aligned initialization by minimizing expected transfer risk. We further introduce a Hessian-free solver with linear complexity to ensure efficient convergence.
- Extensive evaluation on 12 victim models spanning CNNs and Transformers demonstrates improved transferability across **10** baseline attackers and **28** attack configurations under diverse attack and defense settings.

2 Related Work

Transfer-based Adversarial Attacks. Current transfer attacks generally follow an iterative optimization paradigm and refine specific components within this loop. *(1) Gradient and Input Refinement.* To stabilize update directions, Gradient-based methods introduce momentum terms [11, 58], Nesterov acceleration [32], or variance reduction techniques [56] to escape local optima. Simultaneously, Input-based strategies augment the input space to mitigate overfitting, employing diverse transformations such as random resizing [63], translation [12], scale-invariance [32], and patch-wise mixing [57, 19]. *(2) Advanced Objective Design.* Beyond standard losses, recent works design specialized objectives to disrupt intrinsic representations. This includes targeting intermediate feature layers [17, 23], fusing multi-scale attention maps [70], or shifting the perturbation domain from pixels to features [37]. Besides, several works also leverage architecture-specific priors such as eroding feature responses in Ghost networks [29], utilizing skip connections [61], or harnessing the computational redundancy in ViTs [36] and alignment modules [30].

Another line of work improves transferability by modifying the surrogate side or learning generator-based attack mechanisms. Ensemble-based attacks aggregate gradients from multiple models [44, 27] or stochastic variants [3] to reduce model bias. Furthermore, generative approaches [49, 60] and meta-learning frameworks [66, 14] attempt to learn transferable perturbation priors or generalizable attack generators. Beyond adversarial transfer, diffusion models have also been used to model long-horizon trajectories in offline goal-conditioned reinforcement learning [67].

However, these methods still improve transferability mainly through surrogate diversity or learned generation, without explicitly modeling how the adversarial trajectory evolves geometrically. *Summary*. In contrast to existing works that prioritize specific isolated factors, we investigate the optimization process from the perspective of the *adversarial trajectory*. We propose a unified framework that explicitly aligns perturbation evolution with the data manifold, while enabling the initialization to learn transferable geometric priors to rectify the geometric disconnect.

Manifold Perspectives in Deep Learning. Beyond specific attack tactics, the manifold hypothesis [5, 28] also serves as a geometric prior in broader visual tasks. In the security realm, this perspective has been adopted for defense mechanisms. Approaches utilize manifold manipulation to purify adversarial perturbations [71] or detect anomalies that deviate from the natural distribution [33]. A recent theoretical study also analyzes transfer-based attacks from a manifold perspective, explaining transferability through the geometry and curvature of the data manifold [7]. On the offensive front, manifold priors are leveraged to craft stealthy or realistic examples. These methods generally constrain perturbations to bypass detection [2], ensure geometric consistency in 3D domains [18], or perform unrestricted attacks by optimizing adversarial directions in generative latent manifolds [8]. *Summary*. While these studies highlight the relevance of manifold geometry, they mainly use it for purification, detection, theoretical explanation, or realistic example generation. Different from

these approaches, our work uses manifold anchoring as a semantic rectifier to explicitly guide the optimization trajectory for improved transferability.

Bilevel Optimization and Applications. Bilevel optimization provides a unified framework for coupled decision-making [38] and has been widely applied in meta-learning [15], neural architecture search [35], and constraint learning [42]. Recent work has studied non-convex followers [40], lower-level constraints without strong convexity [41], and iterative trajectory construction in bilevel optimization [43]; meanwhile, algorithmic advances have gradually shifted from expensive Hessian-based methods to more efficient first-order solvers [39, 64, 24]. In trustworthy machine learning, bilevel formulations are mainly used to optimize robustness-related objectives [47, 69]. For transfer attacks, recent work [45] has explored bilevel initialization optimization, but it relies on auxiliary victim models to construct the objective and requires high-order gradient estimation, leading to extra dependencies and computational overhead. *Summary*. In contrast, MABT formulates a distributional bilevel optimization problem that minimizes the *expected transfer risk*. This design eliminates the need for extra model accesses and employs a Hessian-free solver, enabling efficient learning of geometry-aligned priors for improved transferability.

3 Methodology

3.1 Perspective

Let $\mathcal{D}$ denote the data distribution supported on a low-dimensional Riemannian manifold $\mathcal{M} \subset \mathbb{R}^d$, embedded within the ambient pixel space $\mathbb{R}^d$. Given a clean pair $(\mathbf{x}, \mathbf{y}) \in \mathcal{M} \times \mathcal{Y}$ and a white-box surrogate $\mathcal{S}$, existing iterative attacks (e.g., PGD) typically generate perturbations by ascending surrogate gradients inside an ℓ_p-norm ball $\Omega = \{\boldsymbol{\delta} \in \mathbb{R}^d \mid \|\boldsymbol{\delta}\|_p \leq \epsilon\}$, without explicitly modeling the manifold geometry of natural images. The update rule is governed by:

$$\boldsymbol{\delta}_{t+1} = \Pi_\Omega \left(\boldsymbol{\delta}_t + \alpha \cdot \operatorname{sign} \left(\nabla_{\mathbf{x}} \mathcal{L}(\mathcal{S}(\mathbf{x} + \boldsymbol{\delta}_t), \mathbf{y}) \right) \right), \tag{1}$$

where Π_Ω projects the update onto the feasible set, and $\operatorname{sign}(\cdot)$ is applied element-wise.

Off-manifold Geometric Misalignment. While Eq. (1) degrades $\mathcal{S}$ via high-frequency gradients, it suffers from intrinsic *geometric misalignment*. Driven by ambient-space gradients, the accumulated perturbation $\boldsymbol{\delta}$ is dominated by orthogonal components ($\mathcal{M}^\perp$) that pull the trajectory away from $\mathcal{M}$. Consequently, the attack drifts into undefined "off-manifold" regions of surrogate-specific artifacts, which fail to transfer to unknown victims $\mathcal{V}$ due to the boundary divergence shown in Fig. 1b.

Manifold Anchoring Projector. To rectify this geometric disconnect, we move beyond standard ℓ_p-bounded attacks defined purely in the ambient space and introduce a *manifold-anchored* evolution process. We propose the MAP, a generalized operator $\mathcal{P}_\mathcal{M} : \mathbb{R}^d \to \mathbb{R}^d$ designed to *anchor* ambient adversarial states toward the intrinsic data manifold by suppressing off-manifold components, while strictly preserving the perturbation budget.

We formulate $\mathcal{P}_\mathcal{M}$ as a *generalized variational anchoring problem*. Given a query state $\mathbf{z} \in \mathbb{R}^d$ (e.g., $\mathbf{z} = \mathbf{x} + \boldsymbol{\delta}$), the ideal anchoring operator seeks an anchored state $\mathbf{u}^*$ by solving:

$$\mathcal{P}_\mathcal{M}[\mathbf{z}] := \underset{\mathbf{u} \in \mathbf{x} + \Omega}{\arg\min} \quad \Big(\underbrace{\Psi(\mathbf{u}, \mathbf{z})}_{\text{Fidelity Potential}} + \underbrace{\mathcal{R}_\mathcal{M}(\mathbf{u})}_{\text{Manifold Prior}} \Big), \tag{2}$$

where $\Psi(\mathbf{u}, \mathbf{z})$ enforces signal fidelity (e.g., Euclidean or perceptual distance), and $\mathcal{R}_\mathcal{M}(\cdot)$ penalizes deviations from the data manifold. Here $\mathbf{x}$ is omitted from $\mathcal{P}_\mathcal{M}$ for notational simplicity. The feasible set $\mathbf{x} + \Omega$ enforces $\|\mathbf{u} - \mathbf{x}\|_p \leq \epsilon$, ensuring strict compliance with the adversarial constraints.

We emphasize that Eq. (2) provides a **generalized formulation of manifold anchoring**, and does not imply that every practical MAP instance exactly solves the variational problem. In practice, efficient operators such as resizing or bit-depth reduction are used as empirical MAP realizations that approximate the anchoring effect by suppressing unstable off-manifold perturbation components while preserving the main semantic structure. Stronger generative operators can provide a closer manifold-anchoring effect at higher computational cost. Different practical MAP instances correspond to different implicit or explicit choices of $\mathcal{R}_\mathcal{M}$, ranging from structural smoothing priors to learned denoising or generative purification priors. Integrating MAP yields the composite attack objective $\mathcal{L}(\mathcal{S}(\mathcal{P}_\mathcal{M}[\mathbf{x} + \boldsymbol{\delta}]), \mathbf{y})$, which regularizes the trajectory away from non-robust orthogonal directions and toward manifold-compatible adversarial vulnerabilities.

Unified and Flexible Implementation. All practical MAP instances are incorporated through a unified operator interface, $\tilde{\mathbf{x}} = \mathcal{P}_{\mathcal{M}}[\mathbf{x}+\boldsymbol{\delta}]$, before surrogate evaluation. Thus, different implementations only instantiate $\mathcal{P}_{\mathcal{M}}$ in different ways, while the subsequent attack objective $\mathcal{L}(\mathcal{S}(\tilde{\mathbf{x}}), \mathbf{y})$ remains unchanged. This interface covers three categories: **(i)** *Explicit Priors*, such as spatial rescaling [20], bit-depth reduction, and frequency filtering, which provide low-cost empirical anchoring; **(ii)** *Discriminative Priors*, such as DnCNN [68], which implement anchoring through learned denoising; and **(iii)** *Generative Priors*, such as Latent Diffusion Models [50], which provide stronger purification-based anchoring at higher cost. This unified view enables a flexible trade-off between anchoring strength and computational efficiency within the same manifold-anchored attack objective. More details are provided in ***Appendix*** *A.2*.

3.2 Distributional Bilevel Optimization

Integrating the geometric constraints of $\mathcal{P}_{\mathcal{M}}$, we formulate transfer attack as DBO. By treating the adversarial initialization $\boldsymbol{\xi}$ as a learnable latent variable, we seek a geometry-aligned state $\boldsymbol{\xi}^*$ whose induced perturbation $\boldsymbol{\delta}^*(\boldsymbol{\xi})$ maximizes the expected adversarial loss under surrogate uncertainty:

$$\min_{\boldsymbol{\xi}\in\Omega} F(\boldsymbol{\xi}, \boldsymbol{\delta}^*(\boldsymbol{\xi})) := -\mathcal{R}_{dist}(\boldsymbol{\delta}^*(\boldsymbol{\xi}); \pi), \text{ s.t. } \boldsymbol{\delta}^*(\boldsymbol{\xi}) = \arg\min_{\boldsymbol{\delta}\in\Omega} \underbrace{\left(-\mathcal{L}\left(\mathcal{S}\big(\mathcal{P}_{\mathcal{M}}[\mathbf{x}+\boldsymbol{\delta}]\big), \mathbf{y}\right)\right)}_{f(\boldsymbol{\delta}, \boldsymbol{\xi}; \mathcal{S})}. \quad (3)$$

Here, $F(\cdot)$ and $f(\cdot)$ denote the upper-level and lower-level objectives, respectively. The mapping $\boldsymbol{\delta}^*(\boldsymbol{\xi})$ is the inner optimal perturbation obtained by solving $f(\cdot)$ from initialization $\boldsymbol{\delta}_0 = \boldsymbol{\xi}$. Although f is written as $f(\boldsymbol{\delta}, \boldsymbol{\xi}; \mathcal{S})$, its dependence on $\boldsymbol{\xi}$ arises through this initialization and the resulting manifold-anchored optimization trajectory.

Lower-level Perturbation Generation. The lower-level problem optimizes $\boldsymbol{\delta}$ under the ℓ_p budget while evaluating the loss on the anchored state $\mathcal{P}_{\mathcal{M}}[\mathbf{x}+\boldsymbol{\delta}]$. This anchoring suppresses surrogate-specific off-manifold artifacts and biases the trajectory toward vulnerabilities that are more likely to remain shared across unknown architectures.

