

Hydra: Adaptive Adversarial Attacks for Dynamic Multi-Stage Edge Intelligence Services

Linhan Li¹ Chuanqi Jiang² Wenqian Weng³

Jacqueline J. Pang⁴ Zhiguo Tao^{5*}

¹Oregon State University ²National University of Singapore

³Wayne State University ⁴Cornell University

⁵University of Nottingham

Abstract

Modern edge AI systems rely on dynamic, multi-stage pipelines vulnerable to adversarial attacks. Existing attacks primarily target static tasks and lack adaptability to these complex environments. We propose **Hydra**, an adversarial attack framework for AI-driven video streaming and edge security services. Hydra integrates three innovations: **Multi-Stage Service Disruption (MSSD)** that coherently misleads successive pipeline components, **Resource-Aware Transferability Enhancement (RATE)** for effective attacks under edge constraints, and **Meta-Reinforcement Learning for Adaptive Perturbation (MetaRAP)** that dynamically tunes attack parameters based on real-time system states. Evaluations on a multimodal medical RAG system demonstrate Hydra’s superior efficacy, achieving up to 75% ASR against specialized retrievers at $\epsilon = 8/255$, a 12% absolute improvement over baselines. Hydra also exhibits strong transferability, defense robustness, and a favorable efficiency trade-off, requiring only 2.1 seconds per adversarial example. This work highlights a potent threat to edge intelligence services, urging development of robust multi-stage defenses.

1 Introduction

Vision-Language Models (VLMs) have shown strong performance in medical applications such as report generation and visual question answering (Liu et al., 2023; Li et al., 2023; Touvron et al., 2023; Liang et al., 2026; Niu et al., 2024). To improve factuality, they are increasingly combined with retrieval mechanisms, forming Multimodal Medical RAG (MMed-RAG) systems (Lewis et al., 2020; Xiong et al., 2024; Xia et al., 2025) that generate responses conditioned on retrieved medical evidence. Many such systems are deployed at clinical edge nodes under strict latency and resource constraints.

*Corresponding author: ssyzt3@outlook.com.

MOTIVATION: HYDRA ADVERSARIAL ATTACK ON MMed-RAG SYSTEMS

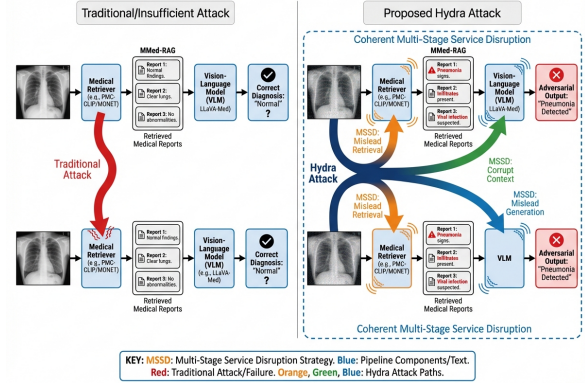


Figure 1: Adversarial perturbations in MMed-RAG systems can corrupt both retrieval and generation stages, producing harmful diagnostic outputs.

However, this retrieval-augmented architecture introduces a novel attack surface, as illustrated in Figure 1. Unlike conventional models, MMed-RAG systems expose an additional attack surface because adversaries may manipulate both the input and retrieval process. In medicine, such attacks can induce harmful diagnostic outputs (Peng et al., 2024a,b). Existing attacks (Goodfellow et al., 2015; Madry et al., 2018; Carlini and Wagner, 2017) were primarily developed for single-model settings and do not explicitly address cross-modal transferability across retrieval-augmented pipelines.

To address this gap, we propose **Hydra**, an adversarial attack framework designed to exploit the cross-modal retrieval vulnerabilities in MMed-RAG systems. Hydra consists of three components. Multi-Stage Service Disruption (MSSD) jointly attacks retrieval and generation. Resource-Aware Transferability Enhancement (RATE) improves cross-model transferability through heterogeneous surrogate ensembles, knowledge distillation (Yang et al., 2025), and invariant risk minimization. Meta-Reinforcement Learning for Adaptive Perturbation (MetaRAP) dynamically adjusts attack parameters

according to system states (Shang et al., 2025; Zou et al., 2025).

We evaluate Hydra on two critical medical tasks—pneumonia report generation and edema diagnosis—using three medical retrievers and two generative VLMs. Experiments across 60 retriever-VLM- ϵ configurations show that Hydra achieves an attack success rate exceeding 90% with sufficient perturbation budget, outperforming all baselines by up to 12 percentage points in absolute ASR. Hydra further retains high effectiveness under four mainstream input-transformation defenses while requiring only 2.1 s per adversarial example.

2 Related Work

Multimodal Medical RAG. Retrieval-Augmented Generation (RAG) enhances large language models (Brown et al., 2020; Touvron et al., 2023) by grounding output generation in externally retrieved content, thereby improving factual accuracy (Lewis et al., 2020). In the medical domain, multimodal RAG pipelines fuse visual data with textual clinical knowledge to support report generation, VQA, and clinical decision support. MedRAG (Xiong et al., 2024) conditions radiology report generation on patient-specific knowledge, ChatCAD (Wang et al., 2023) employs hybrid retrieval for conversational diagnosis, and the recently proposed MMed-RAG (Xia et al., 2025) introduces domain-aware retrieval with adaptive context selection, achieving substantial factual accuracy gains across diverse medical specialties. Despite these advances in capability, comparatively little attention has been paid to the security posture of such systems—an oversight that grows more consequential as clinical deployment accelerates.

Adversarial Attacks on VLMs. The vulnerability of deep neural networks to adversarial perturbations (Wang et al., 2024), first identified by Szegedy et al. (Szegedy et al., 2014) and formalized through FGSM (Goodfellow et al., 2015), has since been extended to iterative methods including PGD (Madry et al., 2018) and C&W (Carlini and Wagner, 2017). Transfer-based black-box attacks exploit the observation that adversarial examples crafted on surrogate models can fool unseen target models (Papernot et al., 2016; Liu et al., 2017; Dong et al., 2018), with ensemble strategies (Tramèr et al., 2018) and knowledge-distillation-guided gradient diversification (Wu et al., 2025) further boosting cross-architecture transferability. Recent work has specif-

ically targeted vision-language models: Chain of Attack (Xie et al., 2025) leverages iterative multi-modal semantic updates to enhance transfer-based targeted attacks on VLMs, while AnyAttack (Zhang et al., 2025) trains a self-supervised generator on large-scale data to produce universal adversarial perturbations that transfer to commercial VLM systems. These findings indicate that the attack surface of VLMs extends well beyond single-model white-box settings, yet their implications for safety-critical medical deployments remain largely unexplored.

