arXiv:2602.01369v1 [cs.CV] 1 Feb 2026

揭示并防御视频混合专家的阿喀琉斯之踵

Songping Wang[1*] **Qinglong Liu**[1*] **Yueming Lyu**[1†] **Ning Li**[2] **ZiwenHe**[3] **Caifeng Shan**[1†]1南京大学 [2]中国移动信息技术有限公司[3]南京信息工程大学

摘要

混合专家（MoE）在视频理解任务中展现出强大的性能，但其对抗鲁棒性仍未得到充分研究。现有攻击方法通常将混合专家视为统一架构，忽视了路由器和专家模块等关键组件各自独立及协同存在的弱点。为填补这一空白，我们提出**T**emporal **L**ipschitz-**G**uided **A**ttacks（**TLGA**），以全面研究视频混合专家模型的组件级漏洞。我们首先针对路由器设计攻击，揭示其独立弱点。在此基础上，我们提出**J**oint **T**emporal **L**ipschitz-**G**uided **A**ttacks（**J-TLGA**），协同扰动路由器和专家。这种联合攻击显著放大了对抗效应，并揭示了混合专家架构的阿喀琉斯之踵（协同弱点）。基于这些发现，我们进一步提出**J**oint **T**emporal**L**ipschitz **A**dversarial **T**raining（**J-TLAT**）。J-TLAT通过联合训练进一步防御协同弱点，提升逐组件鲁棒性。与稠密模型相比，我们的框架即插即用，可将推理成本降低60%以上。该框架在各种数据集和架构上均能稳定提升对抗鲁棒性，有效缓解混合专家的独立及协同弱点。

1 引言

作为人工智能的核心，深度学习近年来取得了快速发展。混合专家（MoE）已成为该领域一种高效且强大的深度学习架构。通过引入路由机制，为每次计算仅激活一小部分专家子网络，混合专家能够在控制推理成本的同时提升模型容量（Jacobs 等人（1991）；Jordan & Jacobs（1994）；Shazeer 等人（2017年））。这种架构优势在视频理解领域尤为突出。视频数据具有复杂的时空结构和长程依赖，这使传统稠密模型在平衡表达能力与计算可行性方面面临根本性挑战。混合专家通过根据帧级语义动态选择专家来应对这一挑战，因此展现出捕捉运动动态和时间上下文的卓越能力。因此，混合专家在动作识别和视频语言建模等任务中取得了出色的表现（Jain 等人（2024）；Wu 等人（2024）；Shaabana 等人（2023年）；Cheng 等人（2024b））。

尽管视频混合专家模型取得了优异的性能，其鲁棒性和安全性却未得到相应的关注。与其他深度学习模型一样，它们仍容易受到对抗样本的影响——精心构造的扰动可能导致高置信度的分类错误（Goodfellow 等人（2014）；Xie 等人（2017）；Wang 等人（2025c；2024a；2025d））。这类漏洞会给自动驾驶和监控等安全关键型视频应用带来严重风险（Kong 等人（2025）；Chen 等人（2022））。

具体而言，视频混合专家模型的漏洞尤为复杂。一方面，视频数据具有很强的时序结构，扰动可以在帧间传播，从而导致累积时序误差。另一方面，混合专家的模块化设计——包括多样化的专家和动态路由——引入了额外的攻击面。攻击者可能针对单个

*贡献相同。 †通讯作者。

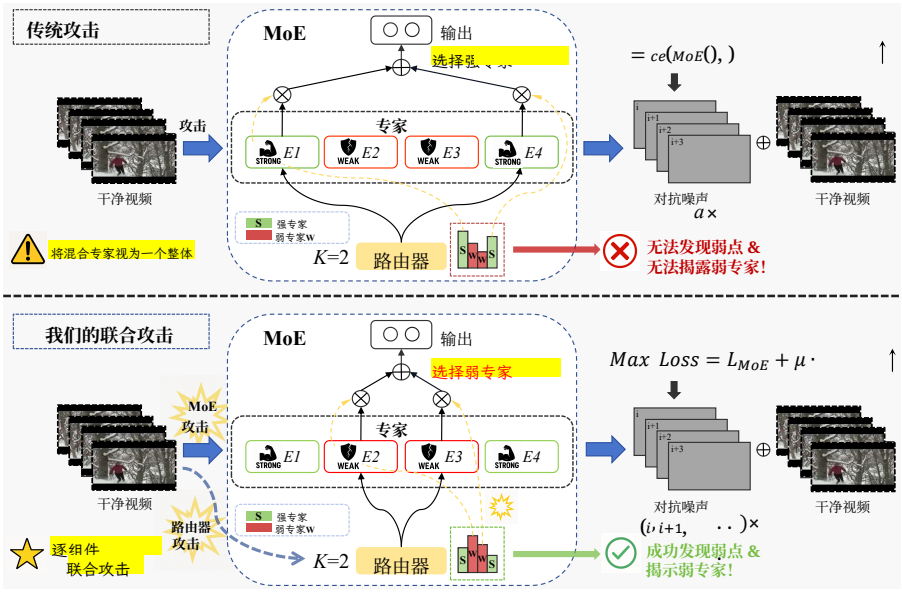


图 1：传统攻击与我们的方法。传统攻击将混合专家模型视为一个整体，忽略其内部架构，因此限制了攻击强度。我们的方法采用针对组件的联合攻击，明确以路由器和专家为目标，引导路由器转向弱专家，同时对其进行扰动。该策略精准揭示了混合专家模型的阿喀琉斯之踵（协同弱点），显著提升了攻击效力。

专家或路由器发起攻击，或破坏它们之间的协同。然而，现有的对抗攻击和训练策略都将混合专家视为一个整体，忽视了内部组件的协同漏洞，尤其是在视频领域。这凸显了亟须发现视频混合专家中的组件级漏洞，并开发专门的对抗训练（AT）机制。由此，我们提出以下问题：

（***Q.A***）对抗攻击会暴露视频混合专家模型的哪些弱点？（***Q.B***）基于这些发现，我们如何开发有效的对抗训练来抵御这些弱点？

为解答上述问题，我们首先研究视频混合专家的组件级漏洞。我们提出了一系列时间 Lipschitz 引导攻击（TLGA），将 Lipschitz 正则性与时间自适应步长相结合，针对路由器、专家以及整体混合专家模型发起攻击。据我们所知，我们首次证明，对路由器施加 TLGA 会导致路由决策崩溃，严重威胁模型的完整性。此外，我们提出联合时序 Lipschitz 引导攻击（J-TLGA），同时针对路由器和专家，表明由于组件间依赖关系，协同攻击会造成更具破坏性的后果。J-TLGA 揭示了协同弱点正是混合专家的阿喀琉斯之踵。

基于 TLGA 揭示的弱点，我们提出两种防御策略：**时**间Lipschitz **对**抗**训**练（**TLAT**）和**联**合**时**间Lipschitz **对**抗**训**练（**J-TLAT**）。TLAT 通过注入 TLGA 提升整个混合专家模型的鲁棒性，而 J-TLAT 则以分层方式将对抗鲁棒性从组件级提升至整体级，从而防御混合专家的致命弱点。此外，我们在附录中推导了混合专家模型的 Lipschitz 上界，为我们的联合攻击与防御框架提供理论支撑。我们的贡献总结如下：

- **组件级漏洞分析：**我们首次对视频混合专家的逐组件漏洞展开系统性调查，揭示了路由器与专家之间独立及协同的失效模式。
- **时间 Lipschitz 引导攻击：**我们提出了 TLGA，它首先揭示了视频混合专家中路由机制的漏洞。我们进一步将其扩展为 J-TLGA，揭示了路由器与专家之间的协同漏洞。