Upper-level Loss Maximization. The upper-level objective $\mathcal{R}_{dist}$ bridges the white-box surrogate and unknown victims by evaluating adversarial loss across surrogate uncertainty. Inspired by the Bayesian perspective on transferability [27], we characterize a local hypothesis family around $\mathcal{S}$ via a surrogate-induced distribution $\pi(\boldsymbol{\theta}|\mathcal{S})$:

$$\mathcal{R}_{dist}(\boldsymbol{\delta}; \pi) := \mathbb{E}_{\boldsymbol{\theta}\sim\pi}\left[\mathcal{L}\left(\mathcal{S}_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\delta}), \mathbf{y}\right)\right]. \quad (4)$$

In practice, $\pi(\boldsymbol{\theta}|\mathcal{S})$ is instantiated as a local Gaussian approximation around a fine-tuned surrogate checkpoint, rather than an exact distribution over unknown victims. The expectation in Eq. (4) is estimated by averaging over a small set of sampled surrogate variants, with implementation details deferred to ***Appendix*** *A.2*. Maximizing this local average discourages single-surrogate overfitting and favors a transferable region under surrogate variation.

3.3 Hessian-Free Solution Strategy

To avoid differentiating through the lower-level dynamics, we adopt a value-function approach based on the *regularized gap function* [64], which encodes lower-level optimality as a differentiable penalty. Let $\mathcal{L}(\boldsymbol{\delta}) := \mathcal{L}(\mathcal{S}(\mathcal{P}_{\mathcal{M}}[\mathbf{x}+\boldsymbol{\delta}]), \mathbf{y})$. The gap function is defined as

$$\mathcal{G}_\gamma(\boldsymbol{\xi}, \boldsymbol{\delta}) := \max_{\hat{\boldsymbol{\delta}}\in\Omega} \left\{\mathcal{L}(\hat{\boldsymbol{\delta}}) - \mathcal{L}(\boldsymbol{\delta}) - \frac{1}{2\gamma}\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|^2\right\}. \quad (5)$$

When $\mathcal{G}_\gamma \to 0$, $\boldsymbol{\delta}$ approaches a stationary point of the lower-level problem. We then optimize the gap-regularized meta objective as follows:

$$\max_{\boldsymbol{\xi}\in\Omega} \phi(\boldsymbol{\xi}) := \frac{1}{\lambda}\mathcal{R}_{dist}(\boldsymbol{\delta}^*(\boldsymbol{\xi}); \pi) - \mathcal{G}_\gamma(\boldsymbol{\xi}, \boldsymbol{\delta}^*(\boldsymbol{\xi})). \quad (6)$$

Exact $\nabla_{\boldsymbol{\xi}}\phi$ requires backpropagation through the inner solver. We therefore use a first-order hypergradient approximation by identifying the dependence on $\boldsymbol{\xi}$ with the current inner state $\boldsymbol{\delta}^*$, i.e., $\nabla_{\boldsymbol{\xi}} \approx \nabla_{\boldsymbol{\delta}}$. By Danskin's theorem, a subgradient of $\mathcal{G}_\gamma$ is given by the maximizer $\hat{\boldsymbol{\delta}}^*$ in Eq. (5). We approximate it by one-step probing with radius σ, $\hat{\boldsymbol{\delta}} = \Pi_\Omega(\boldsymbol{\delta}^* + \sigma\nabla_{\boldsymbol{\delta}}\mathcal{L}(\boldsymbol{\delta}^*))$, yielding

$$\nabla_{\boldsymbol{\xi}}\phi \approx \frac{1}{\lambda}\nabla_{\boldsymbol{\delta}}\mathcal{R}_{dist}(\boldsymbol{\delta}^*; \pi) - \left(\nabla_{\boldsymbol{\delta}}\mathcal{L}(\hat{\boldsymbol{\delta}}) - \nabla_{\boldsymbol{\delta}}\mathcal{L}(\boldsymbol{\delta}^*)\right). \quad (7)$$

Algorithm 1: Manifold Anchored Bilevel Transfer

Input: Image pair $(\mathbf{x}, \mathbf{y})$, surrogate $\mathcal{S}$, MAP $\mathcal{P}_{\mathcal{M}}$. Parameters: perturbation budget ϵ; steps T, K; step sizes α, β; probing radius σ; regularization coeff λ.
Output: Adversarial example $\mathbf{x}_{adv}$.

Initialize $\boldsymbol{\xi}_0 \leftarrow \mathbf{0}$

/* I: Manifold-Constrained Bilevel Learning */
for $t = 0$ **to** $T - 1$ **do**
  $\boldsymbol{\delta}_t \leftarrow \boldsymbol{\xi}_t$
  $\mathbf{g}_t \leftarrow \nabla_{\boldsymbol{\delta}} \mathcal{L}(\mathcal{S}(\mathcal{P}_{\mathcal{M}}[\mathbf{x} + \boldsymbol{\delta}_t]), \mathbf{y})$ // Single-step manifold-anchored attack
  $\boldsymbol{\delta}_{t+1} \leftarrow \Pi_\Omega(\boldsymbol{\delta}_t + \beta \cdot \mathrm{sign}(\mathbf{g}_t))$
  $\mathbf{g}_{t+1} \leftarrow \nabla_{\boldsymbol{\delta}} \mathcal{L}(\mathcal{S}(\mathcal{P}_{\mathcal{M}}[\mathbf{x} + \boldsymbol{\delta}_{t+1}]), \mathbf{y})$ // Probing to approximate the gap maximizer
  $\hat{\boldsymbol{\delta}} \leftarrow \Pi_\Omega(\boldsymbol{\delta}_{t+1} + \sigma \cdot \mathbf{g}_{t+1})$
  $\hat{\mathbf{g}} \leftarrow \nabla_{\boldsymbol{\delta}} \mathcal{L}(\mathcal{S}(\mathcal{P}_{\mathcal{M}}[\mathbf{x} + \hat{\boldsymbol{\delta}}]), \mathbf{y})$ // Gap gradient estimator
  $\nabla \mathcal{G} \leftarrow \hat{\mathbf{g}} - \mathbf{g}_{t+1}$
  $\mathbf{v} \leftarrow \nabla_{\boldsymbol{\delta}} \mathcal{R}_{dist}(\boldsymbol{\delta}_{t+1}; \pi)$ // Distributional transfer risk gradient
  $\boldsymbol{\xi}_{t+1} \leftarrow \Pi_\Omega\big(\boldsymbol{\xi}_t + \alpha \cdot \mathrm{sign}\big(\frac{1}{\lambda}\mathbf{v} - \nabla \mathcal{G}\big)\big)$ // Meta update: maximize $\phi = \frac{1}{\lambda}\mathcal{R}_{dist} - \mathcal{G}_\gamma$
end
/* II: Manifold Adversarial Adaptation */
Reset: $\boldsymbol{\delta}_0 \leftarrow \boldsymbol{\xi}_T$
for $k = 0$ **to** $K - 1$ **do**
  $\boldsymbol{\delta}_{k+1} \leftarrow \Pi_\Omega\big(\boldsymbol{\delta}_k + \alpha \cdot \mathrm{sign}\big(\nabla_{\boldsymbol{\delta}} \mathcal{L}\big(\mathcal{S}(\mathcal{P}_{\mathcal{M}}[\mathbf{x} + \boldsymbol{\delta}_k]), \mathbf{y}\big)\big)\big)$
end
return $\mathbf{x}_{adv} = \mathbf{x} + \boldsymbol{\delta}_K$

This estimator uses only gradients at $\boldsymbol{\delta}^*$ and $\hat{\boldsymbol{\delta}}$, together with $\nabla_{\boldsymbol{\delta}} \mathcal{R}_{dist}$, avoiding any Hessian-vector products. Alg. 1 summarizes the resulting MABT procedure.

Algorithm Analysis. We analyze solver stability using the projected gradient mapping $\mathcal{G}_{proj}(\boldsymbol{\xi}_t) := \frac{1}{\alpha_t}(\Pi_\Omega(\boldsymbol{\xi}_t + \alpha_t \nabla \phi(\boldsymbol{\xi}_t)) - \boldsymbol{\xi}_t)$. Detailed proofs are provided in ***Appendix*** *A.5*.

Theorem 3.1 (Non-asymptotic convergence)**.** *Under standard assumptions, let* $\sigma = \Theta(1/\sqrt{T})$*. With* $\alpha_t = \Theta(1/\sqrt{T})$ *and* $\alpha_t \le \frac{1}{2L_\phi}$ *for sufficiently large* T*, the sequence generated by MABT satisfies:*

$$\min_{0 \le t < T} \mathbb{E}\big[\|\mathcal{G}_{proj}(\boldsymbol{\xi}_t)\|^2\big] \le \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{T}\right) = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \tag{8}$$

Remark. Theorem 3.1 does not claim global optimality of the non-convex bilevel problem; it guarantees projected first-order stationarity of the gap-regularized update under the smooth composite setting. For practical MAP instances that are non-smooth or use surrogate gradients, the same update rule is applied with implementation details and empirical validation provided in the Appendix. For clarity, the analysis uses the projection-based update and omits $\mathrm{sign}(\cdot)$, which is only used in implementation to match the standard ℓ_p-budgeted attack routine; stationarity is defined by $\mathcal{G}_{proj}$.

4 Experiments

Dataset and Models. We follow commonly used settings to select 5,000 correctly classified images from the ImageNet validation set. To evaluate cross-architecture generalization, we employ ResNet-50 and VGG-19 as surrogates. The generated AEs are evaluated on 12 diverse victim models spanning CNNs, robust ensembles, and Transformers: *(i)* **6 Standard CNNs**: Inception-v3 (Inc-v3) [52], IncRes-v2 [53], DenseNet [22], MobileNet [51], PNASNet [34], and SENet [21]; *(ii)* **3 Robust Ensembles**: Inc-v3$_{ens3}$, Inc-v3$_{ens4}$, and IncRes$_{ens}$ [55]; and *(iii)* **3 Transformers**: Visformer [9], DeiT-S [54], and ConViT-B [13]. We use Attack Success Rate (ASR) as the primary metric.

Baselines and Implementation Details. We evaluate MABT as a flexible framework integrated with diverse attackers. We use gradient-based methods (PGD [26], MI [11], VMI [56], GAA [16]) and

Table 1: ASR (%) of MABT across **16** attack configurations on the ResNet-50 surrogate. Results cover 4 base attackers combined with ensemble/model-related strategies. Best results are in **bold**.