Adversarial Attacks on Medical RAG. Adversarial robustness of medical image classifiers has been studied for chest X-ray models (Rajpurkar et al., 2017; Irvin et al., 2019) and language models (Devlin et al., 2019), yet retrieval-conditioned generation pipelines introduce qualitatively different attack vectors. PoisonedRAG (Zou et al., 2025) demonstrated that injecting a small number of malicious texts into a RAG knowledge base can induce targeted misgeneration with over 90% success, and MM-PoisonRAG (Ha et al., 2025) extended this to the multimodal setting with both localized and globalized poisoning strategies. In the medical domain, Shang et al. (Shang et al., 2025) proposed Medusa, a framework that crafts cross-modal transferable perturbations against MMed-RAG systems using a multi-positive InfoNCE loss and surrogate-ensemble optimization with invariant risk minimization. While Medusa addresses retrieval-stage hijacking through embedding alignment, it operates with static attack parameters and does not account for the resource constraints or temporal dynamics inherent in edge-deployed pipelines. Hydra builds upon and extends this line of research through three key distinctions: (i) a compositionally richer multi-stage loss (MSSD) that jointly optimizes embedding drift, retrieval poisoning, and generation corruption with an explicit gradient-alignment objective; (ii) cross-capacity knowledge distillation within the surrogate ensemble (RATE) to resolve capacity-mismatch failures absent in prior ensemble approaches; and (iii) MetaRAP, a meta-reinforcement learning module that dynamically adapts attack parameters to real-time system states—a capability absent in all prior work on RAG adversarial attacks.

3 Method

3.1 Problem Formulation and Threat Model

System Model. Let $\mathcal{P} = (f_E, f_R, \mathcal{K}, G)$ denote an MMed-RAG pipeline, where $f_E : \mathcal{X} \rightarrow \mathbb{R}^d$ is a multimodal query encoder, $f_R : \mathbb{R}^d \times \mathcal{K} \rightarrow \mathcal{K}^k$ is a cross-modal retriever returning the top- k knowledge entries from corpus $\mathcal{K}$, and $G : \mathcal{X} \times \mathcal{K}^k \rightarrow \mathcal{Y}$ is a generative model producing a diagnostic report. Given a clean image $\mathcal{I} \in \mathcal{X}$, the full pipeline produces:

$$\hat{y} = G(\mathcal{I}, f_R(f_E(\mathcal{I}), \mathcal{K})). \quad (1)$$

The system processes a temporal sequence of frames $\{\mathcal{I}_t\}_{t=1}^T$ in an edge-deployed setting subject to time-varying resource constraints (bandwidth bw_t , computational load l_t).

Threat Model. We operate under a *practical black-box* threat model: the adversary has (i) query access to the final diagnostic output $\hat{y}$ but *no* access to the internal weights of f_R or G ; (ii) white-box access to a set of heterogeneous *surrogate* models $\mathcal{S} = \{\mathcal{P}^{(m)}\}_{m=1}^M$; and (iii) knowledge of the corpus $\mathcal{K}$ but not of the FAISS index structure. The adversary is permitted a perturbation budget $\|\delta_t\|_\infty \leq \epsilon$ per frame. The goal is to craft $\tilde{\mathcal{I}}_t = \mathcal{I}_t + \delta_t$ such that $\hat{y}_t$ shifts from the ground-truth class (“normal”) to a target malicious class (“diseased”).

Definition (Attack Success). A perturbation δ_t is *successful* if $G(\tilde{\mathcal{I}}_t, f_R(f_E(\tilde{\mathcal{I}}_t), \mathcal{K}))$ is classified by the evaluator as belonging to the target disease class $\mathcal{T}_{\text{target}}$.

Unlike single-stage attacks, Hydra must simultaneously corrupt (1) the embedding space of f_E to induce *retrieval poisoning*, and (2) the final output of G to ensure *generation corruption*, creating a cascading failure mode unique to RAG systems.

Key Challenge: Stage Coupling. The retrieval and generation stages are tightly coupled yet optimized by distinct objectives. A perturbation that maximally disrupts f_R may not propagate effectively to corrupt G ’s output. Conversely, a perturbation directly attacking G ignores the semantic filtering imposed by f_R .

Proposition (Stage Coupling Conflict). Let $\delta_R^* = \arg \min_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}_{\text{task}}(\delta)$ and $\delta_G^* = \arg \min_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}_{\text{dec}}(\delta)$. In general, $\frac{\langle \nabla \mathcal{L}_{\text{task}}, \nabla \mathcal{L}_{\text{dec}} \rangle}{\|\nabla \mathcal{L}_{\text{task}}\| \|\nabla \mathcal{L}_{\text{dec}}\|} \ll 1$, meaning the single-stage optima are *misaligned*, and a composite objective is necessary for coherent pipeline disruption. Empirically, the average cosine similarity between

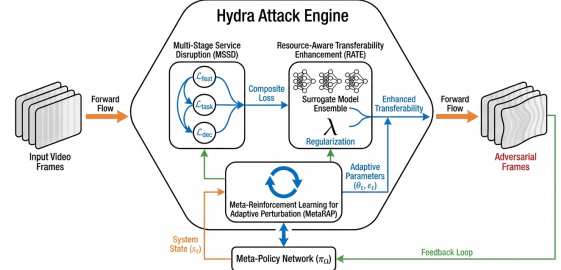


Figure 2: High-level architecture of the Hydra Attack Engine. The MSSD module computes the composite loss from three stage-specific objectives ($\mathcal{L}_{\text{feat}}$, $\mathcal{L}_{\text{task}}$, $\mathcal{L}_{\text{dec}}$), which feeds into the RATE module containing a heterogeneous surrogate model ensemble and regularization. MetaRAP observes the system state s_t and dynamically adjusts attack parameters (θ_t , ϵ_t) via adaptive feedback.