- **弱点引导的防御机制：** 基于已识别的弱点，我们提出了两种对抗训练框架：TLAT 通过 TLGA 增强全局鲁棒性，而 J-TLAT 则通过解决路由器与专家的协同弱点，分层增强逐组件鲁棒性和整体鲁棒性。
- **跨数据集和模型的广泛评估：** 实验表明，J-TLAT 能有效防御混合专家的阿喀琉斯之踵，同时保持推理效率和准确率。我们还在附录中对混合专家的阿喀琉斯之踵进行了理论分析。

2 相关工作

2.1 视频对抗攻击与防御

近期研究表明，视频识别模型仍容易受到对抗扰动的影响。Wei 等人（2019）提出 3D 稀疏扰动，用于在白盒设置下生成对抗样本；Wei 等人（2020）则采用基于启发式的优化，以最小噪声扰动选定的关键帧。Yin 等人（2023）将视频变换纳入损失函数，以增强攻击韧性。Wei 等人（2023）进一步通过 AstFocus 减少时空冗余，同时针对帧间和帧内的关键帧与区域。然而，目前尚无专门针对前景广阔的混合专家而设计的对抗攻击方法。

视频防御机制方面的研究仍然有限，增加了现实应用中的安全风险。OUDefend（Lo 等人（2021年））设计了一种针对对抗视频的修复网络，但其评估仅限于黑盒设置。AAT（Kinfu 和 Vidal（2022年））结合课程学习与自适应对抗训练，以提升对多种攻击的鲁棒性。Yi 等人（2023年）通过时间一致性增强鲁棒性。然而，针对视频混合专家的防御研究也仍不充分。

2.2 混合专家

混合专家（MoE）针对每个输入仅激活 top-k 个子网络（Jacobs 等人（1991）；Wang 等人（2020年）；Fedus 等人（2022年））。尽管混合专家已广泛应用于视频应用（Ma 等人（2025年）），其抵御对抗攻击的能力仍缺乏充分研究：Zhang 等人（2023年）将混合专家的鲁棒性分解为路由器和专家的鲁棒性，但没有专门针对混合专家设计攻击，因此其固有弱点在很大程度上尚未暴露。同样，Han 等人（2024）仅研究卷积神经网络模型，限制了其适用性。Zhang 等人（2025年）提出了一种用于基于图像的混合专家的双模型方法，在鲁棒性与准确率之间实现了更有效的权衡。然而，现有方法均针对图像领域，针对基于视频的混合专家的攻击仍是一个开放问题。此外，由于视频混合专家具有独特结构，其鲁棒性问题十分复杂；将其视为一个整体而忽略各组件之间的相互作用，会使阿喀琉斯之踵难以辨识。此外，视频的特征比图像更复杂，进一步增加了开发针对视频混合专家的对抗攻击方法的难度。解决这些鲁棒性问题对于混合专家在安全关键型应用中的部署至关重要。因此，我们开展系统研究，探讨针对混合专家的攻击与防御策略。

需要指出的是，后续方法均被配置为白盒攻击。我们致力于设计针对混合专家的强力攻击，以最大程度地暴露其根本弱点，并基于这些发现制定有效的防御策略。白盒攻击比灰盒攻击和黑盒攻击更强大，它能够了解目标的架构、参数和梯度，因此可以更深入地探查其内部漏洞。此外，如果一种防御方法能使模型抵御强力的白盒攻击，通常也能更好地防御较弱的灰盒攻击和黑盒攻击。因此，我们设计了一个面向视频混合专家的集成攻防框架，以确保其安全性。

3 预备知识

视频混合专家架构。混合专家是一种神经网络范式，它用一组专门且稀疏激活的子网络（称为专家）替代传统的稠密层。轻量级路由器会针对每个输入动态选择少数几个专家，使模型容量大幅提升，同时将推理成本维持在几乎不变的水平。视频混合专家将

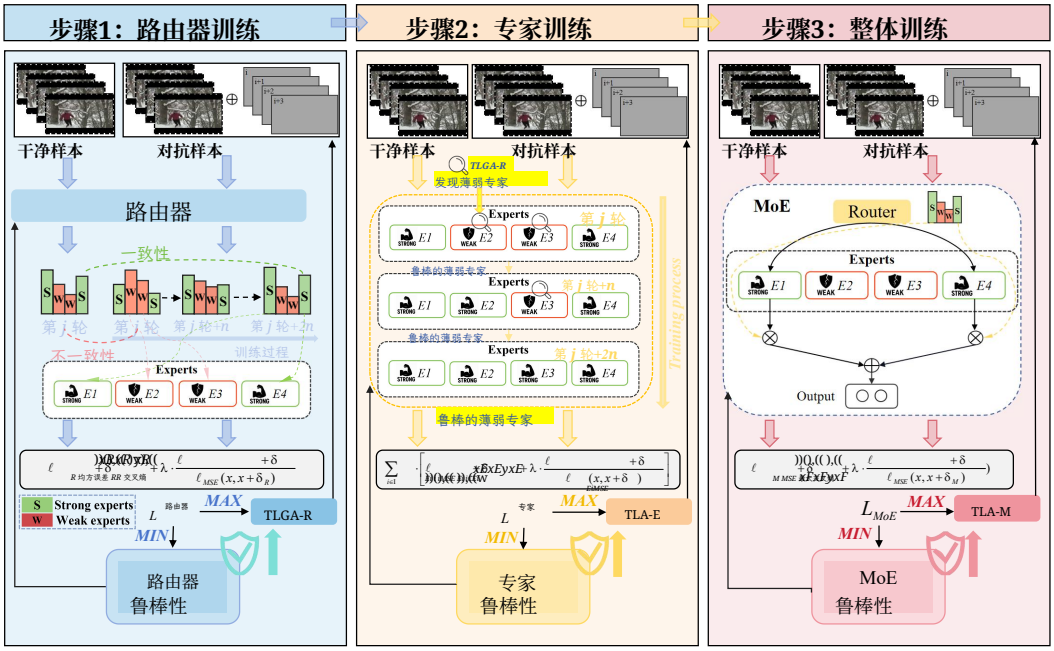


图 2：J-TLAT 框架。面向路由器的时间利普希茨引导攻击（TLGA-R）、面向专家的时间利普希茨攻击（TLA-E）和面向混合专家模型的时间利普希茨攻击（TLA-M）均为混合专家模型而设计。J-TLAT 在每个时期通过三个步骤分层增强逐组件鲁棒性和整体鲁棒性。

视频数据作为输入，视频数据是在图像的基础上扩展出时间轴。视频片段可表示为 $x \in \mathbb{R}^{C\times T\times H\times W}$，其中 C是通道数（例如 RGB），T是帧数，$H\times W$是每帧的空间分辨率。视频混合专家由专家 $\{E_1(\cdot)\ldots,E_M(\cdot)\}$组成。专家 i将视频数据映射为逻辑值 $z_i = E_i(x) \in \mathbb{R}^N$。路由器 $R(\cdot)$为每个专家生成权重： $R(x) = (w_1(x)\ldots,w_M(x))$。混合专家的最终预测是专家输出的加权组合： $F(\mathbf{x}) = \sum_{i=1}^{M} w_i(\mathbf{x})E_i(\mathbf{x}) \in \mathbb{R}^N$。遵循近期的几项研究（Videau 等（2024）；He 等（2024）；Chen 等（2025a））， 我们采用轻量级网络作为路由器 $R(\cdot)$，为给定视频片段中的各个专家分配权重。专家 $E_i(x)$是一种用于捕捉视频特定时空内容的神经网络。更多详情请参见附录。