Untargeted Attack Scenario, ResNet-50 backbone, Average Success Rate (%) ↑														
Method		**CNN**						**CNN Ensemble**			**Transformer**			**Avg.**
		Inc-v3	Inc-Res-v2	DenseNet	MobileNet	PNASNet	SENet	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Visformer-s	DeiT-s	ConViT-b	
PGD	N/A	16.38	13.04	37.52	36.50	12.80	17.10	10.18	9.40	5.86	10.98	7.50	5.56	15.24
	Ours	**34.96**	**24.58**	**70.00**	**69.22**	**27.84**	**33.94**	**19.72**	**16.38**	**10.28**	**17.02**	**11.54**	**7.62**	**28.59**
	GHOST	18.34	14.44	41.56	41.72	13.60	19.78	10.98	10.00	6.02	12.10	7.14	5.90	16.80
	Ours	**37.38**	**27.52**	**74.66**	**72.58**	**30.84**	**36.72**	**21.18**	**17.66**	**10.50**	**18.24**	**11.92**	**7.72**	**30.58**
	MBA	21.66	14.98	49.22	57.74	13.06	20.22	12.28	10.26	5.94	11.00	7.58	4.52	19.04
	Ours	**44.20**	**32.52**	**78.76**	**81.22**	**32.60**	**39.20**	**27.12**	**22.44**	**14.54**	**19.08**	**13.80**	**8.36**	**34.49**
	AWT	21.24	15.92	42.30	42.38	15.58	18.50	13.52	13.10	8.10	12.04	8.78	6.32	18.15
	Ours	**33.94**	**24.58**	**59.56**	**61.92**	**24.68**	**28.60**	**22.52**	**20.76**	**13.52**	**15.52**	**12.58**	**8.12**	**27.19**
MI	N/A	26.58	21.76	51.36	49.54	22.96	30.26	15.92	14.54	8.98	17.96	11.96	9.04	23.41
	Ours	**59.62**	**47.92**	**88.78**	**87.48**	**49.92**	**58.54**	**37.74**	**32.08**	**21.02**	**34.86**	**22.44**	**15.92**	**46.36**
	GHOST	30.10	24.02	58.28	55.82	24.86	34.00	17.00	15.12	9.86	19.40	12.34	9.06	25.82
	Ours	**63.68**	**51.10**	**91.22**	**90.44**	**53.88**	**62.42**	**40.14**	**33.64**	**22.72**	**37.20**	**24.40**	**16.40**	**48.94**
	MBA	40.42	29.78	72.62	78.42	27.72	37.60	23.32	19.90	12.10	21.68	14.04	9.26	32.24
	Ours	**67.28**	**55.56**	**93.00**	**92.86**	**56.52**	**63.10**	**46.84**	**40.76**	**28.42**	**40.22**	**27.42**	**18.70**	**52.56**
	AWT	32.42	25.82	56.02	55.40	27.16	30.36	21.56	19.74	13.94	18.02	13.76	9.82	27.00
	Ours	**54.24**	**43.10**	**81.44**	**82.00**	**43.54**	**48.90**	**37.48**	**33.70**	**23.78**	**30.18**	**22.12**	**15.02**	**42.96**
TI	N/A	17.30	14.48	40.10	38.18	15.98	19.80	11.24	10.80	6.48	11.50	8.50	6.32	16.72
	Ours	**36.22**	**25.38**	**70.08**	**68.32**	**29.58**	**33.10**	**22.12**	**19.00**	**12.14**	**15.36**	**11.86**	**7.24**	**29.20**
	GHOST	19.92	15.58	44.76	43.26	17.30	22.52	12.00	10.76	6.70	12.38	8.68	6.50	18.36
	Ours	**36.56**	**25.78**	**72.32**	**70.04**	**30.16**	**34.80**	**22.22**	**19.26**	**12.22**	**16.02**	**11.94**	**7.24**	**29.88**
	MBA	24.66	18.22	55.76	60.20	17.46	23.72	14.78	13.20	8.60	11.80	9.24	5.62	21.94
	Ours	**44.82**	**32.28**	**78.48**	**79.34**	**34.76**	**38.40**	**29.52**	**25.94**	**17.24**	**17.80**	**14.82**	**8.50**	**35.16**
	AWT	24.22	18.46	47.44	46.50	19.58	21.64	16.16	15.38	10.34	12.80	10.22	7.20	20.83
	Ours	**36.18**	**25.56**	**62.04**	**63.88**	**27.58**	**29.42**	**25.32**	**23.88**	**16.56**	**15.38**	**13.14**	**8.54**	**28.96**
GAA	N/A	31.46	24.98	57.76	56.68	24.42	27.56	21.76	20.02	12.80	18.60	13.54	9.86	26.62
	Ours	**59.30**	**50.86**	**85.72**	**82.34**	**58.12**	**58.14**	**43.64**	**39.00**	**29.44**	**39.60**	**26.48**	**19.34**	**49.33**
	GHOST	35.20	27.04	64.24	63.72	27.60	31.14	24.64	22.48	14.80	21.24	14.56	10.94	29.80
	Ours	**67.16**	**59.24**	**91.30**	**88.76**	**65.78**	**66.78**	**50.98**	**45.44**	**33.86**	**47.72**	**30.72**	**22.72**	**55.87**
	MBA	33.74	23.44	65.86	73.78	22.02	29.68	21.00	19.28	11.62	17.76	11.30	7.10	28.05
	Ours	**63.84**	**51.42**	**90.12**	**90.24**	**53.68**	**57.10**	**47.42**	**42.52**	**31.18**	**34.26**	**24.96**	**16.58**	**50.28**
	AWT	24.80	17.46	46.90	46.76	16.88	18.16	17.06	16.52	10.22	11.94	10.30	7.12	20.34
	Ours	**47.78**	**37.98**	**71.90**	**70.14**	**39.52**	**39.78**	**35.50**	**34.28**	**25.02**	**23.54**	**19.56**	**13.02**	**38.17**

input-transformation methods (SI [32], DI [63], TI [12]) as *base attackers*. We further combine them with ensemble-based methods (GHOST [29], MBA [27]) and model-related strategies (AWT [6]), covering the major design paradigms of transfer attacks. We also compare BETAK [45] as a recent initialization-based transfer attack. Finally, we evaluate transferability against five defenses: HGD [31], R&P [62], NIPS-r3 [25], JPEG [20], and RS [10].

Hyperparameters: We adopt BETAK's official settings with Inc-v3 as its white-box pseudo-victim. For all methods, we set $\epsilon = 0.03$, $K = 10$, and $\alpha = 0.006$. We use spatial resizing (Resize) as the default MAP instantiation, as it provides the best efficiency–performance trade-off according to the analysis in Sec. 4.2. For MABT, we set $\sigma = 0.01$ following [64]. The iteration $T = 3$, regularization $\lambda = 0.01$, and lower-level step size $\beta = 1.0$ are selected based on ablation results.

4.1 Experimental Results

Additional analyses and results are provided in the ***Appendix***. We further examine the full component attribution of MAP and DBO in Tab. 4, followed by a sensitivity analysis of the main hyperparameters in Fig. 6. The appendix further reports results with additional attackers and a VGG-19 surrogate (Tabs. 6 and 7), as well as results for targeted transfer attacks (Tab. 8).

Quantitative Comparison. As shown in Tab. 1, MABT raises the overall average ASR from 22.52% to 39.28%, corresponding to a **74.4%** relative gain, with particularly clear gains on heterogeneous ViTs and robust ensembles. Results on the VGG-19 surrogate (Tab. 7 in ***Appendix***) further support the effectiveness of MABT under a structurally distinct surrogate. Together, these results demonstrate that MABT can serve as a unified transfer framework across different surrogate choices and victim architectures.

Table 2: ASR (%) results against five representative defenses using the ResNet-50 surrogate.

Method		HGD	JPEG	RS	R&P	NIPS-r3	Avg.	Method		HGD	JPEG	RS	R&P	NIPS-r3	Avg.
Defense Mechanism, ResNet-50 backbone, Average Success Rate (%)↑															
PGD	N/A	16.94	11.60	7.96	12.12	14.28	12.58	**MI**	N/A	27.12	17.54	9.44	13.40	23.10	18.12
	Ours	**36.16**	**21.38**	**10.42**	**13.90**	**29.32**	**22.24**		**Ours**	**59.82**	**39.68**	**14.80**	**18.28**	**53.68**	**37.25**
	GHOST	19.34	11.84	8.24	12.70	15.76	13.58		GHOST	30.94	19.02	9.40	13.92	24.80	19.62
	Ours	**38.86**	**21.94**	**10.40**	**14.26**	**31.44**	**23.38**		**Ours**	**63.02**	**41.56**	**15.08**	**19.40**	**56.08**	**39.03**
	MBA	22.22	13.22	9.22	12.84	19.80	15.46		MBA	40.10	25.82	11.94	15.52	35.80	25.84
	Ours	**43.98**	**28.48**	**11.08**	**14.84**	**40.10**	**27.70**		**Ours**	**67.04**	**48.80**	**16.50**	**19.18**	**63.50**	**43.00**
	AWT	20.94	16.28	10.30	14.54	19.90	16.39		AWT	32.58	25.60	13.18	17.46	30.18	23.80
	Ours	**33.80**	**25.70**	**13.28**	**17.06**	**30.68**	**24.10**		**Ours**	**54.82**	**41.88**	**18.20**	**22.58**	**51.32**	**37.76**
TI	N/A	19.62	12.70	8.44	11.88	15.86	13.70	**GAA**	N/A	32.10	24.34	13.10	18.02	29.14	23.34
	Ours	**37.42**	**23.92**	**11.56**	**14.82**	**30.42**	**23.63**		**Ours**	**60.34**	**43.60**	**15.00**	**20.40**	**55.26**	**38.92**
	GHOST	21.22	13.26	8.66	11.86	17.76	14.55		GHOST	36.80	27.20	14.14	19.64	32.78	26.11
	Ours	**39.20**	**23.98**	**11.76**	**14.92**	**31.22**	**24.22**		**Ours**	**68.12**	**50.70**	**15.70**	**21.54**	**62.86**	**43.78**
	MBA	27.06	16.18	10.24	13.16	23.76	18.08		MBA	32.48	23.24	12.58	17.20	31.04	23.31
	Ours	**45.68**	**31.44**	**12.32**	**16.18**	**40.16**	**29.16**		**Ours**	**63.60**	**47.90**	**15.90**	**19.56**	**61.10**	**41.61**
	AWT	25.44	19.32	11.80	15.48	23.26	19.06		AWT	38.20	30.28	15.12	20.86	35.24	27.94
	Ours	**37.10**	**29.46**	**14.60**	**18.52**	**33.92**	**26.72**		**Ours**	**47.60**	**40.22**	**19.58**	**24.74**	**44.40**	**35.31**

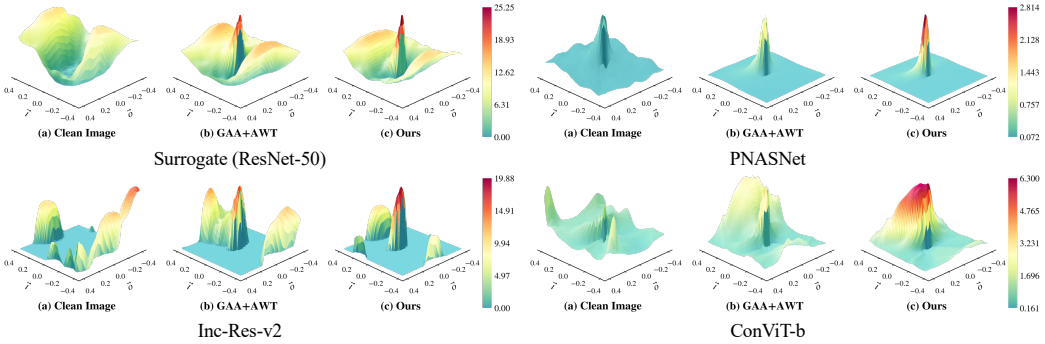


Figure 3: **Adversarial loss landscape**. Surfaces are plotted along the *gradient direction* and a *random orthogonal direction*. Unlike baselines that suffer from landscape collapse on unknown victims, MABT preserves higher victim-side loss basins with relatively flatter surrounding curvature across architectures, which indicates that the learned trajectory better retains transfer-relevant directions.

Visualization of Adversarial Loss Landscapes. To explicate the mechanism of MABT, Fig. 3 visualizes the adversarial loss landscape $\mathcal{L}(\mathbf{x} + a\vec{\boldsymbol{\iota}} + b\vec{\boldsymbol{o}})$, plotted along the gradient direction $\vec{\iota}$ and a random orthogonal direction $\vec{o}$. While mainstream attacks (e.g., GAA+AWT) generate sharp peaks on the surrogate that collapse rapidly on victim models (indicating severe overfitting), MABT maintains a higher loss basin with flatter curvature across architectures. This visualization supports our trajectory-alignment interpretation: MABT tends to preserve higher victim-side loss basins with flatter surrounding curvature, suggesting reduced surrogate-specific overfitting.

Table 3: Comparison with initialization-based BETAK which uses auxiliary *pseudo-victims* (Inc-v3).

Method		Average ASR (%)↑, ResNet-50			Avg.
		CNN	CNN Ensemble	Transformer	
PGD	N/A	22.22	8.48	8.01	15.24
	BETAK	30.87	11.75	10.47	20.99
	Ours	**43.42**	**15.46**	**12.06**	**28.59**
MI	N/A	33.74	13.15	12.99	23.41
	BETAK	43.86	19.74	15.97	30.86
	Ours	**65.38**	**30.28**	**24.41**	**46.36**
TI	N/A	24.31	9.51	8.77	16.72
	BETAK	32.41	13.12	10.75	22.17
	Ours	**43.78**	**17.75**	**11.49**	**29.20**
GAA	N/A	37.14	18.19	14.00	26.62
	BETAK	45.55	23.49	18.46	33.26
	Ours	**65.75**	**37.36**	**28.47**	**49.33**

Evaluation against Defense Mechanisms. We subsequently assess MABT under five representative defense strategies. As reported in Tab. 2, MABT improves the aggregate ASR under these defenses. These results suggest that manifold-anchored perturbations remain more transferable under diverse defense mechanisms.