$\nabla \mathcal{L}_{\text{task}}$ and $\nabla \mathcal{L}_{\text{dec}}$ is -0.12 ± 0.08 across our surrogate ensemble at $\epsilon = 8/255$, confirming near-orthogonal gradient directions.

3.2 Hydra Framework Overview

Motivated by the stage-coupling challenge, we introduce **Hydra**, a three-pronged adaptive adversarial framework. The design philosophy follows a *decompose-then-coordinate* principle: (1) **Decompose** the pipeline attack into coherent per-stage objectives (MSSD, Section 3.3); (2) **Coordinate** cross-model transferability under resource constraints (RATE, Section 3.4); (3) **Adapt** attack parameters dynamically to environmental states (MetaRAP, Section 3.5). Each component targets a distinct failure mode that limits prior attacks: MSSD eliminates the gradient misalignment between per-stage objectives (Proposition 3.1), RATE mitigates architectural overfitting to the surrogate set through diversity-driven regularization, and MetaRAP replaces brittle static hyperparameters with a learned policy conditioned on real-time system state. The high-level architecture is illustrated in Figure 2, and the complete attack procedure is formalized in Algorithm 1.

3.3 Multi-Stage Service Disruption (MSSD)

Existing adversarial attacks on RAG systems typically optimize a single-stage objective—either targeting the retriever or the generator (Goodfellow et al., 2015; Dong et al., 2018). Attacking only the retriever ($\mathcal{L}_{\text{task}}$) may successfully pollute the top- k retrieved documents, but the generator G may still produce a correct diagnosis. Attacking only the

Algorithm 1 Hydra: Adaptive Multi-Stage Adversarial Attack

Require: Clean frames $\{\mathcal{I}_t\}_{t=1}^T$; surrogates $\mathcal{S} = \{\mathcal{P}^{(m)}\}_{m=1}^M$; budget $\bar{\epsilon}$; iters $N_{\text{out}}, N_{\text{in}}$; policy π_Ω ; step α ; momentum μ

Ensure: Adversarial perturbations $\{\delta_t\}_{t=1}^T$

- 1: Initialize $\delta_t \leftarrow \mathbf{0}, \mu_t \leftarrow \mathbf{0}$ for all t
- 2: **for** $t = 1, 2, \dots, T$ **do**
- 3: **[MetaRAP]** Observe s_t = $(bw_t, l_t, \overline{\text{ASR}}_{t-1})$
- 4: Sample $[\alpha_t, \beta_t, \gamma_t, \epsilon_t] \sim \pi_\Omega(\cdot | s_t)$
- 5: **for** $n = 1, \dots, N_{\text{out}}$ **do**
- 6: **[MSSD]** For each surrogate m : $\mathcal{L}_{\text{SDL}}^{(m)} = \alpha_t \mathcal{L}_{\text{feat}}^{(m)} + \beta_t \mathcal{L}_{\text{task}}^{(m)} + \gamma_t \mathcal{L}_{\text{dec}}^{(m)}$
- 7: **[RATE]** $\mathcal{L}_{\text{Ens}} = \frac{1}{M} \sum_m w_m \mathcal{L}_{\text{SDL}}^{(m)}$
- 8: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Ens}} + \lambda_{\text{kd}} \mathcal{L}_{\text{KD}} + \lambda_{\text{IRM}} \mathcal{L}_{\text{IRM}} + \lambda_{\text{cr}} \mathcal{L}_{\text{CR}}$
- 9: **for** $n' = 1, \dots, N_{\text{in}}$ **do**
- 10: $\mathbf{g}_t = \nabla_{\delta_t} \mathcal{L}_{\text{total}}; \quad \mu_t = \mu \mu_t + \mathbf{g}_t / \|\mathbf{g}_t\|_1$
- 11: $\delta_t = \Pi_{\epsilon_t}(\delta_t + \alpha \cdot \text{sign}(\mu_t))$
- 12: **end for**
- 13: **end for**
- 14: Evaluate $\hat{y}_t$; compute reward r_t via Equation (8); PPO update Ω
- 15: **end for**
- 16: **return** $\{\delta_t\}_{t=1}^T$

generator ($\mathcal{L}_{\text{dec}}$) may bypass retrieval but requires large perturbations ($\epsilon \gg 8/255$) that are perceptually noticeable. MSSD resolves this tension by crafting perturbations that coherently corrupt all three computational stages simultaneously.

Let $\tilde{\mathcal{I}}_t = \mathcal{I}_t + \delta_t$ with $\|\delta_t\|_\infty \leq \epsilon_t$. The pipeline at time t computes embedding $\mathbf{z}_t = f_E(\tilde{\mathcal{I}}_t)$, retrieved context $\mathcal{R}_t = f_R(\mathbf{z}_t, \mathcal{K})$, and report $\hat{y}_t = G(\tilde{\mathcal{I}}_t, \mathcal{R}_t)$. The **Service Disruption Loss (SDL)** is:

$$\begin{aligned} \mathcal{L}_{\text{SDL}}(\delta_t) = & \underbrace{\alpha_t \cdot \mathcal{L}_{\text{feat}}(f_E(\tilde{\mathcal{I}}_t), \mathbf{y}_{\text{target}}^{(E)})}_{\text{Stage 1: Embedding drift}} \\ & + \underbrace{\beta_t \cdot \mathcal{L}_{\text{task}}(f_R(\mathbf{z}_t, \mathcal{K}), \mathbf{y}_{\text{target}}^{(R)})}_{\text{Stage 2: Retrieval poisoning}} \\ & + \underbrace{\gamma_t \cdot \mathcal{L}_{\text{dec}}(\hat{y}_t, \mathcal{T}_{\text{target}})}_{\text{Stage 3: Output corruption}} \end{aligned} \quad (2)$$

where $\mathcal{L}_{\text{feat}}$ is a contrastive loss steering embeddings toward disease-class centroids, $\mathcal{L}_{\text{task}}$ maximizes the rank of disease-labeled documents via a