视频混合专家模型上的对抗攻击。深度网络本质上很脆弱，容易受到对抗扰动的影响；基于这一基础构建的视频混合专家模型也不可避免地面临同样的威胁。旨在攻击视频混合专家模型的对抗扰动 $\delta \in \mathbb{R}^{C\times T\times H\times W}$ 经过优化，以在 ℓ_p-范数预算下最大化损失：

$$\delta^* = \arg\max_{\|\delta\|_p \le \epsilon} \ell_{CE}\Big(\sum_{i=1}^{M} w_i(\mathbf{x}+\delta)E_i(\mathbf{x}+\delta),\, y\Big), \tag{1}$$

其中， y 是真实标签， ℓ_{CE} 表示交叉熵损失。不等式 $\|\delta\|_p \le \epsilon$将对抗噪声 δ 的 p-范数限制为小于或等于阈值 ϵ。我们将此类攻击称为传统攻击，它们将混合专家模型视为一个整体。这类攻击并非专为混合专家模型设计，因此忽略了其关键组件（路由器和专家）。然而，我们知道路由器和专家是混合专家模型的核心组件。二者协同工作，实现混合专家架构中分而治之的核心思想。首先，将混合专家模型视为一个整体并忽略其核心组件，可能会限制攻击的有效性。其次，目前缺乏专门针对视频混合专家架构设计的攻击。因此，我们着手探索针对视频混合专家模型核心组件的攻击。

针对路由器和专家的对抗攻击。 为了进一步探索针对混合专家架构的攻击，并聚焦于其两个核心组件，我们设计了两种组件级攻击：路由器攻击和专家攻击。路由器攻击旨在攻击路由器以改变路由决策，从而增大对抗样本与干净样本之间的输出差异： $\delta_R = \arg\ max_\delta\ \ell$ $(R(x+\delta)\ R(x))$。专家攻击则旨在通过攻击专家来生成

错误输出： $\delta_E = \arg\ max_\delta\ \ell_{CE}(E_i(x+\delta)\ y)$。我们后续的研究证明了组件级攻击的有效性：通过攻击专家和路由器生成对抗样本，能够成功威胁混合专家模型，使其产生错误输出。

4 方法

4.1 时间利普希茨引导攻击

目前缺乏针对视频混合专家架构的攻击，这引出了如下问题：

> ***(Q1)*** 如何设计针对视频混合专家核心组件的攻击，以降低其鲁棒性并发现其弱点？

我们的目标是削弱模型组件的鲁棒性，以实施成功的攻击。利普希茨常数 K 是函数 g 对其输入变化敏感程度的度量，能够有效反映对抗鲁棒性（Pauli 等人（2021年）；Zhang 等人（2022年）；Z¨uhlke 与 Kudenko（2025年））。对于定义在域 $\text{dom}(g) \subset \mathbb{R}^d$上的函数 g： $\mathbb{R}^d \to \mathbb{R}^k$，存在常数 $K > 0$，使得对所有 x，$x + \delta \in \text{dom}(g)$，下列不等式成立：

$$\|g(x) - g(x+\delta)\|_p \leq K\|x - (x+\delta)\|_p, \tag{2}$$

利普希茨常数反映了模型对微小扰动的敏感性。模型越鲁棒，对微小扰动就越不敏感，利普希茨常数也就越小。相反，模型越脆弱，对微小扰动就越敏感，利普希茨常数也就越大。这促使我们在设计攻击时考虑利普希茨性质，以提高混合专家的利普希茨常数，使其对微小扰动更加敏感，从而削弱其鲁棒性。

利普希茨常数量化了模型对微小扰动的敏感性。模型越鲁棒，对微小扰动的敏感性就越低，因此利普希茨常数也越小。相反，越脆弱的模型对这类扰动越敏感，因而具有更大的利普希茨常数。这一认识促使我们利用利普希茨性质设计一种攻击，以增大混合专家（MoE）模型的利普希茨常数。这会使模型对微小扰动更加敏感，从而损害其鲁棒性。

我们将这一想法构造为一个可微且可优化的损失目标，记为 $\mathcal{L}_{\text{Lip}}$。它是局部利普希茨常数的直接有限差分近似，其数学表达式为：

$$\mathcal{L}_{\text{Lip}} = \left(\frac{\|g(x+\delta) - g(x)\|_2}{\|\delta\|_2}\right)^2 = \frac{\|g(x) - g(x+\delta)\|_2^2}{\|x - (x+\delta)\|_2^2} = \frac{\ell_{\text{MSE}}(g(x), g(x+\delta))}{\ell_{\text{MSE}}(x, x+\delta)} \tag{3}$$

此处， g 可以表示整个模型 F、路由器 R或某个专家 E_i。通过最大化 $\mathcal{L}_{\text{Lip}}$，我们可以主动搜索输入空间中导致模型输出发生最大变化的方向，即最大化局部利普希茨值的方向。此外，还可以最小化这一损失项，以增强模型的防御能力。具体而言，最小化 $\mathcal{L}_{\text{Lip}}$会直接惩罚并抑制模型在其最脆弱方向上的陡峭程度。这一过程能够有效平滑决策边界，并降低全局利普希茨常数的上界；我们将在后续章节中进一步讨论这一概念。

此外，视频还具有额外的时间维度。以往方法对每一帧都采用相同的攻击步长，忽略了帧间差异，因此限制了攻击效果。因此，我们考虑利用帧间的梯度差异，为每一帧自适应地分配攻击步长，从而从时间视角增强攻击的破坏性。基于上述考虑，我们设计了时序利普希茨攻击（TLA），从时间维度和噪声敏感性两个方面削弱视频混合专家架构的鲁棒性，如下式所示：

$$\ell_{MoE} = \ell_{CE}(F(x+\delta_M), y) + \lambda \cdot \frac{\ell_{MSE}(F(x), F(x+\delta_M))}{\ell_{MSE}(x, x+\delta_M)}, \tag{4}$$

$$V_{t+1} = \beta \cdot V_t + \|\nabla_x \ell_{MoE}(t)\|_2, \tag{5}$$

$$\alpha^* = \frac{\alpha \cdot \epsilon}{1 + \log(1 + \sqrt{V^*})}, \tag{6}$$

$$x_{adv} = Proj(x + \alpha^* \cdot sign(\nabla_x \ell_M)), \tag{7}$$

其中， ℓ_{MSE} 是均方误差损失， λ 用于平衡交叉熵损失与利普希茨正则化， $\nabla_x \ell_{MoE}(t)$表示第 t帧的梯度， $V^* = (V_1$、.....、$V_T)$表示由 β调整的时间动量，用于考虑先前梯度范数的累积效应， α 是基础攻击率， $\alpha^* = (\alpha_1$、.....、$\alpha_T)$表示自适应时间步长， $Proj(\cdot)$表示将扰动限制在 L_{inf}-范数范围内。 $sign(\cdot)$表示符号函数， x_{adv} 表示对抗样本。公式 6 用于沿时间维度平滑地抑制较大的梯度。为了专门对路由器发起有效攻击，我们设计了时间利普希茨路由器攻击（TLA-R），其公式可表示为：

$$\ell_{Router} = \ell_{CE}(R(x + \delta_R), y_R) + \lambda \cdot \frac{\ell_{MSE}(R(x), R(x + \delta_R))}{\ell_{MSE}(x, x + \delta_R)}, \tag{8}$$

其中， y_R 表示干净样本所选专家的索引。通过将利普希茨导数与时间维度深度结合，我们使攻击不仅能够搜索使局部利普希茨值最大的最陡方向，还能智能地在时间维度上分配扰动资源。这一策略确保能够有效实现攻击目标。