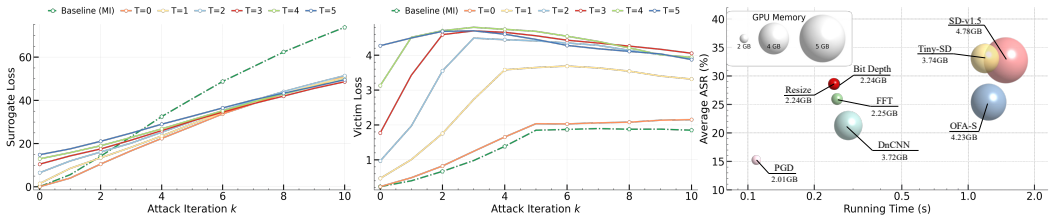


Figure 4: **Convergence and efficiency analysis. Left & Middle:** Increasing T improves geometry-aligned initialization to mitigate surrogate overfitting, thereby ensuring sustained victim loss growth during the subsequent attack. **Right:** Resize provides a favorable low-cost MAP realization, while generative priors achieve stronger anchoring at higher computational cost.

Table 4: Component attribution of MAP and DBO. "w/ DBO, w/o MAP" and "w/o DBO, w/ MAP" retain only the indicated component.

	Method	CNN						CNN Ensemble			Transformer			Avg.
		Inc-v3	Inc-Res-v2	DenseNet	MobileNet	PNASNet	SENet	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Visformer-s	DeiT-s	ConViT-b	
PGD	N/A	16.38	13.04	37.52	36.50	12.80	17.10	10.18	9.40	5.86	10.98	7.50	5.56	15.24
	w/ DBO, w/o MAP	22.64	15.76	52.58	57.24	14.52	22.18	13.02	10.80	6.50	12.26	7.92	5.02	20.04
	w/o DBO, w/ MAP	22.78	18.28	47.94	43.84	20.70	21.92	14.80	14.30	8.90	11.98	10.02	7.02	20.21
	w/ DBO, w/ MAP	**34.96**	**24.58**	**70.00**	**69.22**	**27.84**	**33.94**	**19.72**	**16.38**	**10.28**	**17.02**	**11.54**	**7.62**	**28.59**
	GHOST	18.34	14.44	41.56	41.72	13.60	19.78	10.98	10.00	6.02	12.10	7.14	5.90	16.80
	w/ DBO, w/o MAP	24.24	16.26	54.12	58.32	15.92	23.14	13.20	10.76	5.94	12.46	7.84	5.22	20.62
	w/o DBO, w/ MAP	27.00	20.60	54.90	50.52	24.46	25.96	17.32	15.12	10.36	13.72	10.52	7.60	23.17
	w/ DBO, w/ MAP	**37.38**	**27.52**	**74.66**	**72.58**	**30.84**	**36.72**	**21.18**	**17.66**	**10.50**	**18.24**	**11.92**	**7.72**	**30.58**
	MBA	21.66	14.98	49.22	57.74	13.06	20.22	12.28	10.26	5.94	11.00	7.58	4.52	19.04
	w/ DBO, w/o MAP	32.58	23.54	67.52	72.94	21.02	31.20	17.04	14.24	8.68	16.80	9.76	6.30	26.80
	w/o DBO, w/ MAP	31.26	22.14	64.08	67.16	22.18	26.68	19.96	17.30	11.24	12.02	10.92	6.28	25.94
	w/ DBO, w/ MAP	**44.20**	**32.52**	**78.76**	**81.22**	**32.60**	**39.20**	**27.12**	**22.44**	**14.54**	**19.08**	**13.80**	**8.36**	**34.49**
	AWT	21.24	15.92	42.30	42.38	15.58	18.50	13.52	13.10	8.10	12.04	8.78	6.32	18.15
	w/ DBO, w/o MAP	28.28	19.86	54.62	59.52	18.54	23.84	17.84	16.12	10.30	14.02	10.40	6.88	23.35
	w/o DBO, w/ MAP	27.88	21.52	50.40	48.02	22.00	23.52	18.14	17.22	11.84	13.66	11.28	8.12	22.80
	w/ DBO, w/ MAP	**33.94**	**24.58**	**59.56**	**61.92**	**24.68**	**28.60**	**22.52**	**20.76**	**13.52**	**15.52**	**12.58**	**8.12**	**27.19**

Comparison with Initialization Attackers. We also benchmark MABT against BETAK [45]. While BETAK relies on an auxiliary *pseudo-victim* (i.e., Inc-v3), MABT operates without extra victim model dependencies, anchoring the adversarial trajectory through the intrinsic data geometry. To maintain a consistent evaluation protocol, the mean ASR of the CNN victim suite incorporates results from Inc-v3. On GAA, Tab. 3 shows that MABT increases the average ASR from 33.26% for BETAK to 49.33%, corresponding to a 48.32% relative improvement and supporting the effectiveness of intrinsic geometric alignment.

Targeted Attack Performance. Targeted transfer is intrinsically more challenging than untargeted transfer because an adversarial perturbation must steer unseen victim models toward a specified target class. Accordingly, the absolute ASR values in Tab. 8 of ***Appendix*** remain low for all methods. Under the same evaluation protocol, MABT improves the corresponding attacks, indicating that geometry-aware optimization can also provide measurable benefits in the targeted setting.

4.2 Ablation Results

Convergence Analysis. Fig. 4 shows distinct optimization behaviors. PGD rapidly improves the surrogate loss but saturates early on the Inc-v3 victim, indicating severe surrogate-specific overfitting. In contrast, larger T in MABT keeps the surrogate trajectory controlled while increasing the victim-side loss, suggesting better transfer-oriented initialization. Even $T = 0$ (MAP-only) outperforms PGD, and $T = 3$ provides a strong trade-off between warm-up strength and transferability.

Effectiveness of MAP and DBO. We evaluate each component through quantitative ablation and geometric feature-shift correlation. *(i)* **Component Gain.** As shown in Tab. 4, both MAP and DBO improve transferability when applied individually across the evaluated attackers. Their joint use consistently achieves the strongest results, suggesting that MAP rectifies the adversarial trajectory

while DBO provides a more transferable initialization for the final attack. *(ii)* **Feature-space Alignment.** Fig. 5 compares surrogate-side and victim-side feature shifts across heterogeneous victims. PGD shows weaker correlation and larger variance, suggesting that stronger surrogate distortion does not reliably transfer to victims. MAP improves this correlation by rectifying the trajectory, while DBO further shifts the initialization toward more transferable regions. Together, the two components yield more consistent surrogate-victim feature-shift alignment across architectures, indicating that MABT better preserves transfer-relevant trajectory directions.

Proxy-based Manifold Evidence. To provide an independent trajectory diagnostic, we additionally evaluate the states produced by PGD and PGD with the default Resize MAP using a pretrained and frozen SD-VAE [50]. The VAE is used only for evaluation and is neither part of MAP construction nor involved in attack optimization. As summarized in Tab. 5 of ***Appendix***, the residual of the ordinary PGD trajectory increases over ten steps, whereas the MAP-anchored trajectory remains close to the clean-image reference. At the same time, the anchored states yield higher loss on the held-out Inc-v3 victim at every recorded step. The paired trend is informative: a lower reconstruction residual alone could arise from merely attenuating the perturbation, but its coexistence with higher victim loss indicates that the anchored trajectory retains, and better exposes, transfer-relevant attack directions.

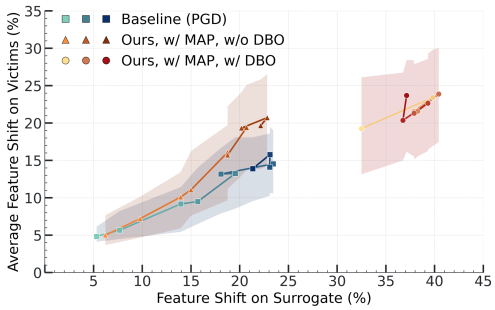


Figure 5: **Feature shift consistency.** We plot surrogate-side (ResNet-50) feature shift against the average victim-side shift. Shaded regions denote the standard deviation over four heterogeneous victims (Inc-v3, DenseNet, DeiT-s and ConViT-b).

Efficiency vs. MAP Performance. As shown in Fig. 4 (Right), the default Resize MAP already yields nearly $\times 2$ ASR over PGD with negligible overhead, making it an efficient low-cost realization. Some stronger generative priors, such as SD-v1.5 and Tiny-SD, further improve ASR by about 4–5 points, but introduce higher latency. At the same time, the results show that stronger or more restrictive anchoring is not a monotonic route to better transferability: different purification operators can suppress surrogate-specific components, but may also attenuate perturbation directions that remain useful across victim models. DBO is therefore complementary to MAP, as it adapts the attack initialization to the constraint induced by a chosen MAP rather than assuming that more aggressive purification is invariably beneficial. Overall, MAP provides a scalable anchoring interface rather than a single fixed operator.

Ablation of Hyperparameters (T, β, λ). Fig. 6 in ***Appendix*** *A.3.2* examines the sensitivity of MABT to its main hyperparameters. The results show that the ASR remains stable over broad ranges of β and λ. Varying λ from 10^{-3} to 10^{-1} leads to only limited changes in performance. In contrast, T controls how extensively DBO refines the attack initialization before the final attack stage. Even small values of T already provide effective initialization refinement, while increasing T yields progressively smaller transfer gains at a higher computational cost. Accordingly, we use $T = 3$ as the default setting throughout the experiments.

5 Conclusion

We propose MABT to rectify the geometric disconnect in adversarial transferability. By synergizing MAP with a distributional bilevel framework, MABT encourages adversarial trajectories to remain closer to the shared semantic subspace. Furthermore, our Hessian-free solver enables low-overhead integration with diverse attacks, achieving strong performance across heterogeneous architectures.

Limitation. Current evaluation mainly focuses on image classification, while extending MABT to dense prediction tasks remains future work. This study may provide useful insights for more reliable robustness evaluation under black-box transfer settings.

Acknowledgements and Disclosures

Funding. This research received no external funding. **Competing interests.** We declare no competing interests.

References

[1] Fengshuo Bai, Runze Liu, Yali Du, Ying Wen, and Yaodong Yang. Rat: Adversarial attacks on deep reinforcement agents for targeted behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[2] Ofir Bar Tal, Adi Haviv, and Amit H. Bermano. Omg-attack: Self-supervised on-manifold generation of transferable evasion attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2023.

[3] Zikui Cai, Yaoteng Tan, and M Salman Asif. Ensemble-based blackbox attacks on dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017.

[5] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. A discussion of semi-supervised learning and transduction. In *Semi-supervised learning*. 2006.

[6] Jiahao Chen, Zhou Feng, Rui Zeng, Yuwen Pu, Chunyi Zhou, Yi Jiang, Yuyou Gan, Jinbao Li, and Shouling Ji. Enhancing adversarial transferability with adversarial weight tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[7] Yanbo Chen and Weiwei Liu. A theory of transfer-based black-box attacks: Explanation and implications. *Advances in Neural Information Processing Systems*, 36:13887–13907, 2023.

[8] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, 36: 51719–51733, 2023.

[9] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[10] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.

[11] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[12] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[13] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, 2021.

[14] Shuman Fang, Jie Li, Xianming Lin, and Rongrong Ji. Learning to learn transferable attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[15] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimilano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. 2018.

[16] Fuquan Gan and Yan Wo. Boosting the transferability of adversarial examples through gradient aggregation. *IEEE Transactions on Information Forensics and Security*, 2025.

[17] Aditya Ganeshan, Vivek BS, and R Venkatesh Babu. Fda: Feature disruptive attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[18] Banibrata Ghosh, Haripriya Harikumar, Svetha Venkatesh, and Santu Rana. Targeted manifold manipulation against adversarial attacks. In *IEEE Conference on Secure and Trustworthy Machine Learning*, 2025.

[19] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision*, 2022.

[20] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[23] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[24] Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, 2021.

[25] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*, 2018.

[26] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*. 2018.

[27] Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. Making substitute models more bayesian can enhance transferability of adversarial examples. In *The Eleventh International Conference on Learning Representations*, 2023.

[28] Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

[29] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[30] Zhiwei Li, Qi Li, Min Ren, Yiwei Ru, and Zhenan Sun. Enhancing adversarial transferability with alignment network. *IEEE Transactions on Information Forensics and Security*, 2025.

[31] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[32] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

[33] Wei-An Lin, Chun Pong Lau, Alexander Levine, Rama Chellappa, and Soheil Feizi. Dual manifold adversarial robustness: Defense against lp and non-lp adversarial attacks. In *Advances in Neural Information Processing Systems*, 2020.

[34] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European Conference on Computer Vision*, 2018.

[35] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations*, 2018.