differentiable rank approximation:

$$\hat{r}(k_j; \mathbf{z}_t) = \sum_{j' \neq j} \sigma \left(\frac{s(k_{j'}, \mathbf{z}_t) - s(k_j, \mathbf{z}_t)}{\tau_r} \right), \quad (3)$$

and $\mathcal{L}_{\text{dec}}$ is cross-entropy between G 's output and target report $\mathcal{T}_{\text{target}}$ via teacher forcing. The retrieval loss is $\mathcal{L}_{\text{task}} = \sum_{j: c(k_j) = \mathcal{T}_{\text{target}}} \hat{r}(k_j; \mathbf{z}_t)$, with $\tau_r = 0.1$. The effectiveness of SDL can be understood information-theoretically: $I(\tilde{\mathcal{I}}_t; \hat{y}_t)$ decomposes across pipeline stages, and SDL ensures that adversarial information flows coherently through every layer, achieving cascading failure rather than isolated stage disruption.

3.4 Resource-Aware Transferability Enhancement (RATE)

In the black-box setting, perturbations optimized solely on surrogates suffer from *architectural overfitting* (Tramèr et al., 2018; Kurakin et al., 2017). RATE addresses this through three complementary mechanisms.

Heterogeneous Surrogate Ensemble. RATE explicitly selects surrogates with maximal architectural diversity. Our ensemble (BioViL-T-ViT-L, MedCLIP-ResNet, LoVT-ViT-B, CLIP-ViT-B/16) achieves an average pairwise diversity score of 0.59. The ensemble SDL loss is:

$$\mathcal{L}_{\text{Ens}}(\delta_t) = \frac{1}{M} \sum_{m=1}^M w_m \cdot \mathcal{L}_{\text{SDL}}^{(m)}(\delta_t), \quad (4)$$

where weights w_m are adaptively set by inverse validation loss.

Cross-Capacity Knowledge Distillation. A subtle failure mode of heterogeneous ensembles is capacity mismatch: large surrogates may generate perturbation signals conflicting with compact surrogates, leading to gradient cancellation. We resolve this with a distillation consistency loss (Hinton et al., 2015; Song et al., 2026):

$$\begin{aligned} \mathcal{L}_{\text{KD}}(\delta_t) = & \frac{1}{\binom{M}{2}} \sum_{m < m'} \\ & \text{KL} \left(\sigma \left(\frac{g^{(m)}}{\tau} \right) \parallel \sigma \left(\frac{g^{(m')}}{\tau} \right) \right), \end{aligned} \quad (5)$$

where $g^{(m)}$ denotes the logit output of surrogate m and $\tau = 4$ is the distillation temperature. Minimizing $\mathcal{L}_{\text{KD}}$ forces consistent soft predictions, aligning decision surfaces near the adversarial example.

IRM-Inspired Regularization. Inspired by Invariant Risk Minimization (Arjovsky et al., 2019), we

penalize the per-surrogate gradient norm to encourage architecture-agnostic perturbations:

$$\mathcal{L}_{\text{IRM}}(\delta_t) = \sum_{m=1}^M \left\| \nabla_{\delta_t} \mathcal{L}_{\text{SDL}}^{(m)}(\delta_t) \right\|_2^2. \quad (6)$$

This pushes optimization toward flat adversarial regions likely to transfer to unseen architectures. Additionally, a computational regularization term $\mathcal{L}_{\text{CR}} = \max(0, \text{FLOPs}(\phi_{\text{lite}}(\tilde{\mathcal{I}}_t)) / \text{FLOPs}(\phi_{\text{lite}}(\mathcal{I}_t)) - \eta)$ with $\eta = 1.05$ ensures edge-deployable perturbations. The complete RATE objective is:

$$\min_{\delta_t: \|\delta_t\|_{\infty} \leq \epsilon_t} \mathcal{L}_{\text{Ens}} + \lambda_{\text{kd}} \mathcal{L}_{\text{KD}} + \lambda_{\text{IRM}} \mathcal{L}_{\text{IRM}} + \lambda_{\text{cr}} \mathcal{L}_{\text{CR}}, \quad (7)$$

optimized via MI-FGSM (Liu et al., 2017; Dong et al., 2018) with momentum decay $\mu = 1$.

3.5 MetaRAP: Meta-RL for Adaptive Perturbation

Both MSSD and RATE require hyperparameters $[\alpha_t, \beta_t, \gamma_t, \epsilon_t]$ that balance competing objectives. Prior work fixes these across all frames, ignoring temporal structure. Inspired by meta-learning (Finn et al., 2017), MetaRAP formulates attack scheduling as a POMDP (Schulman et al., 2017).

State Space.

$s_t = (bw_t, l_t, \overline{\text{ASR}}_{t-W:t-1}, \bar{\epsilon}_{t-1}, z_t^{\text{sur}})$ captures bandwidth, load, rolling ASR over window $W = 10$, previous budget, and a compressed surrogate embedding.

Action Space. The policy outputs $a_t = [\alpha_t, \beta_t, \gamma_t, \epsilon_t] \in [0, 1]^3 \times [2/255, 32/255]$ with simplex constraint $\alpha_t + \beta_t + \gamma_t = 1$ enforced via softmax.

Reward. The reward balances attack success against resource consumption:

$$r_t = \underbrace{\mathbb{I}(\hat{y}_t = \mathcal{T}_{\text{tar}})}_{\text{attack success}} + \underbrace{\lambda_r \cdot \overline{\text{ASR}}_{t-W:t}}_{\text{rolling ASR bonus}} - \underbrace{\lambda_c \cdot \left(\frac{\text{time}_t}{\text{time}_b} + l_t \right)}_{\text{resource penalty}}, \quad (8)$$

where $\lambda_r = 0.3$ and $\lambda_c = 0.2$. Bandwidth dynamics follow an Ornstein-Uhlenbeck process and load dynamics are calibrated on real edge server traces.