针对路由器的**时**间**利**普希茨**攻**击（**路**由器攻击（**TLA-R** ）的其余流程与式 5 至式 7 类似，以下各节中的攻击也同样如此。我们采用不同的组件级攻击方法和扰动预算 ϵ ，攻击一个经对抗训练训练的混合专家模型。我们使用 3D ResNet-18 作为专家，并采用 Top-1 路由器。如表 1 所示，针对路由器的传统攻击 PGD-R 难以利用路由器的弱点，混合专家模型仍能保持超过 42% 的鲁棒准确率。然而，我们的 TLA-R攻击将攻击强度提高了近 20%，有效触发混合专家模型的路由崩溃，从而导致鲁棒准确率急剧下降。我们的方法表明，仅针对路由器的组件级攻击就能对经传统对抗训练训练的混合专家模型构成严重威胁。我们针对专家的攻击方法 TLA-E 也优于传统的专家攻击 PGD-E。

我们进一步推测，对于同一个样本，路由器分配的置信度较高的专家可能更为强大，而分配的置信度较低的专家可能更为脆弱。我们可以设计攻击，引导路由器将样本分配给置信度最低的专家，从而暴露最脆弱的专家并提升攻击性能。基于此，我们进一步设计了时间利普希茨引导路由器攻击（TLGA-R），公式如下：

$$\ell^*_{Router} = \ell_{Router} - \gamma_1 \cdot \ell_{CE}(R(x + \delta_R), \hat{y}_R), \tag{9}$$

其中，$\hat{}y_R$ 表示路由器针对干净样本输出的置信度最低的专家的索引。如表 1 所示，TLGA-R 的性能比 PGD-R 高出近 24%，在路由器攻击中实现了最佳攻击性能。此外，我们的方法不仅证明了组件级攻击的可行性，还为专门针对并暴露混合专家架构弱点提供了方案。我们得到以下洞察：

洞察 ***1***：仅针对路由器本身，组件级攻击就能严重威胁采用传统对抗训练训练的混合专家模型。

4.2 联合时序 LIPSCHITZ 引导攻击

基于洞察 1，我们进一步考察混合专家架构，其中各组件并非彼此孤立，而是协同工作以取得更好的性能。攻击是否也同样如此？因此，我们提出以下问题：

（***Q2***）能否针对混合专家的各组件发起联合攻击，以取得更好的攻击效果？

这促使我们探索针对混合专家架构开发联合攻击。联合攻击的目标包括两个方面：一是攻击路由器，二是攻击

整体架构。对于路由器攻击，我们旨在迫使路由器以较低置信度将输入路由至脆弱专家。对于整体攻击，我们希望削弱整个混合专家架构的鲁棒性，并迫使其产生错误输出。我们在路由器攻击中借鉴 TLGA-R 攻击的思路，在整体架构攻击中借鉴 TLA 攻击的思路。因此，我们分别设计了联合时间利普希茨攻击（J-TLA）和联合时间 Lipschitz 引导攻击（J-TLGA），具体如下：

$$\ell^*_{MoE} = \ell_{MoE} + \gamma_2 \cdot \ell_{Router}, \ \ell^\star_{MoE} = \ell_{MoE} + \gamma_2 \cdot \ell^*_{Router}, \tag{10}$$

其中， ℓ^*_{MoE} 和 $\ell^\star_{MoE}$ 分别表示 J-TLA 和 J-TLGA 的损失。我们进一步在与表 1 一致的设置下进行实验，将我们的联合攻击方法与传统攻击方法进行比较。

如表 1 所示，在 $\epsilon = 14/255$ 的预算下，联合攻击 J-TLA 将经过对抗训练的混合专家模型的鲁棒性大幅降低至仅 4.73%。J-TLGA 进一步增强了攻击能力。在 J-TLGA 的攻击下，鲁棒混合专家模型的准确率仅为 2.54%，远低于传统攻击，揭示了混合专家的阿喀琉斯之踵。这表明：1）传统攻击方法已不再适用于混合专家架构。忽略其内部组件会严重限制性能，也难以深入揭示混合专家架构的弱点。2）组件之间的协作是一把双刃剑，会使联合攻击更具破坏性。基于这些发现，我们得到以下见解：

洞见***2***：组件的漏洞会产生累积效应。实施联合攻击可以提升攻击效果，从而揭示其弱点。

4.3 联合时间 Lipschitz 对抗训练

传统的对抗训练方法已不再适用于视频混合专家架构。由于其采用端到端范式，难以精细修复组件级弱点以及组件协作中的弱点。联合攻击严重威胁其鲁棒性，因此亟须根据混合专家架构的弱点，开发更适用的对抗训练方法。

（***Q3***）我们能否根据攻击发现的弱点，为视频混合专家开发合适的对抗训练机制？

TLA攻击通过针对时间维度和扰动敏感性维度来削弱模型的鲁棒性。反过来，我们利用 TLA 生成的对抗样本对模型进行对抗训练，降低模型对扰动的敏感性，从而增强其对抗鲁棒性。基于此，我们设计了时间 Lipschitz 对抗训练（TLAT），其公式如下：

$$\min_{\theta_{MoE}} \max_{x_{adv}} \ell_{MoE}, \tag{11}$$

其中， θ_{MoE} 表示混合专家的参数。尽管时间 Lipschitz 对抗训练可以增强视频混合专家架构的整体鲁棒性，但无法在组件级实现细粒度的鲁棒性提升。为同时实现混合专家架构的细粒度组件级鲁棒性和整体鲁棒性增强，我们进一步设计了联合时间利普希茨对抗训练（J-TLAT）。联合攻击利用组件的协同弱点发起攻击，而 J-TLAT 则修复攻击发现的弱点。如图 2 所示，J-TLAT 在每个训练时期进行分层鲁棒训练。首先，它增强路由器的鲁棒性，使其在不同扰动下保持一致的输出。其次，基于 TLA-R攻击识别弱专家，并对其进行对抗训练，以解决这些弱点。最后，J-TLAT 对整个混合专家模型进行对抗训练，进一步增强整体协作的鲁棒性，抵御 MoE 的致命弱点。公式如下：

$$step1: \min_{\theta_{Router}} \max_{x_{adv}} \ell_{Router}, \tag{12}$$

$$\ell_{Expert} = \sum_{i \in \mathcal{I}} w_i \cdot \Big[\ell_{CE}\left(E_i(x + \delta_{Ei}), y\right) + \lambda \cdot \frac{\ell_{MSE}\left(E_i(x), E_i(x + \delta_{Ei})\right)}{\ell_{MSE}\left(x, x + \delta_{Ei}\right)} \Big], \tag{13}$$

$$step2: \min_{\theta_{Expert}} \max_{x_{adv}} \ell_{Expert}, \quad step3: \min_{\theta_{MoE}} \max_{x_{adv}} \ell_{MoE}, \tag{14}$$

其中， $\mathcal{I} = \text{Top-2}(Router(x_{adv}))$ 是通过 TLGA-R 攻击获得的弱专家集合。θ_{Router} 其中， θ_{Expert} 分别表示路由器和专家的参数。有关 J-TLAT 的更多详细信息，请参见附录。基于后续实验，我们得到以下见解：

洞见 ***3***：通过修复组件级攻击发现的弱点，联合对抗训练能够以分层方式提升整体和局部对抗鲁棒性。

表 1：UCF-101 上不同扰动预算下各种混合专家攻击的鲁棒准确率（%）。鲁棒准确率表示攻击下预测正确的百分比。

ϵ	路由器攻击			专家攻击		整体攻击			联合攻击	
	PGD-R	TLA-R	TLGA-R	PGD-E	TLA-E	FGSM	PGD	TLA-M	J-TLA	J-TLGA
$8/255$	42.20%	23.52%	**18.13%**	31.98%	**28.11%**	30.00%	22.09%	**15.05%**	7.03%	**4.95%**
$10/255$	41.98%	22.31%	**16.48%**	29.01%	**25.59%**	27.25%	17.58%	**12.42%**	6.81%	**3.41%**
$12/255$	39.78%	21.98%	**16.22%**	26.26%	**22.08%**	25.05%	15.71%	**11.54%**	5.39%	**2.75%**
$14/255$	39.34%	21.65%	**15.82%**	22.97%	**19.89%**	24.29%	14.84%	**11.10%**	4.73%	**2.54%**