[36] Jiani Liu, Zhiyuan Wang, Zeliang Zhang, Chao Huang, Susan Liang, Yunlong Tang, and Chenliang Xu. Harnessing the computation redundancy in vits to boost adversarial transferability. *arXiv preprint arXiv:2504.10804*, 2025.

[37] Renpu Liu, Hao Wu, Jiawei Zhang, Xin Cheng, Xiangyang Luo, Bin Ma, and Jinwei Wang. Pixel2feature attack (p2fa): Rethinking the perturbed space to enhance adversarial transferability. In *Forty-second International Conference on Machine Learning*, 2025.

[38] Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10045–10067, 2021.

[39] Risheng Liu, Xuan Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A value-function-based interior-point method for non-convex bi-level optimization. In *International Conference on Machine Learning*, 2021.

[40] Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34:8662–8675, 2021.

[41] Risheng Liu, Yaohua Liu, Wei Yao, Shangzhi Zeng, and Jin Zhang. Averaged method of multipliers for bi-level optimization without lower-level strong convexity. In *International Conference on Machine Learning*, pages 21839–21866. PMLR, 2023.

[42] Risheng Liu, Jiaxin Gao, Xuan Liu, and Xin Fan. Learning with constraint learning: New perspective, solution strategy and various applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):5026–5043, 2024.

[43] Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. Augmenting iterative trajectory for bilevel optimization: Methodology, analysis and extensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[44] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

[45] Yaohua Liu, Jiaxin Gao, Xuan Liu, Xianghao Jiao, Xin Fan, and Risheng Liu. Advancing generalized transfer attack with initialization derived bilevel optimization and dynamic sequence truncation. In *International Joint Conference on Artificial Intelligence*, 2024.

[46] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In *European Conference on Computer Vision*, 2022.

[47] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[48] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[49] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[51] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[52] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[53] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2021.

[55] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[56] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[57] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[58] Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting adversarial transferability through enhanced momentum. *arXiv preprint arXiv:2103.10609*, 2021.

[59] Baoyuan Wu, Zihao Zhu, Li Liu, Qingshan Liu, Zhaofeng He, and Siwei Lyu. Attacks in adversarial machine learning: A systematic survey from the life-cycle perspective, 2024.

[60] Shangbo Wu, Yu-an Tan, Ruinan Ma, Wencong Ma, Dehua Zhu, and Yuanzhang Li. Boosting generative adversarial transferability with self-supervised vision transformer features. *arXiv preprint arXiv:2506.21046*, 2025.

[61] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[62] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

[63] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[64] Wei Yao, Haian Yin, Shangzhi Zeng, and Jin Zhang. Overcoming lower-level constraints in bilevel optimization: A novel approach with regularized gap functions. In *The Thirteenth International Conference on Learning Representations*, 2024.

[65] Jia-Li Yin, Weijian Wang, Wei Lin, Ximeng Liu, et al. Adversarial-inspired backdoor defense via bridging backdoor and adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[66] Zheng Yuan, Jie Zhang, Yunpei Jia, Chuanqi Tan, Tao Xue, and Shiguang Shan. Meta gradient adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[67] Hengrui Zhang, Yuhu Cheng, C. L. Philip Chen, and Xuesong Wang. Diffusion subgoal planning for long-horizon offline goal-conditioned reinforcement learning, 2026. URL `https://arxiv.org/abs/2609.34575`.

[68] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 2017.

[69] Yihua Zhang, Guanhua Zhang, Prashant Khanduri, Mingyi Hong, Shiyu Chang, and Sijia Liu. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *International Conference on Machine Learning*, 2022.

[70] Desheng Zheng, Wuping Ke, Xiaoyu Li, Yaoxin Duan, Guangqiang Yin, and Fan Min. Enhancing the transferability of adversarial attacks via multi-feature attention. *IEEE Transactions on Information Forensics and Security*, 2025.

[71] Jianli Zhou, Chao Liang, and Jun Chen. Manifold projection for adversarial defense on face recognition. In *European Conference on Computer Vision*, 2020.

A Appendix

All the above experiments were conducted on a server equipped with an NVIDIA H100 GPU with 80GB memory.

A.1 Section Overview

Below is a brief summary of the Appendix.

- **Method Details:** ***Appendix*** *A.2* provides practical details of MABT, including the local Gaussian instantiation of the surrogate distribution in DBO, finite-sample estimation of the expected transfer risk, and practical MAP realizations with their backward handling.
- **Runtime and Ablation Results:** ***Appendix*** *A.3* reports runtime-fair comparisons, ablations of different MAP priors, component studies of MAP and DBO, and hyperparameter sensitivity analyses.
- **Comparative Results:** ***Appendix*** *A.4* provides extended empirical evaluations, including full results on additional attackers, VGG-19 surrogate results, targeted attack results, and defense scenarios.
- **Algorithm Analysis:** ***Appendix*** *A.5* presents the assumptions and proof for the convergence result of the proposed gap-regularized solver.

A.2 Method Details

DBO Instantiation. In the main text, the upper-level distribution $\pi(\boldsymbol{\theta}|\mathcal{S})$ is introduced to model local surrogate uncertainty. In practice, we do not assume access to the true distribution of unknown victim models. Instead, we instantiate $\pi(\boldsymbol{\theta}|\mathcal{S})$ as a local Gaussian approximation around a fine-tuned surrogate checkpoint. Starting from the source surrogate $\mathcal{S}$, we first perform lightweight fine-tuning and then sample a small family of surrogate variants by adding Gaussian parameter perturbations around the resulting checkpoint. The sampled variants are fixed during attack optimization, and the expectation in Eq. (4) is computed by finite-sample averaging:

$$\mathcal{R}_{dist}(\boldsymbol{\delta};\pi) \approx \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(\mathcal{S}_{\boldsymbol{\theta}_i}(\mathbf{x}+\boldsymbol{\delta}),\mathbf{y}), \quad \boldsymbol{\theta}_i \sim \pi(\boldsymbol{\theta}|\mathcal{S}). \tag{9}$$

We use $N = 3$ surrogate variants in our implementation to keep the upper-level risk estimation efficient while capturing local surrogate uncertainty. This finite-sample approximation serves as a practical realization of the distributional objective, where $\pi(\boldsymbol{\theta}|\mathcal{S})$ represents a surrogate-induced local hypothesis family rather than the exact distribution of unknown victim models. Practically, the surrogate is fine-tuned following MBA [27] for 10 epochs at a batch size of 1024, utilizing the SGD optimizer with a learning rate of 0.001, momentum of 0.9, and weight decay of 5e-4.

MAP Realizations and Practical Optimization. All MAP realizations are implemented through the same forward anchoring interface, $\tilde{\mathbf{x}} = \mathcal{P}_{\mathcal{M}}[\mathbf{x}+\boldsymbol{\delta}]$, where $\tilde{\mathbf{x}}$ denotes the anchored input used for surrogate evaluation. They differ only in how the operator $\mathcal{P}_{\mathcal{M}}$ is instantiated. We consider the following three categories:

- **Explicit structural priors.** This category includes Resize, Bit-Depth, and FFT filtering. Resize applies bilinear downsampling followed by upsampling with a scale factor of 0.9 to attenuate unstable high-frequency perturbations. Bit-Depth removes small-magnitude variations through low-bit quantization, and FFT filtering suppresses high-frequency components in the frequency domain. These lightweight operators provide efficient empirical anchoring with negligible computational overhead.
- **Discriminative denoising prior.** We instantiate this category with DnCNN [68]. Given an input $\mathbf{u}$, DnCNN predicts a residual $r_\psi(\mathbf{u})$, and the anchored output is computed as $\mathbf{u} - r_\psi(\mathbf{u})$. This learned denoising prior suppresses noise-like perturbation components while preserving the main image structure.
- **Generative purification priors.** This category includes SD-v1.5 [50], Tiny-SD, and OFA-S. These diffusion-based operators map the perturbed input to a purified output, offering potentially stronger but more computationally expensive anchoring realizations.

For practical optimization, differentiable realizations are handled by their native autograd pipelines, while non-smooth or generative realizations use standard surrogate-gradient approximations for update estimation. All variants share the same forward anchoring interface, whereas the theoretical analysis in the main text is stated for smooth composite settings satisfying the required assumptions.

A.3 Ablation Results

A.3.1 Proxy-based Manifold Evidence

We use the independently pretrained checkpoint `stabilityai/sd-vae-ft-mse` as a frozen learned proxy for the natural-image prior. For an image $\mathbf{u}$, let $\hat{\mathbf{u}}$ denote its VAE reconstruction and define the reconstruction residual as

$$r(\mathbf{u}) = \frac{1}{CHW} \sum_{c,h,w} |u_{c,h,w} - \hat{u}_{c,h,w}| \,. \tag{10}$$

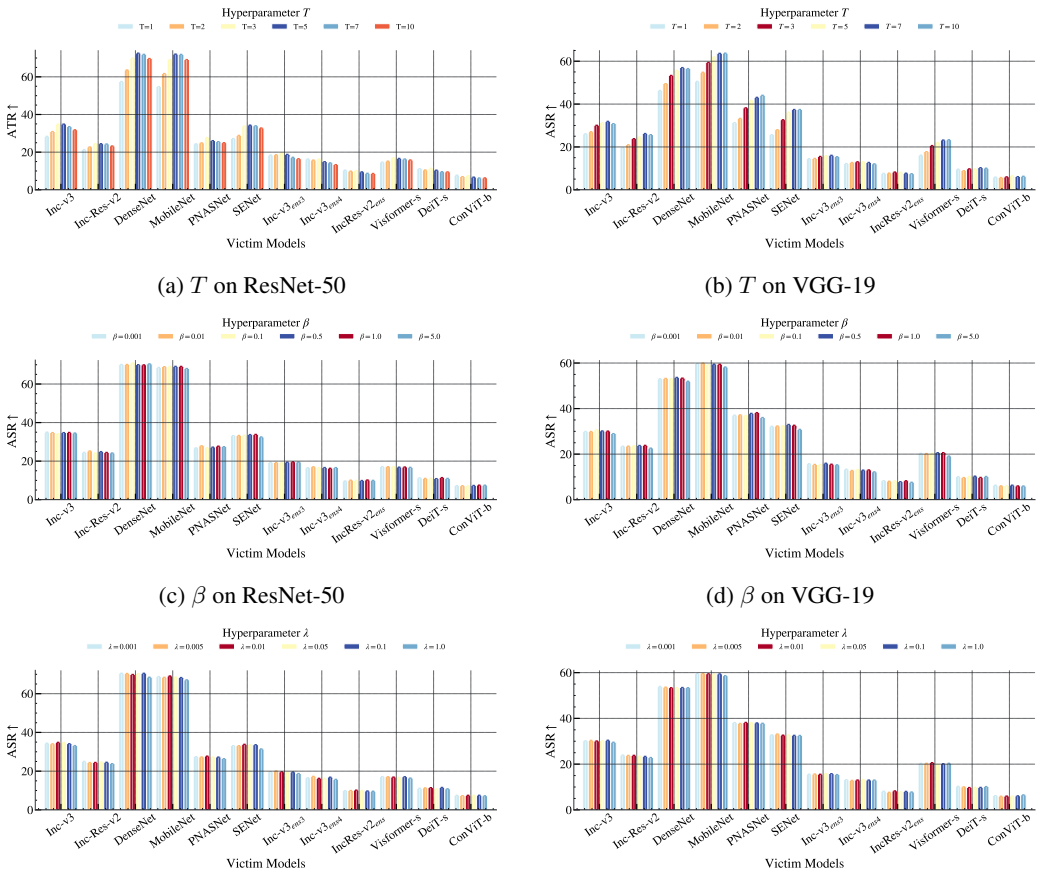
The proxy is evaluated only after each attack update and is not used to construct MAP or compute attack gradients. Tab. 5 reports the trajectory diagnostic for the default Resize realization: compared with PGD, applying MAP after each update keeps the proxy residual close to the clean-image reference while increasing the loss on the Inc-v3 victim.

Table 5: Full proxy-based trajectory diagnostic. "PGD + Resize MAP" denotes the post-MAP state after each PGD update; the frozen SD-VAE is used only for evaluation.