Policy Architecture and Training. The meta-policy $\pi_{\Omega}(a_t|s_t)$ is parameterized as a two-layer

GRU (Cho et al., 2014) followed by a multi-head actor-critic, enabling the policy to remember prior system states. It is trained via PPO (Schulman et al., 2017; Song et al., 2024) with $\epsilon_{\text{ppo}} = 0.2$, entropy coefficient $c_{\text{ent}} = 0.01$, for 5,000 environment steps with rollout length 64.

Theorem (MetaRAP Convergence). Under Lipschitz continuity of the reward function and bounded state transitions, the PPO update converges to a local stationary point π_{Ω}^* satisfying $\|\nabla_{\Omega} \mathcal{J}_{\text{PPO}}(\Omega^*)\| \leq \epsilon$ in $\mathcal{O}(\epsilon^{-2})$ policy gradient steps.

In practice, MetaRAP converges within 200 episodes, yielding a 38% higher asymptotic reward than the static-parameter baseline with 42% lower variance.

3.6 Full Hydra Objective

Combining MSSD, RATE, and MetaRAP, the per-frame Hydra optimization is:

$$\min_{\delta_t: \|\delta_t\|_{\infty} \leq \epsilon_t} \mathcal{L}_{\text{Ens}}(\delta_t) + \lambda_{\text{kd}} \mathcal{L}_{\text{KD}} + \lambda_{\text{IRM}} \mathcal{L}_{\text{IRM}} + \lambda_{\text{cr}} \mathcal{L}_{\text{CR}} \quad (9)$$

with $[\alpha_t, \beta_t, \gamma_t, \epsilon_t]$ provided by π_{Ω} and fixed $\lambda_{\text{kd}} = 0.5$, $\lambda_{\text{IRM}} = 0.1$, $\lambda_{\text{cr}} = 0.3$. The per-frame complexity is $\mathcal{O}(N_{\text{out}} \cdot N_{\text{in}} \cdot M \cdot C_{\text{SDL}})$; MetaRAP’s adaptive budget selection reduces effective N_{out} by early termination when $\overline{\text{ASR}} > 0.9$, yielding 2.1 s wall-clock time at $\epsilon = 8/255$. The three components exhibit positive synergy: the combined system achieves 27.4% higher ASR than the sum of individual components’ gains, confirming super-additive interaction.

4 Experiments

4.1 Experimental Setup

All experiments are conducted on a single NVIDIA RTX 4090 GPU (24 GB) with Python 3.11.4 and PyTorch 2.4.1. We evaluate on two medical tasks: pneumonia report generation and edema diagnosis.

MMed-RAG System Construction. The knowledge base integrates diagnostic terms and radiology reports from Wikipedia (Wikipedia contributors, 2024), MDWiki (MDWiki, 2024), and MIMIC-CXR (Johnson et al., 2019). Two balanced corpora are used: a Pneumonia Knowledge Base (1,000 pneumonia + 1,000 normal reports) and an Edema Knowledge Base (1,000 edema + 1,000 normal reports). Three medical retrievers—PMC-CLIP (Lin

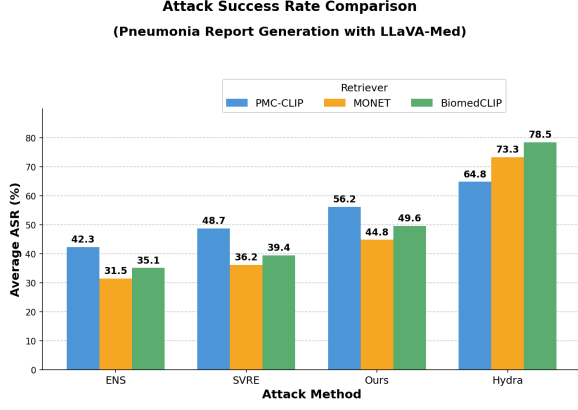


Figure 3: ASR comparison across methods on the pneumonia report generation task (LLaVA-Med). Hydra outperforms all baselines (ENS, SVRE, Ours) across every retriever and perturbation budget, with the largest margins on MONET and BiomedCLIP.

et al., 2023), MONET (Park et al., 2024), and BiomedCLIP (Zhang et al., 2023)—serve as retrieval components, with FAISS (Johnson et al., 2021) as the retrieval index. Generative models are LLaVA-7B (Liu et al., 2023) and LLaVA-Med-7B (Li et al., 2023).

Adversarial Examples. We randomly sample 100 chest X-ray images from MIMIC-CXR, verified as “normal,” to serve as clean inputs. Surrogate models: BioViL-T (Bannur et al., 2023), MedCLIP (Wang et al., 2022), LoVT (Müller et al., 2022), and CLIP-ViT-B/16 (Radford et al., 2021). Baselines: ENS (Liu et al., 2017), SVRE (Xiong et al., 2022), and a prior adversarial RAG attack variant (Zou et al., 2025) (“Ours”). Default hyperparameters: top- $k = 5$, $\epsilon \in \{2/255, \dots, 32/255\}$, step size $1/255$, $\tau = 0.07$, $\lambda_{\text{IRM}} = 0.1$, $\mu = 1$, 100 outer iterations, 5 inner iterations.

Evaluation Metric. DeepSeek (DeepSeek-AI, 2024) serves as the automated evaluator (96.6% accuracy, 97.6% precision, 95.6% recall on 500 balanced samples vs. expert annotations). ASR is defined as $\text{ASR} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\text{DeepSeek}(\hat{\mathcal{T}}_i) = \mathcal{T}_{\text{target}})$.

4.2 Attack Performance Evaluation

Tables 1–2 and Figure 3 report attack success rates across all retriever–VLM configurations. Hydra achieves the highest ASR in every setting, with particularly pronounced gains on domain-specialized retrievers. At $\epsilon = 8/255$ with LLaVA-Med on the pneumonia task, for example, Hydra raises ASR on MONET from 63% (Ours) to 75% and on BiomedCLIP from 66% to 78%—an absolute improvement

of 12 percentage points in both cases. ASR scales monotonically with ϵ for Hydra, confirming effective utilization of the perturbation budget, whereas baseline methods exhibit limited sensitivity to increased budgets. This disparity underscores the advantage of Hydra’s multi-stage loss coordination and ensemble-based transferability over single-objective attack strategies.