5 实验

5.1 实验设置

数据集与识别模型。遵循 Wei 等人（2023）的做法，我们在广泛使用的 UCF-101（Soomro 等人，2012年）和 HMDB-51（Kuehne 等人，2011年）数据集上进行实验。遵循 Wong等人（2020年）的做法，我们采用三种经典模型作为专家网络，包括 3D ResNet-18（Hara 等人，2018）、TSM（Lin 等人，2019年）、Slowfast（Feichtenhofer 等人，2019年）和 R(2+1)D（Huang 等人，2021年），并使用包含四个专家的 Top-1 路由器，以权衡效率与准确率。

基线。为验证我们的方法能否更有效地揭示混合专家的漏洞，我们选取经典的对抗鲁棒性基准，并将其扩展以适用于视频数据：快速梯度符号法（Goodfellow 等人，2014年）、投影梯度下降（Madry 等人，2017年）和 AutoAttack（Croce 和 Hein，2020年）作为基线；此外，我们还纳入主流的白盒视频对抗攻击 TT（Wei 等人，2022年）作为扩展基线，以便进行更全面的比较。在防御评估中，我们使用以下方法作为基线：AT-S 和 AT-D（分别代表稀疏和密集专家网络，均通过对抗训练进行训练）、AT-M、OUD-M 和 AAT-M（均为混合专家架构，分别通过对抗训练、OUD（Lo 等人，2021年）和 AAT（Kinfu 和 Vidal，2022年）进行训练）。

5.2 主要结果

攻击方法评估。表 2 中的实验表明，TLGA-R 攻击显著降低了除 J-TLAT 以外所有方法的鲁棒性，凸显了混合专家架构中路由器的明显弱点。在 J-TLGA 的强攻击（$\epsilon = 8/255$）下，OUD-M 的准确率仅为 2.64%，而 AAT-M 则降至 0%，这表明联合攻击能够有效揭示混合专家的协同弱点，并大幅提升攻击效果。

训练方法评估。与 AT-D（稠密模型）相比，**J-TLAT** 实现了极低的利普希茨常数（0.823）。在所有扰动强度下，面对最强联合攻击 J-TLGA，J-TLAT 的表现均比 AAT-M 提升近 34%，且几乎不增加推理开销，充分证明了联合训练范式在增强混合专家架构对抗鲁棒性方面的有效性。

5.3 消融研究与更多结果

内部攻击。图 3 比较了不同对抗训练配置（PGD、LIP、T-PGD、TLA）下，快速梯度符号法、投影梯度下降和 J-TLGA 攻击的鲁棒准确率。LIP 表示.....的组合

表 2：以 3D ResNet 为骨干网络，对 UCF-101 上的对抗鲁棒性进行全面评估。表中报告了干净准确率（%），以及在四种扰动预算下针对投影梯度下降、AutoAttack、TT、快速梯度符号法、TLA-M、TLA-E、TLGA-R 和 J-TLGA 攻击的鲁棒准确率（%）。此外，GFLOPs 表示计算开销，Lips-R 和 Lips-J 分别表示 TLGA-R 和 J-TLGA 攻击下的利普希茨常数值。

方法	干净 (%)	投影梯度下降 (%)				AutoAttack (%)			
		$\epsilon = 8/255$	$\epsilon = 10/255$	$\epsilon = 12/255$	$\epsilon = 14/255$	$\epsilon = 8/255$	$\epsilon = 10/255$	$\epsilon = 12/255$	$\epsilon = 14/255$
◦AT-S	49.89	23.85	19.45	16.81	13.19	22.75	19.45	16.26	12.97
◦AT-D	54.51	24.84	21.98	18.57	14.73	23.30	20.11	17.25	14.18
◦AT-M	49.23	19.23	18.46	15.93	14.62	21.43	20.00	17.69	15.82
◦OUD-M	51.67	15.16	13.08	11.10	9.12	18.13	16.59	16.26	15.60
◦AAT-M	49.67	23.08	20.66	19.45	18.57	22.20	19.78	18.24	13.41
◦TLAT	51.65	30.22	27.14	22.75	19.34	22.64	20.22	19.89	18.46
•J-TLAT	54.29	36.37	33.63	29.23	26.92	34.29	30.11	27.03	23.63

方法	十亿次浮点运	训练时间 (%)				快速梯度符号法 (%)			
		$\epsilon = 8/255$	$\epsilon = 10/255$	$\epsilon = 12/255$	$\epsilon = 14/255$	$\epsilon = 8/255$	$\epsilon = 10/255$	$\epsilon = 12/255$	$\epsilon = 14/255$
◦AT-S	2.534	27.91	23.74	20.44	17.03	27.47	24.40	21.21	18.79
◦AT-D	4.790	29.23	26.15	22.42	19.01	28.57	24.84	22.75	20.22
◦AT-M	1.831	25.05	22.09	21.54	18.79	29.56	27.36	24.73	24.73
◦OUD-M	19.94	22.75	20.99	20.11	17.47	24.73	24.95	22.97	23.63
◦AAT-M	1.831	24.51	23.41	20.44	19.34	26.92	26.59	26.48	26.37
◦TLAT	1.831	35.16	30.99	27.03	24.84	35.16	33.08	29.56	27.69
•J-TLAT	1.831	37.36	34.29	30.66	28.57	38.68	36.59	33.52	32.20

方法	Lips-R ↓	TLA-M (%)				TLA-E (%)			
		$\epsilon = 8/255$	$\epsilon = 10/255$	$\epsilon = 12/255$	$\epsilon = 14/255$	$\epsilon = 8/255$	$\epsilon = 10/255$	$\epsilon = 12/255$	$\epsilon = 14/255$
◦AT-M	261.8	18.90	16.70	15.05	15.82	29.23	27.36	22.86	19.89
◦OUD-M	1389	8.02	6.70	7.03	5.82	22.53	19.01	16.37	13.63
◦AAT-M	953.3	13.41	11.43	9.45	8.24	20.55	16.70	13.19	12.09
◦TLAT	3.500	34.51	31.21	26.04	23.30	31.98	29.12	24.95	21.10
•J-TLAT	0.823	34.95	31.76	28.24	24.95	34.29	31.43	25.93	23.41

方法	Lips-J ↓	TLGA-R (%)				J-TLGA (%)			
		$\epsilon = 8/255$	$\epsilon = 10/255$	$\epsilon = 12/255$	$\epsilon = 14/255$	$\epsilon = 8/255$	$\epsilon = 10/255$	$\epsilon = 12/255$	$\epsilon = 14/255$
◦AT-M	596.0	18.24	16.15	17.36	15.82	6.15	5.06	4.73	5.17
◦OUD-M	1474	4.29	4.29	4.29	4.40	2.64	2.42	1.87	1.76
◦AAT-M	1157	5.93	5.50	5.93	5.71	0.00	0.11	0.22	0.00
◦TLAT	208.0	26.37	28.13	26.59	27.36	9.78	10.11	9.67	7.36
•J-TLAT	2.343	51.87	52.42	51.76	51.87	33.96	30.33	24.95	21.98