Reconstruction residual $r(\mathbf{u})$ ↓										
State	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**	**Step 7**	**Step 8**	**Step 9**	**Step 10**
Clean image	0.0232	0.0232	0.0232	0.0232	0.0232	0.0232	0.0232	0.0232	0.0232	0.0232
PGD	0.0245	0.0258	0.0270	0.0280	0.0287	0.0294	0.0301	0.0303	0.0307	0.0311
PGD + Resize MAP	**0.0194**	**0.0202**	**0.0211**	**0.0218**	**0.0223**	**0.0229**	**0.0231**	**0.0234**	**0.0237**	**0.0238**
Victim loss on Inc-v3 ↑										
State	**Step 1**	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**	**Step 7**	**Step 8**	**Step 9**	**Step 10**
Clean image	0.1561	0.1561	0.1561	0.1561	0.1561	0.1561	0.1561	0.1561	0.1561	0.1561
PGD	0.1857	0.1908	0.2347	0.2099	0.6499	0.4155	0.9666	0.5821	1.0612	0.7644
PGD + Resize MAP	**0.5322**	**0.8179**	**0.7836**	**1.2244**	**0.9602**	**1.3950**	**1.3806**	**2.0527**	**1.8440**	**2.1038**

A.3.2 Hyperparameter Analysis

Fig. 6 visualizes the sensitivity of T, β, and λ on ResNet-50 and VGG-19 surrogates. MABT is stable to β and λ over broad ranges, suggesting that the gap-regularized update is not sensitive to their exact values. By contrast, T controls the bilevel warm-up strength: small T may under-align the initialization, whereas overly large T may overfit the warm-up to surrogate-specific directions. The consistent trends across both backbones support the default setting $T = 3$, $\beta = 1.0$, and $\lambda = 0.01$.



(a) T on ResNet-50 (b) T on VGG-19

(c) β on ResNet-50 (d) β on VGG-19

(e) λ on ResNet-50 (f) λ on VGG-19

Figure 6: **Hyperparameter sensitivity on ResNet-50 and VGG-19 surrogates.** We evaluate the effects of the warm-up iteration T, lower-level step size β, and gap coefficient λ.

We note that σ determines the probing point for estimating $\nabla\mathcal{G}$, rather than serving as an explicit update coefficient. Because the final meta update applies an element-wise sign operation to $\lambda^{-1}v - \nabla\mathcal{G}$, the gap term can still provide coordinate-level rectification where the primary risk-gradient signal is weak, sign-unstable, or close to a sign transition. Its practical effect is reflected in the component ablation and convergence behavior.

A.4 Comparative Results

Additional Quantitative Results on ResNet-50 Backbone. To provide a comprehensive evaluation, Tab. 6 also includes extended untargeted ASR results for VMI, DI, and SI, leading to 12 method combinations. The aggregate results support the compatibility of MABT with diverse gradient stabilizers and input transformations. Moreover, MABT maintains strong transferability on heterogeneous ViTs and robust ensembles, consistent with the role of anchoring adversarial trajectories to the shared semantic subspace.

Quantitative Results on VGGNet-19 Backbone. Tab. 7 further investigates the cross-architecture transferability of MABT using VGGNet-19 as the surrogate model. Despite the significant structural differences between VGGNet-19 and the diverse victim models, the results provide additional evidence that MABT is effective with a VGG-style surrogate. For instance, when combined with the competitive GAA+MBA baseline, MABT achieves an average ASR of **57.21%**, compared with 32.83% for the original baseline. These results mirror the gains of MABT observed on the ResNet-50 backbone.

Targeted Attack Performance. Tab. 8 reports targeted transfer results with a ResNet-50 surrogate. The absolute success rates remain low across methods, reflecting the difficulty of steering heteroge-

Table 6: Comparison of untargeted ASR (%) by implementing 3 base attackers under MABT, including VMI, DI, and SI. We integrate and compare MABT with different method combinations. The best results are highlighted in **bold**.

Untargeted Attack Scenario, ResNet-50 backbone, ASR (%) ↑															
Method		CNN						CNN Ensemble			Transformer			Avg.
		Inc-v3	Inc-Res-v2	DenseNet	MobileNet	PNASNet	SENet	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Visformer-s	DeiT-s	ConViT-b	
VMI	N/A	36.98	31.44	62.64	59.84	33.42	39.52	23.08	20.34	14.10	25.96	16.78	13.18	31.44
	Ours	**63.64**	**51.00**	**90.96**	**89.84**	**53.10**	**61.86**	**40.50**	**34.10**	**22.98**	**37.46**	**23.44**	**16.88**	**48.81**
	GHOST	39.58	32.42	67.82	64.70	34.50	42.10	24.38	21.74	14.34	26.64	17.38	12.84	33.20
	Ours	**66.22**	**53.82**	**92.64**	**91.62**	**56.46**	**65.62**	**41.76**	**35.72**	**24.20**	**41.26**	**26.08**	**18.48**	**51.16**
	MBA	44.16	32.06	76.32	80.92	30.80	40.30	26.00	22.04	14.08	23.52	16.10	10.40	34.73
	Ours	**66.16**	**53.96**	**92.70**	**93.74**	**54.06**	**62.80**	**43.28**	**36.36**	**24.68**	**40.40**	**25.38**	**17.52**	**50.92**
	AWT	39.90	32.40	63.78	62.12	33.44	35.52	27.70	25.74	18.02	23.02	17.64	13.12	32.70
	Ours	**57.86**	**46.52**	**84.48**	**85.10**	**46.86**	**52.78**	**40.36**	**36.38**	**25.08**	**33.38**	**24.58**	**17.10**	**45.87**
DI	N/A	37.12	31.96	68.84	64.00	36.54	40.00	22.96	20.10	13.16	21.76	13.58	10.48	31.71
	Ours	**56.72**	**44.88**	**87.70**	**84.24**	**52.94**	**53.82**	**37.30**	**30.60**	**20.82**	**27.10**	**17.46**	**11.76**	**43.78**
	GHOST	38.66	30.66	70.12	65.38	35.30	39.12	23.10	19.42	12.76	20.72	13.12	10.38	31.56
	Ours	**57.12**	**45.24**	**87.90**	**85.58**	**51.46**	**53.64**	**36.60**	**30.64**	**20.02**	**26.90**	**17.08**	**11.88**	**43.67**
	MBA	44.84	32.40	78.94	81.34	32.28	38.90	27.16	23.04	15.08	20.08	14.76	8.86	34.81
	Ours	**60.56**	**47.46**	**89.48**	**89.78**	**49.64**	**52.56**	**42.42**	**37.94**	**26.22**	**26.90**	**19.94**	**12.30**	**46.27**
	AWT	38.98	31.42	66.24	62.38	33.54	34.18	26.96	24.14	16.76	20.56	14.70	10.50	31.70
	Ours	**46.06**	**35.06**	**72.78**	**71.98**	**36.32**	**37.64**	**32.46**	**30.42**	**21.02**	**20.42**	**15.56**	**11.06**	**35.90**
SI	N/A	22.12	17.70	48.76	46.46	17.66	23.38	13.14	11.50	7.02	15.32	9.20	7.14	19.95
	Ours	**45.36**	**33.20**	**80.16**	**76.48**	**35.92**	**41.38**	**26.96**	**21.70**	**13.96**	**21.34**	**13.68**	**9.44**	**34.97**
	GHOST	23.82	18.38	53.70	51.80	19.32	25.80	14.52	11.88	7.60	16.18	9.20	6.88	21.59
	Ours	**48.04**	**35.98**	**82.92**	**79.48**	**38.58**	**43.86**	**28.94**	**23.84**	**14.96**	**23.36**	**15.08**	**9.74**	**37.07**
	MBA	24.70	18.10	51.82	59.76	14.10	22.66	13.98	11.06	7.04	12.16	7.60	4.76	20.65
	Ours	**48.20**	**36.12**	**79.10**	**81.22**	**34.56**	**40.76**	**31.10**	**26.46**	**18.26**	**20.34**	**15.20**	**8.66**	**36.67**
	AWT	26.48	19.66	50.60	49.24	19.42	22.84	17.68	16.42	10.76	14.84	10.96	7.70	22.22
	Ours	**41.88**	**31.50**	**70.02**	**69.82**	**31.84**	**34.88**	**28.90**	**26.70**	**18.20**	**19.94**	**15.10**	**10.46**	**33.27**

neous unseen victims toward a prescribed target class. Under matched settings, MABT improves the corresponding baseline attacks and turns several near-zero baseline outcomes into nonzero targeted successes. We therefore view these results as evidence that manifold-aware optimization remains beneficial in the targeted setting, while recognizing that targeted transfer remains a challenging open problem.

A.5 Algorithm Analysis

Scope and notation. The following analysis characterizes projected stationarity of the gap-regularized solver under a smooth composite setting. For concise notation, we write

$$\mathcal{R}_{dist}(\boldsymbol{\xi};\pi) := \mathcal{R}_{dist}(\boldsymbol{\delta}^*(\boldsymbol{\xi});\pi), \quad \mathcal{G}_\gamma(\boldsymbol{\xi}) := \mathcal{G}_\gamma(\boldsymbol{\xi},\boldsymbol{\delta}^*(\boldsymbol{\xi})).$$

Thus,

$$\phi(\boldsymbol{\xi}) = \frac{1}{\lambda}\mathcal{R}_{dist}(\boldsymbol{\xi};\pi) - \mathcal{G}_\gamma(\boldsymbol{\xi}).$$

The practical use of $\mathrm{sign}(\cdot)$ and non-smooth or surrogate-gradient MAP instances is treated as an implementation-level approximation, with details provided in Appendix A.2.

A.5.1 Assumptions

Assumption A.1. The feasible set Ω is non-empty, closed, convex, and compact. The reduced transfer-risk term $\mathcal{R}_{dist}(\boldsymbol{\xi};\pi)$ is L_R-smooth on Ω. The regularized gap term $\mathcal{G}_\gamma(\boldsymbol{\xi})$ is continuously differentiable and L_G-smooth along the analyzed trajectory. Consequently, $\phi(\boldsymbol{\xi})$ is L_ϕ-smooth on Ω, where $L_\phi = \frac{1}{\lambda}L_R + L_G$. These are local analytical conditions for characterizing stationarity of the smooth gap-regularized solver, rather than global convexity assumptions on the practical adversarial loss.

Assumption A.2. The gradient estimator $\mathbf{s}_t$ is stochastic. We assume its conditional variance decays with an effective sample size B_t:

$$\mathbb{E}_t\left[\|\mathbf{s}_t - \mathbb{E}_t[\mathbf{s}_t]\|^2\right] \le \frac{\nu^2}{B_t}, \tag{11}$$

Table 7: Comparison of untargeted ASR (%) using a VGG-19 surrogate. We implement 4 base attackers along with 3 ensembles attacks. The best results in each category are highlighted in **bold**.