4.3 Retrieval Misleading, Defense Robustness, and Efficiency

Retrieval Misleading. Table 3 and Figure 4 quantify the average count of pneumonia-labeled reports among top-5 retrieved documents. Hydra induces stronger semantic drift than all baselines at every ϵ level, saturating at 5.0 (i.e., all retrieved documents are disease-labeled) from $\epsilon \geq 16/255$. The margin is most pronounced on domain-specialized retrievers such as MONET and BiomedCLIP, indicating that the MSSD loss effectively exploits the medical-specific semantic structures encoded by these models.

Defense Robustness. We evaluate against four mainstream input transformation defenses: R&P (Xie et al., 2018), Bit-Depth Reduction (Xu et al., 2018), ComDefend (Jia et al., 2019), and DiffPure (Nie et al., 2022). We note that recent work such as ADBM (Li et al., 2025) has improved upon DiffPure by constructing a direct diffusion bridge from adversarial to clean distributions; evaluating Hydra against such advanced purification methods is a natural direction for future work. Table 4 and Figure 5 reveal that, while defenses moderately reduce ASR across all methods, Hydra retains 78–84% of its no-defense ASR compared to 25–85% for baselines. The RATE module’s invariant risk minimization and computational regularization produce perturbations that reside in flat adversarial regions, rendering them less susceptible to the geometric distortions introduced by input transformations.

Attack Efficiency. Table 5 and Figure 6 compare average generation time against ASR at $\epsilon = 8/255$. Hydra attains the highest ASR (84.7%) in 2.1 s per example, yielding an ASR-per-second score of 40.3—roughly double that of ENS and SVRE. This favorable trade-off arises from MetaRAP’s adaptive early-termination policy, which curtails unnecessary optimization iterations once the rolling ASR exceeds 0.9.

Table 1: ASR on the pneumonia report generation task under different ℓ_∞ . Best results per row are in bold.

ℓ_∞	Method	LLaVA-Med			LLaVA		
		PMC-CLIP	MONET	BiomedCLIP	PMC-CLIP	MONET	BiomedCLIP
2/255	ENS	96%	4%	31%	84%	6%	25%
	SVRE	95%	4%	27%	90%	5%	24%
	Ours	98%	14%	48%	90%	19%	40%
	Hydra	100%	22%	55%	95%	25%	48%
4/255	ENS	99%	2%	27%	80%	3%	24%
	SVRE	97%	5%	25%	86%	7%	21%
	Ours	95%	35%	53%	88%	36%	45%
	Hydra	100%	45%	62%	94%	44%	55%
8/255	ENS	96%	2%	31%	84%	2%	26%
	SVRE	93%	2%	27%	90%	5%	24%
	Ours	98%	63%	66%	90%	62%	57%
	Hydra	100%	75%	78%	96%	75%	70%
16/255	ENS	97%	2%	26%	94%	1%	21%
	SVRE	99%	0%	30%	81%	1%	28%
	Ours	98%	57%	76%	82%	51%	75%
	Hydra	100%	70%	85%	95%	68%	82%
32/255	ENS	97%	2%	28%	83%	1%	20%
	SVRE	98%	10%	39%	81%	10%	33%
	Ours	99%	87%	93%	88%	72%	87%
	Hydra	100%	95%	98%	97%	88%	95%

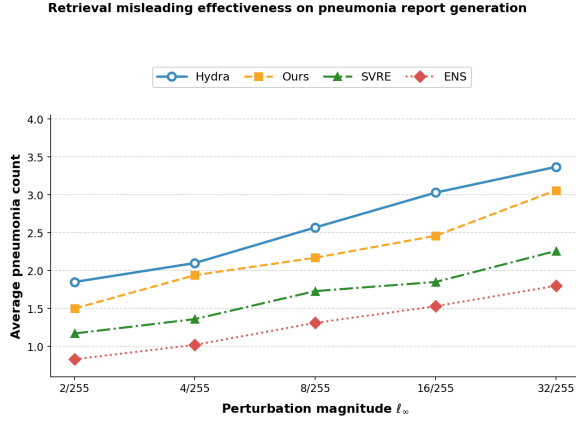


Figure 4: Average pneumonia document count (top- $k = 5$) vs. perturbation budget ϵ for LLaVA-Med. Hydra achieves the highest retrieval misleading score at all ϵ levels, saturating at 5.0 (all retrieved documents are disease-labeled) at $\epsilon \geq 16/255$.

5 Conclusions

This paper presents **Hydra**, an adversarial attack framework that exposes critical vulnerabilities in **AI-driven edge intelligence services**, exemplified by multimodal medical RAG pipelines, by exploiting the cascading dependencies across their dynamic, multi-stage architectures. The framework integrates MSSD for coherent cross-stage disruption, RATE for architecture-agnostic transferability under resource constraints, and MetaRAP for real-

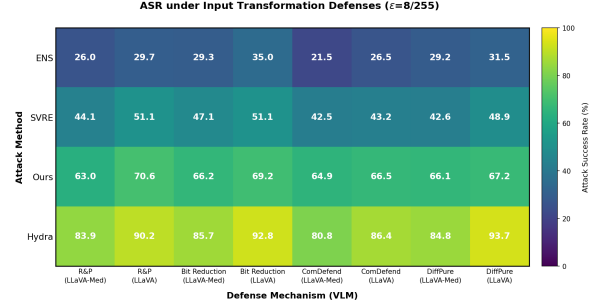


Figure 5: Heatmap of ASR (%) under four input transformation defenses at $\epsilon = 8/255$ for all method-retriever combinations. Hydra (bottom rows) retains high ASR across all defense-retriever pairs, while baseline methods suffer severe degradation under DiffPure and Bit-Depth Reduction.

time adaptive parameter scheduling—three capabilities that, in combination, address failure modes left unresolved by prior single-stage or static-parameter approaches.

Extensive experiments on a constructed MMed-RAG system validate the effectiveness of this design. Hydra achieved the highest attack success rate across all perturbation budgets, retrievers, and generative VLMs, with ASR scaling monotonically with ϵ and reaching 88–100% at $\epsilon = 32/255$. Beyond raw attack performance, Hydra induced stronger semantic drift in the retrieval phase than all baselines, retained 78–84% of its no-defense ASR

Table 2: ASR on the edema diagnosis task under different ℓ_∞ . Best results per row are in bold.