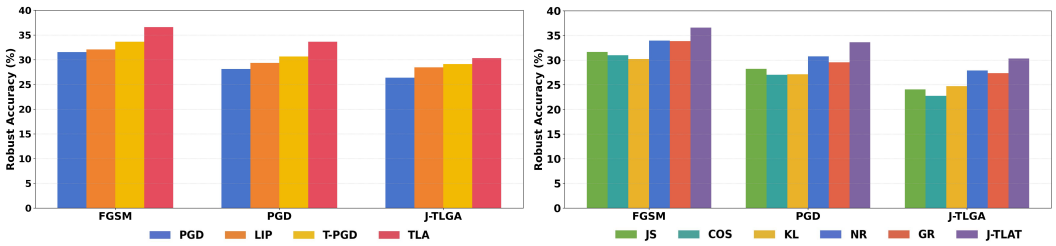


图 3：左图展示了在 J-TLAT 中将不同攻击用作内部攻击时的鲁棒性。右图展示了在 UCF-101 上使用不同正则化方法对 J-TLAT 进行外部训练时的鲁棒性。

PGD攻击与Lipschitz正则化的组合，而 T-PGD 则表示 PGD 与时间步长自适应的组合。TLA 优于其他方法，在 PGD 攻击下实现了 33.63% 的鲁棒准确率（比传统 PGD 高 5.5%），并在 J-TLGA 下保持了 30.33% 的鲁棒准确率，凸显了其抵御结构化攻击的有效性。

正则化策略。我们比较了正则化策略（JS、COS、KL、NR（Tram`er等人（2017年））、GR（Wong等人（2020年））））以及**J-TLAT** 在三种攻击下的表现。其中，JS表示使用Jensen-Shannon散度，COS表示使用余弦损失，KL表示使用Kullback-Leibler散度。**J-TLAT**在所有场景中均取得了最高的鲁棒性（36.59%、33.63%、30.33%），表明其通过降低模型对扰动的敏感性来增强鲁棒性的优势。图4(a)和图4(b)表明，我们的攻击方法TLGA-R和J-TLGA相对于AT-M能够更好地分离对抗样本与干净样本的分布，从而实现更强的破坏力。图4(c)和图4(d)说明J-TLAT降低了交并比



图4：UCF-101 在不同攻击下的交并比分布。交并比分数是用于衡量混合专家或路由器输出一致性的指标。与干净数据的交并比分布偏差越大，对抗数据的攻击成功率就越高。

数据集：UCF-101 模型：AT-M 干净准确率：49.78%				
攻击	$\epsilon = 8/255$	$\epsilon = 10/255$	$\epsilon = 12/255$	$\epsilon = 14/255$
○PGD	22.09%	17.58%	15.71%	14.84%
○LA	20.66%	17.24%	15.38%	14.73%
○T-PGD	18.79%	16.48%	14.18%	13.63%
○TLA	15.05%	12.42%	11.54%	11.10%
●J-TLGA	4.950%	3.410%	2.750%	2.540%

图 5：我们提出的 TLGA 在 UCF-101 上针对 AT-M 的消融研究。LA 表示 Lipschitz 攻击，T-PGD 表示结合时间自适应步长的 PGD。

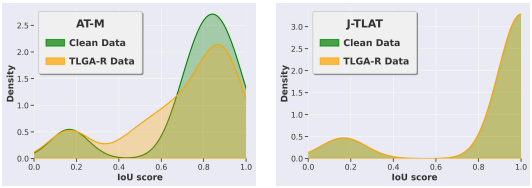


图 6：AT-M 和 J-TLAT 在 UCF-101 上受到 TLGA-R 攻击时的 IoU分数分布。

分布差异，并具有更高的对抗鲁棒性。图5表明，与其他攻击相比，J-TLGA有效地暴露了 MoE 的致命弱点。图6表明，J-TLAT在TLGA-R下确保了更稳定的路由机制。

源模型	方法	3D ResNet	SlowFast	R(2+1)D
3D ResNet	PGD	94.80%	95.00%	85.61%
	FGSM	69.35%	95.58%	87.46%
	TT	66.04%	94.32%	84.78%
	J-TLGA	**20.15%**	**81.79%**	**66.90%**
SlowFast	PGD	87.89%	77.56%	88.48%
	FGSM	88.87%	92.26%	89.06%
	TT	86.84%	89.04%	87.34%
	J-TLGA	**74.00%**	**49.09%**	**75.91%**
R(2+1)D	PGD	84.64%	95.57%	67.90%
	FGSM	87.55%	96.01%	84.00%
	TT	84.16%	94.68%	78.31%
	J-TLGA	**66.52%**	**73.59%**	**43.10%**

表3：迁移攻击性能。数值表示模型在不同迁移攻击下剩余的准确率（%）。数值越低，表示攻击效果越好。J-TLGA 具有更强的泛化能力。

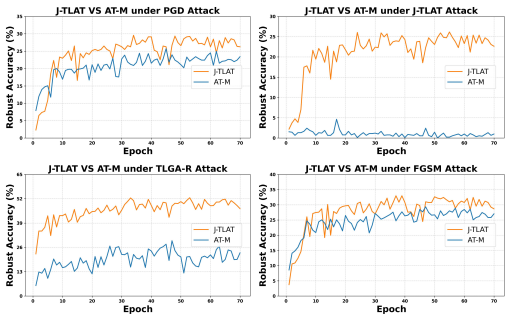


图7：在四种攻击设置下，对比 J-TLAT 和 AT-M 经 70 个训练轮次后的鲁棒准确率（%）训练动态。J-TLAT 在整个训练过程中始终表现出更高且更稳定的鲁棒性，并保持强大的防御能力。

更多结果。图3展示了不同攻击在各种模型上的迁移攻击性能。实验表明，在黑盒迁移设置下，J-TLGA 优于其他方法，表现出更强的泛化能力。图7显示，AT-M 在训练过程中性能崩溃，而 J-TLAT 始终保持强大的防御能力。更多结果和详情请参见附录。

6 结论

混合专家架构在视频理解领域展现出巨大潜力。然而，目前针对视频混合专家架构的攻击与防御方法仍存在空白。为解决这些问题，我们提出 TLGA，专门攻击视频混合专家的核心组件，并进一步开发联合攻击方法 J-TLGA，以揭示视频混合专家的阿喀琉斯之踵。基于这些弱点，我们提出 J-TLAT联合训练方法，从组件级到整体级分层提升混合专家架构的对抗鲁棒性。大量实验验证了我们方法的有效性，为未来视频混合专家领域的对抗攻击与防御研究奠定了基础。

参考文献

Bowen Chen、Keyan Chen、Mohan Yang、Zhengxia Zou 和 Zhenwei Shi。《用于遥感图像超分辨率的异构专家混合》。*IEEE*地球科学与遥感快报，2025a。

Chubin Chen、Sujie Hu、Jiashu Zhu、Meiqi Wu、Jintao Chen、Yanxun Li、Nisha Huang、Chengyu Fang、Jiahong Wu、Xiangxiang Chu 等人。通过扩散强化学习中的方向性解耦对齐，抑制偏好模式坍塌。*arXiv* 预印本 *arXiv:2512.24146*，2025b。

Chubin Chen、Jiashu Zhu、Xiaokun Feng 等人。S²-引导：用于训练 e 扩散模型增强的随机自引导。*arXiv* 预印本 *arXiv:2508.12880*，2025c。

Fanglin Chen、Weihang Wang、Huiyuan Yang、Wenjie Pei 和 Guangming Lu。用于监控视频诊断的多尺度特征融合。知识型系统，第 108103 页，2022年。