Untargeted Attack on ImageNet, VGGNet-19 backbone, ASR (%) ↑														
Method		CNN						CNN Ensemble			Transformer			Avg.
		Inc-v3	Inc-Res-v2	DenseNet	MobileNet	PNASNet	SENet	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Visformer-s	DeiT-s	ConViT-b	
PGD	N/A	12.84	9.54	25.50	32.44	13.06	13.04	7.96	7.46	4.18	9.52	5.10	3.08	11.98
	Ours	**29.74**	**23.50**	**52.56**	**58.34**	**37.58**	**32.70**	**15.38**	**12.92**	**7.94**	**19.84**	**10.28**	**6.28**	**25.59**
	GHOST	12.60	9.64	25.18	32.52	13.46	13.78	7.96	7.42	4.10	9.54	5.30	3.02	12.04
	Ours	**29.94**	**23.22**	**52.50**	**58.46**	**37.38**	**32.02**	**15.26**	**12.94**	**8.16**	**19.88**	**10.14**	**6.42**	**25.53**
	MBA	25.26	19.96	49.56	58.96	29.78	28.40	13.16	11.60	6.78	20.46	8.90	5.18	23.17
	Ours	**46.38**	**38.46**	**72.78**	**76.90**	**58.46**	**49.20**	**25.64**	**20.66**	**13.20**	**32.96**	**15.46**	**9.98**	**38.34**
	AWT	16.24	11.86	28.86	34.70	15.04	12.56	10.32	10.00	5.76	8.90	5.32	3.34	13.58
	Ours	**26.20**	**18.94**	**42.46**	**50.36**	**24.76**	**22.48**	**14.94**	**13.44**	**8.22**	**14.14**	**9.18**	**5.14**	**20.86**
MI	N/A	19.72	15.46	37.08	43.34	21.34	23.16	11.62	10.12	5.86	14.50	8.26	4.94	17.95
	Ours	**38.80**	**30.30**	**62.22**	**67.34**	**46.24**	**42.72**	**19.56**	**16.28**	**10.46**	**27.12**	**14.16**	**9.12**	**32.03**
	GHOST	19.84	15.48	36.46	43.14	21.48	23.18	11.50	10.04	6.20	14.22	8.42	5.00	17.91
	Ours	**38.80**	**30.30**	**62.66**	**67.28**	**46.46**	**42.60**	**19.80**	**16.18**	**10.72**	**27.18**	**14.28**	**8.96**	**32.10**
	MBA	43.28	34.32	68.86	74.34	47.50	44.82	22.80	19.50	12.20	32.00	16.00	10.10	35.48
	Ours	**56.96**	**47.14**	**79.76**	**82.26**	**65.44**	**59.20**	**32.16**	**25.94**	**16.64**	**41.10**	**20.96**	**14.38**	**45.16**
	AWT	25.72	18.34	40.64	46.12	23.66	21.08	15.36	14.58	9.46	13.08	8.32	5.08	20.12
	Ours	**37.68**	**28.22**	**56.94**	**61.96**	**37.40**	**34.40**	**20.84**	**18.68**	**11.58**	**23.84**	**13.62**	**8.40**	**29.46**
TI	N/A	14.12	10.88	28.50	35.80	17.28	17.70	8.78	8.08	4.90	10.96	6.22	3.64	13.91
	Ours	**28.60**	**21.44**	**49.48**	**55.28**	**36.28**	**30.60**	**15.32**	**12.86**	**8.26**	**17.72**	**10.42**	**6.00**	**24.36**
	GHOST	13.92	10.64	28.38	35.78	17.30	17.38	9.02	7.84	4.88	10.42	6.14	3.46	13.76
	Ours	**28.16**	**21.88**	**49.72**	**55.28**	**36.72**	**30.96**	**15.24**	**13.34**	**8.58**	**17.98**	**10.68**	**6.06**	**24.55**
	MBA	27.50	22.24	52.56	61.38	37.96	34.52	15.18	12.50	7.74	21.18	10.50	6.76	25.84
	Ours	**41.82**	**34.08**	**68.04**	**72.16**	**55.44**	**45.20**	**23.32**	**19.92**	**12.52**	**28.28**	**15.12**	**9.46**	**35.45**
	AWT	19.46	13.42	33.04	39.14	17.80	15.28	11.56	11.42	6.92	8.76	6.74	3.70	15.60
	Ours	**27.40**	**19.12**	**43.62**	**50.22**	**25.30**	**22.42**	**16.10**	**14.70**	**9.34**	**13.14**	**9.64**	**5.36**	**21.36**
GAA	N/A	25.66	16.88	42.58	50.36	22.26	19.54	14.14	13.60	7.90	13.16	8.72	5.42	20.02
	Ours	**53.76**	**44.50**	**75.06**	**77.14**	**61.80**	**53.76**	**32.72**	**28.60**	**20.00**	**38.86**	**21.38**	**14.78**	**43.53**
	GHOST	25.50	17.00	42.76	49.96	21.62	19.52	14.80	13.76	7.90	13.24	8.78	5.14	20.00
	Ours	**53.80**	**43.86**	**75.04**	**77.04**	**60.98**	**53.74**	**32.84**	**28.14**	**19.70**	**38.76**	**21.14**	**15.18**	**43.35**
	MBA	41.92	29.26	65.90	74.42	40.84	35.80	23.04	21.96	13.44	25.16	13.80	8.42	32.83
	Ours	**70.50**	**60.52**	**88.64**	**89.40**	**78.30**	**69.90**	**47.60**	**43.16**	**31.70**	**53.66**	**30.62**	**22.46**	**57.21**
	AWT	30.42	20.94	49.02	56.40	28.10	22.80	19.24	18.56	12.00	15.62	10.42	6.34	24.16
	Ours	**39.68**	**28.90**	**57.16**	**61.90**	**37.36**	**31.76**	**25.34**	**23.56**	**15.80**	**20.72**	**13.90**	**8.62**	**30.39**

where B_t is non-decreasing and $B_t \to \infty$ as $t \to \infty$. Moreover, for a T-step run, we assume

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{B_t} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \tag{12}$$

Assumption A.3. Let $\mathbf{b}_t := \mathbb{E}_t[\mathbf{s}_t] - \nabla\phi(\boldsymbol{\xi}_t)$ denote the systematic bias induced by the one-step probing approximation with radius σ and the structural error from the first-order hypergradient approximation. Assume that the local curvature variation is Lipschitz bounded and the structural error ϵ_t satisfies $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\epsilon_t\|^2 \le \mathcal{O}(1/\sqrt{T})$ under the manifold-anchoring constraint. Then the induced bias satisfies

$$\|\mathbf{b}_t\| \le C_{\text{bias}}\sigma + \|\epsilon_t\|,$$

where C_{bias} is independent of t and σ.

A.5.2 Detailed Proof Derivation

Define the *stochastic gradient mapping* at step t as $\tilde{\mathbf{g}}_t := \frac{1}{\alpha_t}(\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t)$, and write $\nabla\phi_t := \nabla\phi(\boldsymbol{\xi}_t)$. Thus, the update rule can be written as $\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t + \alpha_t\tilde{\mathbf{g}}_t$. Since ϕ is L_ϕ-smooth, we have:

$$\begin{aligned}\phi(\boldsymbol{\xi}_{t+1}) &\ge \phi(\boldsymbol{\xi}_t) + \langle\nabla\phi_t, \boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\rangle - \frac{L_\phi}{2}\|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\|^2 \\ &= \phi(\boldsymbol{\xi}_t) + \alpha_t\langle\nabla\phi_t, \tilde{\mathbf{g}}_t\rangle - \frac{L_\phi\alpha_t^2}{2}\|\tilde{\mathbf{g}}_t\|^2,\end{aligned} \tag{13}$$

Recall $\boldsymbol{\xi}_{t+1} = \Pi_\Omega(\boldsymbol{\xi}_t + \alpha_t\mathbf{s}_t)$. By the property of the projection operator Π_Ω onto a convex set Ω, for any $\mathbf{e} \in \Omega$, we have $\langle(\boldsymbol{\xi}_t + \alpha_t\mathbf{s}_t) - \boldsymbol{\xi}_{t+1}, \mathbf{e} - \boldsymbol{\xi}_{t+1}\rangle \le 0$. Setting $\mathbf{e} = \boldsymbol{\xi}_t$, we get:

$$\langle\boldsymbol{\xi}_t + \alpha_t\mathbf{s}_t - \boldsymbol{\xi}_{t+1}, \boldsymbol{\xi}_t - \boldsymbol{\xi}_{t+1}\rangle \le 0. \tag{14}$$

Table 8: Comparative results of targeted ASR against 12 victim architectures.

Method		Targeted Attack on ImageNet, ResNet-50 backbone, ASR (%) ↑												
		CNN						CNN Ensemble			Transformer			Avg.
		Inc-v3	Inc-Res-v2	DenseNet	MobileNet	PNASNet	SENet	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Visformer-s	DeiT-s	ConViT-b	
PGD	N/A	0.02	0.00	0.04	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	Ours	**0.08**	**0.10**	**1.32**	**0.74**	**0.22**	**0.30**	**0.06**	**0.02**	**0.00**	**0.04**	**0.02**	**0.00**	**0.24**
	GHOST	0.02	0.00	0.14	0.08	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.02
	Ours	**0.32**	**0.22**	**2.94**	**1.30**	**0.50**	**0.50**	**0.06**	**0.00**	**0.00**	**0.14**	**0.02**	**0.00**	**0.50**
	MBA	0.10	0.06	1.04	0.98	0.20	0.22	0.02	0.00	0.00	0.06	0.02	0.00	0.23
	Ours	**0.98**	**0.70**	**7.50**	**5.38**	**1.10**	**1.30**	**0.28**	**0.22**	**0.12**	**0.44**	**0.10**	**0.06**	**1.52**
	AWT	0.00	0.00	0.06	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	Ours	**0.24**	**0.04**	**1.62**	**0.90**	**0.16**	**0.18**	**0.04**	**0.00**	**0.00**	**0.06**	**0.00**	**0.00**	**0.27**
MI	N/A	0.04	0.02	0.14	0.14	0.04	0.04	0.02	0.02	0.00	0.02	0.00	0.00	0.04
	Ours	**0.74**	**0.58**	**4.92**	**3.06**	**0.82**	**1.04**	**0.22**	**0.10**	**0.02**	**0.38**	**0.14**	**0.04**	**1.01**
	GHOST	0.06	0.02	0.40	0.24	0.06	0.12	0.02	0.00	0.00	0.02	0.00	0.00	0.08
	Ours	**1.22**	**0.68**	**7.82**	**4.46**	**1.42**	**1.74**	**0.40**	**0.12**	**0.06**	**0.54**	**0.18**	**0.06**	**1.56**
	MBA	0.32	0.14	2.34	1.78	0.46	0.44	0.06	0.04	0.00	0.20	0.04	0.02	0.49
	Ours	**2.28**	**1.66**	**12.34**	**9.36**	**2.12**	**2.24**	**1.08**	**0.46**	**0.22**	**1.00**	**0.34**	**0.24**	**2.78**
	AWT	0.06	0.08	0.20	0.14	0.08	0.04	0.04	0.00	0.00	0.02	0.00	0.00	0.06
	Ours	**0.86**	**0.60**	**5.00**	**3.40**	**0.88**	**1.02**	**0.30**	**0.14**	**0.14**	**0.26**	**0.16**	**0.02**	**1.07**
TI	N/A	0.04	0.02	0.10	0.08	0.02	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.03
	Ours	**0.14**	**0.14**	**1.60**	**0.72**	**0.28**	**0.34**	**0.04**	**0.02**	**0.02**	**0.02**	**0.02**	**0.00**	**0.28**
	GHOST	0.04	0.02	0.18	0.12	0.02	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.04
	Ours	**0.34**	**0.24**	**2.60**	**1.16**	**0.40**	**0.62**	**0.08**	**0.04**	**0.02**	**0.12**	**0.02**	**0.04**	**0.47**
	MBA	0.14	0.10	1.66	0.92	0.30	0.26	0.06	0.04	0.02	0.12	0.04	0.00	0.31
	Ours	**1.14**	**0.76**	**7.58**	**4.80**	**1.48**	**1.14**	**0.46**	**0.22**	**0.16**	**0.38**	**0.14**	**0.12**	**1.53**
	AWT	0.04	0.00	0.14	0.06	0.00	0.04	0.00	0.00	0.00	0.02	0.00	0.00	0.03
	Ours	**0.24**	**0.14**	**1.94**	**1.32**	**0.46**	**0.44**	**0.12**	**0.08**	**0.06**	**0.12**	**0.06**	**0.04**	**0.42**
GAA	N/A	0.04	0.04	0.34	0.30	0.08	0.06	0.02	0.00	0.04	0.02	0.02	0.00	0.08
	Ours	**1.46**	**1.32**	**7.36**	**3.96**	**2.14**	**2.10**	**0.64**	**0.30**	**0.26**	**0.94**	**0.22**	**0.14**	**1.74**
	GHOST	0.06	0.04	0.70	0.48	0.10	0.14	0.02	0.02	0.02	0.04	0.04	0.02	0.14
	Ours	**2.26**	**1.84**	**11.20**	**5.82**	**3.28**	**3.20**	**0.96**	**0.56**	**0.36**	**1.30**	**0.40**	**0.22**	**2.62**
	MBA	0.26	0.20	2.24	2.26	0.24	0.24	0.04	0.08	0.00	0.10	0.08	0.02	0.48
	Ours	**2.42**	**1.48**	**10.98**	**8.04**	**2.42**	**1.98**	**1.08**	**0.66**	**0.50**	**0.78**	**0.36**	**0.10**	**2.57**
	AWT	0.10	0.10	0.44	0.32	0.14	0.08	0.04	0.00	0.02	0.02	0.02	0.00	0.11
	Ours	**0.90**	**0.54**	**3.34**	**1.82**	**1.00**	**0.80**	**0.38**	**0.24**	**0.12**	**0.22**	**0.12**	**0.12**	**0.80**