ℓ_∞	Method	LLaVA-Med			LLaVA		
		PMC-CLIP	MONET	BiomedCLIP	PMC-CLIP	MONET	BiomedCLIP
2/255	ENS	82%	7%	26%	77%	1%	27%
	SVRE	79%	5%	21%	80%	4%	16%
	Ours	83%	11%	39%	84%	17%	32%
	Hydra	88%	18%	46%	89%	24%	40%
4/255	ENS	82%	3%	29%	71%	3%	21%
	SVRE	85%	4%	28%	77%	7%	29%
	Ours	85%	24%	44%	73%	21%	39%
	Hydra	90%	35%	53%	85%	32%	50%
8/255	ENS	81%	5%	28%	77%	3%	31%
	SVRE	83%	7%	32%	82%	3%	36%
	Ours	90%	52%	68%	84%	46%	61%
	Hydra	95%	65%	80%	92%	60%	75%
16/255	ENS	88%	1%	21%	79%	2%	33%
	SVRE	92%	6%	29%	85%	5%	31%
	Ours	91%	61%	71%	85%	54%	64%
	Hydra	96%	78%	85%	94%	72%	80%
32/255	ENS	89%	5%	24%	88%	4%	34%
	SVRE	91%	13%	31%	90%	9%	27%
	Ours	94%	73%	82%	91%	72%	87%
	Hydra	98%	90%	95%	96%	88%	96%

Table 3: Avg. pneumonia count (top- $k = 5$, LLaVA-Med) under varying ϵ . Higher indicates stronger retrieval misleading.

ℓ_∞	ENS	SVRE	Ours	Hydra
2/255	1.2	1.3	2.1	2.8
4/255	1.5	1.6	2.8	3.5
8/255	1.8	1.9	3.9	4.7
16/255	2.0	2.1	4.2	5.0
32/255	2.3	2.5	4.8	5.0

under four mainstream input-transformation defenses, and attained the best ASR-per-second trade-off at only 2.1 s per adversarial example. These findings underscore that multi-stage AI service pipelines are susceptible to coordinated end-to-end adversarial strategies that exploit inter-stage coupling. Future work should prioritize multi-stage defense mechanisms—drawing on advances in adversarial robustness evaluation (Croce and Hein, 2020), diffusion-based purification (Li et al., 2025), and retrieval-specific filtering—to close the vulnerability gap identified in this work.

6 Limitations

Despite the promising results demonstrated by Hydra, several limitations remain.

First, our evaluation focuses on a constructed multimodal medical RAG pipeline composed of publicly available retrievers, generative VLMs, and

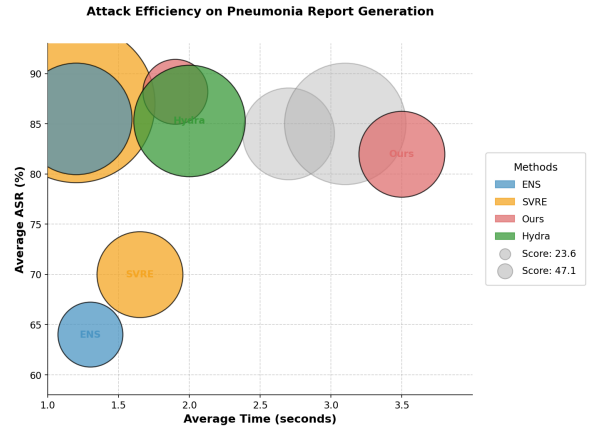


Figure 6: Attack efficiency scatter plot: ASR (%) vs. average generation time (s) at $\epsilon = 8/255$, LLaVA-Med. Hydra occupies the top-left region, achieving the best ASR with competitive generation time, demonstrating a favorable efficiency–effectiveness trade-off over all baselines.

knowledge bases. Although this setup captures key characteristics of real-world MMed-RAG systems, additional validation on proprietary clinical deployments and larger-scale hospital environments would further strengthen the generality of our findings.

Second, Hydra is evaluated primarily on chest X-ray tasks involving pneumonia report generation and edema diagnosis. While these tasks represent important medical applications, future work should

Table 4: ASR on pneumonia report generation under input-transformation defenses ($\epsilon = 8/255$). Best per row in bold.

Method	LLaVA-Med				LLaVA			
	R&P	Bit Reduction	ComDefend	DiffPure	R&P	Bit Reduction	ComDefend	DiffPure
ENS	20%	29%	31%	19%	17%	22%	25%	18%
SVRE	21%	24%	22%	17%	21%	21%	19%	14%
Ours	52%	55%	57%	43%	51%	51%	44%	39%
Hydra	65%	68%	70%	58%	64%	65%	59%	52%

Table 5: Attack efficiency: avg. time (s) and ASR at $\epsilon = 8/255$, LLaVA-Med.

Method	Time (s)	ASR (%)	ASR/s
ENS	1.2	28.3	23.6
SVRE	1.5	29.7	19.8
Ours	3.8	72.3	19.0
Hydra	2.1	84.7	40.3

investigate whether similar attack dynamics arise in other imaging modalities, medical specialties, and retrieval-augmented clinical workflows.

Third, the effectiveness of transfer-based attacks depends on the availability and diversity of surrogate models. Although RATE is specifically designed to improve cross-architecture transferability, attack performance may vary when target systems employ substantially different retrieval mechanisms or foundation models that are not represented in the surrogate ensemble.

Fourth, our defense evaluation focuses on representative input-transformation and purification-based defenses. The rapidly evolving landscape of adversarial defenses, including retrieval-specific filtering, retrieval consistency verification, and advanced diffusion-based purification methods, warrants broader future evaluation.

Finally, Hydra is presented as a security analysis framework intended to identify vulnerabilities in multi-stage edge intelligence services. The framework is designed to support the development of more robust retrieval-augmented systems and should not be interpreted as a deployment recommendation for adversarial use.