Shaoshen Chen、Peijie Huang、Zhanbiao Zhu、Yexing Zhang 和 Yuhong Xu。面向零样本跨域槽位填充的鲁棒多原型感知集成。*IEEE*信号处理快报，2024。

Shaoshen Chen、Yangning Li、Zishan Xu、Yinghui Li、Xin Su、Zifei Shan 和 Hai-Tao Zheng。DAST：通过软令牌动态分配实现大语言模型的上下文感知压缩。载于计算语言学协会年会，2025d。网址 https://api. semanticscholar.org/CorpusID:276409340。

Shuyu Cheng、Yibo Miao、YinpengDong、XiaoYang、Xiao-ShanGao 和JunZhu。通过函数先验引导的贝叶斯优化实现高效黑盒对抗攻击。载于第*41*届国际机器学习会议论文集，第 8163–8183 页，2024a。

Ying Cheng、Yang Li、Junjie He 和 Rui Feng。用于视听学习的专家混合。神经信息处理系统进展，37:219–243，2024b。

Francesco Croce 和 Matthias Hein。通过多种无参数攻击的集成可靠地评估对抗鲁棒性。载于国际机器学习会议，第 2206–2216 页。PMLR，2020年。

William Fedus、Barret Zoph 和 Noam Shazeer。Switch Transformer：利用简单高效的稀疏性扩展至万亿参数模型。机器学习研究杂志，23(120):1–39，2022年。

Christoph Feichtenhofer、Haoqi Fan、Jitendra Malik 和 何恺明。用于视频识别的 SlowFast 网络。载于*IEEE/CVF* 国际计算机视觉大会论文集，第 6202–6211 页，2019年。

Ian J Goodfellow、Jonathon Shlens 和 Christian Szegedy。阐释并利用对抗样本。*arXiv* 预印本 *arXiv:1412.6572*，2014年。

Feijiang Han。《超越检测：面向细粒度 Webshell 家族分类的表征学习综合基准与研究》。载于 *AAAI* 人工智能会议论文集（*AAAI-26*），2026 年。网址 https://arxiv.org/abs/2512.05288。

Feijiang Han、Hengtao Cui、Licheng Guo、Zelong Wang 和 Zhiyuan Lyu。《三思而后读：通过分步阅读缓解大语言模型理解失败》。载于 *IEEE* 国际声学、语音与信号处理会议论文集（*ICASSP*），2026a 年。网址 https://arxiv.org/abs/2504.09402。

Feijiang Han、Zelong Wang、Bowen Wang、Xinxin Liu、Skyler Cheung、Delip Rao、Chris Callison-Burch 和 Lyle Ungar。《Latex2layout：用于布局检测的高保真、可扩展文档布局标注流水线》。发表于*AAAI* 人工智能会议论文集（*AAAI-26*），2026b。网址：https://www.cis.upenn.edu/~ccb/publications/ latex2layout.pdf。

Feijiang Han、Xiaodong Yu、Jianheng Tang、Delip Rao、Weihua Du 和 Lyle Ungar。《Zerotuning：释放初始词元的力量，无需训练即可增强大语言模型》。发表于第十四届国际学习表征会议（*ICLR 2026*），2026c。网址：https://openreview.net/forum?id=EdkQ14FiiO。Qiao Han、Yiteng Zhai、Yu Qin、Yao Yang 等。《增强混合专家网络的"免疫力"，以实现对抗防御》。*arXiv* 预印本 *arXiv:2402.18787*，2024。

Kensho Hara、Hirokatsu Kataoka 和 Yutaka Satoh。时空 3D 卷积神经网络能否追溯 2D 卷积神经网络和 ImageNet 的发展历程？载于*IEEE* 计算机视觉与模式识别会议论文集，第 6546–6555 页，2018。

Shaofeng He、Qiu Cheng、Yu Huai、Zhongke Zhu 和 Jie Ding。用于遥感图像语义分割的混合专家模型。载于国际图像处理与人工智能会议（*ICIPAI 2024*），第 13213 卷，第 478–483 页。SPIE，2024。

Min Huang、Huimin Qian、Yi Han 和 Wenbo Xiang。用于视频中人体活动识别的 R (2+ 1) d 双流卷积神经网络。载于*2021*年第*40*届中国控制会议（*CCC*），第 7932–7937 页。IEEE，2021年。

Robert A Jacobs、Michael I Jordan、Steven J Nowlan 和 Geoffrey E Hinton。自适应局部专家混合。神经计算，3(1):79–87，1991。

Gagan Jain、Nidhi Hegde、Aditya Kusupati、Arsha Nagrani、Shyamal Buch、Prateek Jain、Anurag Arnab 和 Sujoy Paul。嵌套专家混合：视觉词元的自适应处理。神经信息处理系统进展，37:58480–58497，2024。

Teng Jin、Ziwen He、Zhangjie Fu、Songping Wang、Yueming Lyu 和 Yufei Shi。频域分布式扰动：迈向查询高效的黑盒对抗视频攻击。载于第 *33* 届 *ACM* 国际多媒体会议论文集，第 8360–8368 页，2025年。

Michael I Jordan 和 Robert A Jacobs。分层专家混合与期望最大化算法。神经计算，6(2):181–214，1994。

Kaleab A Kinfu 和 Ren´e Vidal。视频分类中对抗训练的分析与扩展。载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集，第 3416–3425 页，2022年。

Lingdong Kong、Xiang Xu、Jiawei Ren、Wenwei Zhang、Liang Pan、Kai Chen、Wei Tsang Ooi 和 Ziwei Liu。面向自动驾驶的多模态数据高效三维场景理解。*IEEETransactions on Pattern Analysis and MachineIntelligence*，2025年。

Hildegard Kuehne、Hueihan Jhuang、Est´ıbaliz Garrote、Tomaso Poggio 和 Thomas Serre。HMDB：用于人体动作识别的大型视频数据库。载于 *2011*年国际计算机视觉会议，第 2556–2563 页。IEEE，2011年。

Ji Lin、Chuang Gan 和 Song Han。TSM：用于高效视频理解的时序移位模块。载于 *IEEE*国际计算机视觉会议论文集，2019年。

Renyang Liu、Xin Jin、Dongting Hu、Jinhong Zhang、Yuanyu Wang、Jin Zhang 和 Wei Zhou。Dualflow：通过流场和归一化流模型生成难以察觉的对抗样本。《神经机器人学前沿》，17：1129720，2023a年。

Renyang Liu、Jinhong Zhang、Haoran Li、Jin Zhang、Yuanyu Wang 和 Wei Zhou。Aflow：在极度噪声受限环境下开发对抗样本。载于国际信息与通信安全会议，第502–518页。施普林格，2023b年。Renyang Liu、Wei Zhou、Tianwei Zhang、Kangjie Chen、Jun Zhao 和 Kwok-Yan Lam。利用条件扩散模型增强针对深度神经网络的黑盒攻击。*IEEE*信息取证与安全汇刊，19：5207–5219，2024。

Renyang Liu、Wenjie Feng、Tianwei Zhang、Wei Zhou、Xueqi Cheng 和 See-Kiong Ng。《重新思考图像生成模型中的机器遗忘》。载于*2025*年*ACM SIGSAC*计算机与通信安全会议论文集，第 993–1007 页，2025年。

Shao-Yuan Lo、Jeya Maria Jose Valanarasu 和 Vishal M Patel。《对抗视频的过完备表示》。载于*2021*年*IEEE*国际图像处理会议（*ICIP*），第 1939–1943 页。IEEE，2021年。

Yueen Ma、Yuzheng Zhuang、Jianye Hao 和 Irwin King。《3d-moe：一种通过整流流实现三维视觉和姿态扩散的混合专家多模态大语言模型》。*arXiv* 预印本 *arXiv:2501.16698*，2025年。

Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。《迈向能够抵御对抗攻击的深度学习模型》。*arXiv* 预印本 *arXiv:1706.06083*，2017年。

Hui Meng、Haochen Zhao、Deqian Yang、Songping Wang 和 Zhenpeng Li。《由粗到细的分割方法实现腹部CT中器官和肿瘤的精准高效分割》。载于*MICCAI*快速低资源半监督腹部器官分割挑战赛，第 115–129 页。施普林格，2023年。

Yibo Miao、Yinpeng Dong、Jun Zhu 和 Xiao-Shan Gao。在物理世界中的等距三维对抗样本。神经信息处理系统进展，35：19716–19731，2022年。

Yibo Miao、Yinpeng Dong、Jinlai Zhang、Lijia Yu、Xiao Yang 和 Xiao-Shan Gao。从傅里叶视角提升三维点云识别的鲁棒性。神经信息处理系统进展，37：68183–68210，2024。

Patricia Pauli、Anne Koch、Julian Berberich、Paul Kohler 和 Frank Allg¨ower。利用利普希茨界训练鲁棒神经网络。*IEEE*控制系统快报，6：121–126，2021年。

Ala Shaabana、Zahra Gharaee 和 Paul Fieguth。使用专家混合进行视频关系检测。*IEEE Access*，11：26048–26058，2023年。

Noam Shazeer、Azalia Mirhoseini、Krzysztof Maziarz、Andy Davis、Quoc Le、Geoffrey Hinton 和 Jeff Dean。超大规模神经网络：稀疏门控专家混合层。载于国际学习表征会议（*ICLR*），2017年。

Khurram Soomro、Amir Roshan Zamir 和 Mubarak Shah。UCF101：来自野外视频的101个人类动作类别数据集。*arXiv* 预印本 *arXiv:1212.0402*，2012年。

Florian Tram`er、Alexey Kurakin、Nicolas Papernot、Ian Goodfellow、Dan Boneh 和 Patrick Mc-Daniel。集成对抗训练：攻击与防御。*arXiv* 预印本 *arXiv:1705.07204*，2017年。

Mathurin Videau、Alessandro Leite、Marc Schoenauer 和 Olivier Teytaud。图像分类中的专家混合：最佳平衡点在哪里？*arXiv* 预印本 *arXiv:2411.18322*，2024。

Songping Wang、Hanqing Liu 和 Haochen Zhao。用于提升视频攻击可迁移性的公共领域定位器。载于*2024 IEEE*多媒体与博览会国际会议（*ICME*），第 1–6 页。IEEE，2024a。

Songping Wang、Haoxiang Rao、Xiantao Hu、Yueming Lyu 和 Caifeng Shan。一种用于多模态动作识别的有效端到端解决方案。载于模式识别国际会议，第 324–338 页。施普林格，2024b。

Songping Wang、Hanqing Liu、Yueming Lyu、Xiantao Hu、Ziwen He、Wei Wang、Caifeng Shan 和 Liang Wang。在频域中对视频进行具有由弱到强的时空一致性的快速对抗训练。*IEEE*信息取证与安全汇刊，21：681–696，2025a。

Songping Wang、Yueming Lyu、Shiqi Liu、Ning Li、Tong Tong、Hao Sun 和 Caifeng Shan。反审美：保护面部隐私，抵御定制文本到图像合成。*arXiv* 预印本 *arXiv:2504.12129*，2025b。

Songping Wang、Rufan Qian、Yueming Lyu、Qinglong Liu、Linzhuang Zou、Jie Qin、Songhua Liu 和 Caifeng Shan。《Runawayevil：使图像到视频生成模型越狱》。*arXiv* 预印本 *arXiv:2512.06674*，2025c。

Songping Wang、Xinquan Yue、Yueming Lyu 和 Caifeng Shan。《探索 Kolmogorov-Arnold 网络之间的对抗迁移性》。*arXiv* 预印本 *arXiv:2503.06276*，2025d。Xin Wang、Fisher Yu、Lisa Dunlap、Yi-An Ma、Ruth Wang、Azalia Mirhoseini、Trevor Darrell 和 Joseph E Gonzalez。《通过浅层嵌入实现深度专家混合》。载于人工智能中的不确定性，第 552–562 页。PMLR，2020。Xingxing Wei、Jun Zhu、Sha Yuan 和 Hang Su。《视频稀疏对抗扰动》。载于*AAAI* 人工智能会议论文集，第 33 卷，第 8973–8980 页，2019。Xingxing Wei、Songping Wang 和 Huanqian Yan。《通过视频上的对抗性时空聚焦高效评估鲁棒性》。*IEEE* 模式分析与机器智能汇刊，45(9):10898–10912，2023。Zhipeng Wei、Jingjing Chen、Xingxing Wei、Linxi Jiang、Tat-Seng Chua、Fengfeng Zhou 和 Yu-Gang Jiang。《针对视频识别模型的启发式黑盒对抗攻击》。载于*AAAI* 人工智能会议论文集，第 34 卷，第 12338–12345 页，2020。Zhipeng Wei、Jingjing Chen、Zuxuan Wu 和 Yu-Gang Jiang。《通过时间平移提升视频对抗样本的迁移性》。载于*AAAI* 人工智能会议论文集，第 36 卷，第 2659–2667 页，2022。Eric Wong、Leslie Rice 和 J Zico Kolter。《快速胜过免费：重新审视对抗训练》。*arXiv* 预印本 *arXiv:2001.03994*，2020。Shiwei Wu、Joya Chen、Kevin Qinghong Lin、Qimeng Wang、Yan Gao、Qianli Xu、Tong Xu、Yao Hu、Enhong Chen 和 Mike Zheng Shou。《VideoLLM-MoD：通过深度混合实现高效的视频语言流式处理与视觉计算》。神经信息处理系统进展，37:109922–109947，2024。Cihang Xie、Jianyu Wang、Zhishuai Zhang、Zhou Ren 和 Alan Yuille。《通过随机化减轻对抗效应》。*arXiv* 预印本 *arXiv:1711.01991*，2017。Chenyu Yi、Siyuan Yang、Yufei Wang、Haoliang Li、Yap-Peng Tan 和 Alex C Kot。《面向鲁棒视频分类的时间一致性测试时优化》。*arXiv* 预印本 *arXiv:2302.14309*，2023。Mingyong Yin、Yixiao Xu、Teng Hu 和 Xiaolei Liu。《一种基于视频增强的鲁棒对抗样本攻击》。应用科学，13(3):1914，2023。Bohang Zhang、Du Jiang、Di He 和 Liwei Wang。《重新思考 Lipschitz 神经网络与可认证鲁棒性：布尔函数视角》。神经信息处理系统进展，35:19398–19413，2022。Xu Zhang、Kaidi Xu、Ziqing Hu 和 Ren Wang。《优化专家混合中的鲁棒性与准确率：一种双模型方法》。*arXiv* 预印本 *arXiv:2502.06832*，2025。Yihua Zhang、Ruisi Cai、Tianlong Chen、Guanhua Zhang、Huan Zhang、Pin-Yu Chen、Shiyu Chang、Zhangyang Wang 和 Sijia Liu。《面向卷积神经网络的鲁棒专家混合训练》。载于*IEEE/CVF* 国际计算机视觉大会论文集，第 90–101 页，2023。Zhiqin Zhou、Songping Wang 和 Caifeng Shan。《基于时间差分 Transformer 的多模态视频心率估计》。2025。Monty-Maximilian Z¨uhlke 和 Daniel Kudenko。《从 Lipschitz 微积分视角看神经网络的对抗鲁棒性：综述》。*ACM* 计算调查，57(6):1–41，2025。