Substituting $\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t+1} = -\alpha_t \tilde{\mathbf{g}}_t$, this inequality becomes:

$$\langle \alpha_t \mathbf{s}_t - \alpha_t \tilde{\mathbf{g}}_t, -\alpha_t \tilde{\mathbf{g}}_t \rangle \leq 0 \implies -\alpha_t^2 \langle \mathbf{s}_t - \tilde{\mathbf{g}}_t, \tilde{\mathbf{g}}_t \rangle \leq 0. \tag{15}$$

Dividing by α_t^2 and rearranging yields:

$$\langle \mathbf{s}_t, \tilde{\mathbf{g}}_t \rangle \geq \|\tilde{\mathbf{g}}_t\|^2. \tag{16}$$

We substitute $\nabla\phi_t = \mathbf{s}_t + (\nabla\phi_t - \mathbf{s}_t)$ into Eq.(13):

$$\phi_{t+1} \geq \phi_t + \alpha_t \langle \mathbf{s}_t, \tilde{\mathbf{g}}_t \rangle + \alpha_t \langle \nabla\phi_t - \mathbf{s}_t, \tilde{\mathbf{g}}_t \rangle - \frac{L_\phi \alpha_t^2}{2} \|\tilde{\mathbf{g}}_t\|^2. \tag{17}$$

Using Eq.(16) to bound $\alpha_t \langle \mathbf{s}_t, \tilde{\mathbf{g}}_t \rangle \geq \alpha_t \|\tilde{\mathbf{g}}_t\|^2$ and reversing the sign of the second term on the right-hand side:

$$\phi_{t+1} \geq \phi_t + \alpha_t \|\tilde{\mathbf{g}}_t\|^2 - \alpha_t \langle \mathbf{s}_t - \nabla\phi_t, \tilde{\mathbf{g}}_t \rangle - \frac{L_\phi \alpha_t^2}{2} \|\tilde{\mathbf{g}}_t\|^2. \tag{18}$$

Using Young's inequality $\langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{1}{2}\|\mathbf{a}\|^2 + \frac{1}{2}\|\mathbf{b}\|^2$ on the cross term:

$$\alpha_t \langle \mathbf{s}_t - \nabla\phi_t, \tilde{\mathbf{g}}_t \rangle \leq \frac{\alpha_t}{2} \|\mathbf{s}_t - \nabla\phi_t\|^2 + \frac{\alpha_t}{2} \|\tilde{\mathbf{g}}_t\|^2. \tag{19}$$

Substituting back:

$$\phi_{t+1} \geq \phi_t + \left(\frac{\alpha_t}{2} - \frac{L_\phi \alpha_t^2}{2} \right) \|\tilde{\mathbf{g}}_t\|^2 - \frac{\alpha_t}{2} \|\mathbf{s}_t - \nabla\phi_t\|^2. \tag{20}$$

Rearranging the terms, we have:

$$\phi_{t+1} - \phi_t \geq \left(\frac{\alpha_t}{2} - \frac{L_\phi \alpha_t^2}{2} \right) \|\tilde{\mathbf{g}}_t\|^2 - \frac{\alpha_t}{2} \|\mathbf{s}_t - \nabla\phi_t\|^2. \tag{21}$$

Taking the total expectation of both sides. Assuming $\alpha_t \leq \frac{1}{2L_\phi}$, we have $\frac{\alpha_t}{2} - \frac{L_\phi \alpha_t^2}{2} \geq \frac{\alpha_t}{4}$. Then we can get:

$$\frac{\alpha_t}{4}\mathbb{E}[\|\tilde{\mathbf{g}}_t\|^2] \leq \mathbb{E}[\phi_{t+1}] - \mathbb{E}[\phi_t] + \frac{\alpha_t}{2}\mathbb{E}[\|\mathbf{s}_t - \nabla\phi_t\|^2]. \tag{22}$$

Using Assumptions A.2 and A.3, we decompose

$$\begin{aligned} \mathbf{s}_t - \nabla\phi_t &= (\mathbf{s}_t - \mathbb{E}_t[\mathbf{s}_t]) + (\mathbb{E}_t[\mathbf{s}_t] - \nabla\phi_t) \\ &= (\mathbf{s}_t - \mathbb{E}_t[\mathbf{s}_t]) + \mathbf{b}_t, \end{aligned} \tag{23}$$

Taking the total expectation of both sides and obtain the bound

$$\begin{aligned} \mathbb{E}[\|\mathbf{s}_t - \nabla\phi_t\|^2] &\leq 2\mathbb{E}[\|\mathbf{s}_t - \mathbb{E}_t[\mathbf{s}_t]\|^2] + 2\mathbb{E}[\|\mathbf{b}_t\|^2] \\ &\leq \frac{2\nu^2}{B_t} + 4C_{\text{bias}}^2\sigma^2 + 4\mathbb{E}\|\epsilon_t\|^2, \end{aligned} \tag{24}$$

Summing from $t = 0$ to $T - 1$ and dividing by $T\alpha_t/4$, we can get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\tilde{\mathbf{g}}_t\|^2] \leq \frac{4(\phi^* - \phi_0)}{T\alpha_t} + 8C_{\text{bias}}^2\sigma^2 + \frac{8}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\epsilon_t\|^2 + \frac{4\nu^2}{T}\sum_{t=0}^{T-1}\frac{1}{B_t}. \tag{25}$$

By the non-expansiveness of the projection operator $\|\Pi_\Omega(\mathbf{a}) - \Pi_\Omega(\mathbf{b})\| \leq \|\mathbf{a} - \mathbf{b}\|$:

$$\|\tilde{\mathbf{g}}_t - \mathcal{G}_{\text{proj}}(\boldsymbol{\xi}_t)\| = \frac{1}{\alpha_t}\|(\Pi_\Omega(\boldsymbol{\xi}_t + \alpha_t\mathbf{s}_t) - \boldsymbol{\xi}_t) - (\Pi_\Omega(\boldsymbol{\xi}_t + \alpha_t\nabla\phi_t) - \boldsymbol{\xi}_t)\| \tag{26}$$

$$= \frac{1}{\alpha_t}\|\Pi_\Omega(\boldsymbol{\xi}_t + \alpha_t\mathbf{s}_t) - \Pi_\Omega(\boldsymbol{\xi}_t + \alpha_t\nabla\phi_t)\| \tag{27}$$

$$\leq \frac{1}{\alpha_t}\|(\boldsymbol{\xi}_t + \alpha_t\mathbf{s}_t) - (\boldsymbol{\xi}_t + \alpha_t\nabla\phi_t)\| \tag{28}$$

$$= \|\mathbf{s}_t - \nabla\phi_t\|, \tag{29}$$

Using $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$ and substituting the result from the previous equation, we have:

$$\|\mathcal{G}_{\text{proj}}(\boldsymbol{\xi}_t)\|^2 = \|\tilde{\mathbf{g}}_t + \mathcal{G}_{\text{proj}}(\boldsymbol{\xi}_t) - \tilde{\mathbf{g}}_t\|^2 \tag{30}$$

$$\leq 2\|\tilde{\mathbf{g}}_t\|^2 + 2\|\tilde{\mathbf{g}}_t - \mathcal{G}_{\text{proj}}(\boldsymbol{\xi}_t)\|^2$$

$$\leq 2\|\tilde{\mathbf{g}}_t\|^2 + 2\|\mathbf{s}_t - \nabla\phi_t\|^2. \tag{31}$$

Taking expectation of both sides and using Eq. (24), we obtain:

$$\mathbb{E}\left[\|\mathcal{G}_{\text{proj}}(\boldsymbol{\xi}_t)\|^2\right] \leq 2\mathbb{E}\left[\|\tilde{\mathbf{g}}_t\|^2\right] + \frac{4\nu^2}{B_t} + 8C_{\text{bias}}^2\sigma^2 + 8\mathbb{E}\left[\|\epsilon_t\|^2\right]. \tag{32}$$

Averaging over $t = 0, \ldots, T - 1$ and using Eq. (25) yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathcal{G}_{\text{proj}}(\boldsymbol{\xi}_t)\|^2] \leq \frac{8(\phi^* - \phi_0)}{T\alpha_t} + 24C_{\text{bias}}^2\sigma^2 + \frac{24}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\epsilon_t\|^2 + \frac{12\nu^2}{T}\sum_{t=0}^{T-1}\frac{1}{B_t}. \tag{33}$$

By utilizing the property that the minimum value of a sequence is bounded above by its arithmetic mean, we conclude

$$\min_{0\leq t<T}\mathbb{E}[\|\mathcal{G}_{\text{proj}}(\boldsymbol{\xi}_t)\|^2] \leq \frac{8(\phi^* - \phi_0)}{T\alpha_t} + 24C_{\text{bias}}^2\sigma^2 + \frac{24}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\epsilon_t\|^2 + \frac{12\nu^2}{T}\sum_{t=0}^{T-1}\frac{1}{B_t}. \tag{34}$$

Using Assumptions A.2 and A.3, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{B_t} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \quad \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\epsilon_t\|^2 = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \tag{35}$$

Together with $\alpha_t = \Theta(1/\sqrt{T})$ and $\sigma = \Theta(1/\sqrt{T})$, Eq. (34) yields

$$\min_{0\le t<T}\mathbb{E}[\|\mathcal{G}_{\text{proj}}(\boldsymbol{\xi}_t)\|^2] \le \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{T}\right) = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \tag{36}$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction summarize the proposed manifold-anchored transfer framework, including MAP, DBO, the Hessian-free solver, and the main experimental findings. The scope is further clarified by the method details, theoretical assumptions, empirical evaluations, and the limitation statement in the main paper and Appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: In conclusion, we emphasize that the current evaluation mainly focuses on image classification and that extending MABT to dense prediction tasks remains future work. The theoretical scope is also clarified in the method section and Appendix by specifying that the convergence analysis applies to the smooth composite setting.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: The main paper states the non-asymptotic convergence result and clarifies its scope. The Appendix provides the assumptions, reduced notation, and complete proof for the gap-regularized solver.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: The paper specifies the dataset, model zoo, surrogate and victim models, attack budget, iteration numbers, step sizes, MAP instantiations, and key hyperparameters. Additional implementation details, ablations, runtime analysis, and extended results are provided in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: The experiments use the public ImageNet validation set and publicly available model architectures. The paper and Appendix provide algorithmic details, hyperparameters, implementation choices, and runtime information. The official implementation will be released through the repository listed in the abstract.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The experiments describe the ImageNet validation subset, surrogate and victim models, attack budget, iterations, step sizes, MAP choices, and hyperparameters. The Appendix further provides DBO/MAP implementation details, runtime analysis, ablations, and hyperparameter sensitivity.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

Justification: The main paper reports standard-deviation bands in the feature-shift consistency analysis over four heterogeneous victim models. For ASR tables, we evaluate on 5,000 ImageNet validation images under fixed settings and report deterministic aggregate ASR following common practice in transfer-attack evaluation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: The paper reports the GPU setting and provides a runtime-fair comparison in the Appendix, including runtime measurements under a fixed batch size. These details characterize the computational overhead of MABT relative to PGD and other MAP realizations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The study uses public benchmark data and publicly available models for adversarial robustness evaluation. The work is intended for understanding and evaluating black-box robustness, and we acknowledge the dual-use nature of stronger transfer attacks.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: This work studies adversarial transfer attacks, which have clear dual-use implications. Positively, the work can support robustness evaluation and security auditing under black-box settings; negatively, stronger transfer attacks may be misused to attack deployed vision systems, so the work should be used responsibly for evaluation and defense development.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: This paper does not release a new dataset, pretrained model, or high-risk generative system. The work uses public benchmark data and existing models for robustness evaluation; any future code release will be intended for research and evaluation purposes.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: The paper cites the datasets, models, baselines, defenses, and prior methods used in the experiments. The experiments rely on public benchmark data and publicly available model architectures, and all external methods are credited through citations.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper does not introduce a new dataset, benchmark, pretrained model, or other new asset. The contribution is a method and evaluation framework for adversarial transfer attacks.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing, human-subject studies, user studies, or participant data collection.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human subjects, crowdsourcing, user studies, or personally identifiable participant data; therefore IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core method, experiments, and theoretical analysis do not use LLMs as an important, original, or non-standard component. LLMs were used only for language editing and formatting assistance, which did not affect the scientific methodology, experimental results, or theoretical claims.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.