这份文档已经为你更新，在保留 **EATA** 核心“演化+对齐”逻辑的基础上，全面整合了**多模型集成攻击（Ensemble Attack）**的架构。这种结合能显著提升对抗样本在面对未知黑盒模型时的泛化能力。

演化对齐集成变换攻击 (EATA-Ens) 技术方案

1. 背景与动机

目前的输入变换攻击（如 BSR、DIM）主要依赖随机采样。虽然 BSR 通过打乱图像内在关系来提升迁移性，但其均匀随机采样存在两个痛点：

1. **梯度方差与噪声**：包含大量对攻击贡献较小的冗余变换。
2. **单模型过拟合**：即使结合了 BSR，对抗样本仍容易陷入特定白盒模型的局部最优。

EATA-Ens 的核心思想：

- **演化集成搜索**：在变换空间进行在线优化，寻找能同时激发**多个白盒模型**最大 Loss 增长的“广谱最优变换”。
- **跨模型梯度对齐**：通过度量集成梯度方向与动量的一致性，滤除单一模型的过拟合噪声，保留具有强迁移性的核心梯度方向。

2. 算法原理

2.1 变换参数空间定义

定义 BSR 变换为 $\mathcal{T}(x;\pi)$，其中参数 $\pi=\{S,\mathrm{B}\}$：

- S：块打乱的置换矩阵。
- $\mathrm{B}=\{\beta_1,\ldots,\beta_{n^2}\}$：每个块独立的旋转角度。

2.2 集成演化搜索策略 (Evolutionary Ensemble Strategy)

我们不再针对单模型优化，而是定义一个集成目标函数。假设有 N 个白盒模型 $\mathcal{F}=\{f_1,f_2,\ldots,f_N\}$，集成损失函数为：

$$\pi^* = \arg\max_{\pi} \sum_{n=1}^{N} \omega_n \cdot J(f_n(\mathcal{T}(x^{adv};\pi)),y)$$

其中 ω_n 是各模型的权重（通常为 $1/N$）。通过变异和选择，我们在推理侧筛选出对整个模型簇最具“杀伤力”的变换形态。

2.3 跨模型梯度对齐 (Cross-Model Gradient Alignment)

定义第 k 个优选变换在集成模型下产生的梯度为：

$$\mathbf{g}_{\pi_k} = \sum_{n=1}^{N} \omega_n \nabla_{x^{adv}} J(f_n(\mathcal{T}(x^{adv};\pi_k),y)$$

最终综合梯度 $\bar{g}$ 采用余弦相似度加权聚合：

$$\bar{g} = \sum_{k=1}^{K} w_k \cdot \mathbf{g}_{\pi_k}, \quad w_k = \mathrm{Softmax}\left(\frac{\mathbf{g}_{\pi_k} \cdot m_{t-1}}{\|\mathbf{g}_{\pi_k}\|_2 \cdot \|m_{t-1}\|_2} / \tau \right)$$

其中 τ 为温度系数，用于调节权重的平滑度。

3. 算法流程 (Algorithm)

算法 1：EATA-Ens 迭代攻击

输入：白盒模型集合 $\mathcal{F}$，Loss 函数 J，原图 x，扰动限制 ϵ，迭代次数 T，种群规模 M，精英规模 K。

输出：对抗样本 x^{adv}。

1. 初始化 $x_0^{adv} = x, m_0 = 0, \alpha = \epsilon / T$。
2. **For** $t = 0$ **to** $T - 1$:
    - **A. 集成采样与评估**:
        - 随机生成 M 组参数 $\{\pi_1, \ldots, \pi_M\}$。
        - 在**所有白盒模型**上计算推理损失：$L_m = \sum_{f \in \mathcal{F}} \omega_n J(f(\mathcal{T}(x_t^{adv}; \pi_m)), y)$。
    - **B. 精英演化**:
        - 选取 L_m 最高的 K 组参数。
        - 施加微小变异产生后代并重新评估，锁定最终优胜变换集 Π_{best}。
    - **C. 集成梯度计算与对齐**:
        - 计算 Π_{best} 中每个变换在集成模型下的梯度 $\mathbf{g}_{\pi_k}$。
        - 根据 $\mathbf{g}_{\pi_k}$ 与全局动量 m_{t-1} 的余弦相似度计算权重 w_k。
        - 聚合梯度 $\bar{g}_t = \sum w_k \mathbf{g}_{\pi_k}$。
    - **D. 扰动更新**:
        - 更新动量 $m_{t+1} = \mu \cdot m_t + \bar{g}_t / \|\bar{g}_t\|_1$。
        - 更新对抗样本 $x_{t+1}^{adv} = \mathrm{Clip}(x_t^{adv} + \alpha \cdot \mathrm{sign}(m_{t+1}))$.
3. **End For**
4. **Return** x_T^{adv}

4. 理论优势分析

1. **多模型共性挖掘**：EATA-Ens 的演化过程是在寻找不同架构模型（如 ResNet, Inception, ViT）共同的“脆弱结构”。相比单模型演化，它能有效避免陷入单一架构的决策陷阱。
2. **高性能梯度平滑**：通过集成梯度与对齐机制的双重过滤，生成的扰动方向更加稳健，极大提升了对防御模型（如对抗训练模型、输入变换防御）的穿透力。
3. **计算效率优化**：虽然涉及多模型，但**演化阶段（占 80% 的采样量）仅需前向推理。**在显存允许的情况下，多模型并行推理的开销远小于多次反向传播，实现了在固定计算预算下的攻击效果最大化。

给师兄汇报时的核心卖点 (Tips)：

- **“变换空间的元学习”**：强调我们不是在图像空间学习，而是在变换参数空间（置换+旋转）学习。
- **“动态权重分发”**：可以提到在演化中，如果某个模型已经被完全攻破，我们可以通过调整权重 ω_n 迫使算法去攻击剩下的更强模型。
- **“结构化攻击”**：BSR 是随机打乱，而 EATA-Ens 是学习“如何最优地打乱”。