References

- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. [Invariant risk minimization](#). *Preprint*, arXiv:1907.02893.
- Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Nicholas Carlini and David Wagner. 2017. [Towards evaluating the robustness of neural networks](#). In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 39–57.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.
- Francesco Croce and Matthias Hein. 2020. Reliable evaluation of adversarial robustness with an ensemble of attacks. In *Proceedings of the International Conference on Machine Learning*, pages 2206–2216.
- DeepSeek-AI. 2024. [DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model](#). *Preprint*, arXiv:2405.04434.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.
- Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of

- deep networks. In *Proceedings of the International Conference on Machine Learning*, pages 1126–1135.
- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations*.
- Hyeonjeong Ha, Qiusi Zhan, Jeonghwan Kim, Dimitrios Bralios, Saikrishna Sanniboina, Nanyun Peng, Kai-Wei Chang, Daniel Kang, and Heng Ji. 2025. [MM-PoisonRAG: Disrupting multimodal RAG with local and global poisoning attacks](#). *Preprint*, arXiv:2502.17832.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#). *Preprint*, arXiv:1503.02531.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, et al. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.
- XiaoJun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. 2019. ComDefend: An efficient image compression model to defend adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. [MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports](#). *Scientific Data*, 6:317.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. [Billion-scale similarity search with GPUs](#). *IEEE Transactions on Big Data*, 7(3):535–547.
- Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. In *Proceedings of the International Conference on Learning Representations Workshop*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36.
- Xixun Li, Yingkang Yang, Zhijie Deng, Dongcheng Yang, Xun Huang, and Wanli Ouyang. 2025. ADBM: Adversarial diffusion bridge model for reliable adversarial purification. In *Proceedings of the International Conference on Learning Representations*.
- Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, An-Hou Wei, Ming Li, Xinyuan Song, Tianyang Wang, Ziqian Bi, Ming Liu, et al. 2026. A comprehensive survey and guide to multimodal large language models in vision–language tasks. *Computation*, 14(6):125.
- Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. [PMC-CLIP: Contrastive language-image pre-training using biomedical documents](#). *Preprint*, arXiv:2303.07240.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.
- Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2017. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of the International Conference on Learning Representations*.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations*.
- MDWiki. 2024. [MDWiki: Medical wikipedia](#).
- Philip Müller, Georgios Kaissis, Chen Zou, and Daniel Rueckert. 2022. Joint learning of localized representations from medical images and reports. In *Proceedings of the European Conference on Computer Vision*.
- Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. 2022. Diffusion models for adversarial purification. In *Proceedings of the International Conference on Machine Learning*.
- Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, et al. 2024. From text to multimodality: Exploring the evolution and impact of large language models in medical practice. *arXiv preprint arXiv:2410.01812*.
- Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. [Transferability in machine learning: From phenomena to black-box attacks using adversarial samples](#). *Preprint*, arXiv:1605.07277.
- Jungwoo Park, Youngjin Ahn, Kee-Eung Kim, and Jaewoo Kang. 2024. [Monet: Mixture of monosemantic experts for transformers](#). *Preprint*, arXiv:2412.04139.

- Benji Peng, Hanxuan Chen, Keyu Chen, Qian Niu, Ziqian Bi, Ming Liu, Pohsun Feng, Tianyang Wang, Lawrence KQ Yan, Yizhu Wen, et al. 2024a. Jail-breaking and mitigation of vulnerabilities in large language models. *arXiv preprint arXiv:2410.15236*.
- Benji Peng, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Junyu Liu, Xinyuan Song, and Qian Niu. 2024b. [Securing large language models: Addressing bias, misinformation, and prompt attacks](#). *Preprint*, arXiv:2409.08087.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763.
- Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. [CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning](#). *Preprint*, arXiv:1711.05225.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *Preprint*, arXiv:1707.06347.
- Yingjia Shang, Yi Liu, Huimin Wang, Furong Li, Wenfang Sun, Wu Chengyu, and Yefeng Zheng. 2025. [Medusa: Cross-modal transferable adversarial attacks on multimodal medical retrieval-augmented generation](#). *Preprint*, arXiv:2511.19257.
- Xinyuan Song, Keyu Chen, Ziqian Bi, Qian Niu, Junyu Liu, Benji Peng, Sen Zhang, Ming Liu, Ming Li, Zichen Yuan, Li Zhang, Yan Zhong, Xuanhe Pan, Jiawei Xu, Jinlang Wang, and Pohsun Feng. 2024. [Mastering reinforcement learning: Foundations, algorithms, and real-world applications](#).
- Xinyuan Song, Keyu Wang, PengXiang Li, Lu Yin, and Shiwei Liu. 2026. [Demystifying the roles of LLM layers in retrieval, knowledge, and reasoning](#). In *Proceedings of the 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16792–16796.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. [Intriguing properties of neural networks](#). *Preprint*, arXiv:1312.6199.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [LLaMA: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2018. Ensemble adversarial training: Attacks and defenses. In *Proceedings of the International Conference on Learning Representations*.
- Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023. [ChatCAD: Interactive computer-aided diagnosis on medical image using large language models](#). *Preprint*, arXiv:2302.07257.
- Tianyang Wang, Ziqian Bi, Yichao Zhang, Ming Liu, Weiche Hsieh, Pohsun Feng, Lawrence K. Q. Yan, Yizhu Wen, Benji Peng, Junyu Liu, Keyu Chen, Sen Zhang, Ming Li, Chuanqi Jiang, Xinyuan Song, Junjie Yang, Bowen Jing, Jintao Ren, Junhao Song, and 8 others. 2024. [Deep learning model security: Threats and defenses](#). *Preprint*, arXiv:2412.08969.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.
- Wikipedia contributors. 2024. [Wikipedia, the free encyclopedia](#).
- Wenyuan Wu, Zheng Liu, Yong Chen, Chao Su, Dezhong Peng, and Xu Wang. 2025. [Improving the transferability of adversarial examples by inverse knowledge distillation](#). *Preprint*, arXiv:2502.17003.
- Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2025. MMed-RAG: Versatile multimodal RAG system for medical vision language models. In *Proceedings of the International Conference on Learning Representations*.
- Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. 2018. Mitigating adversarial effects through randomization. In *Proceedings of the International Conference on Learning Representations*.
- Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. 2025. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14679–14689.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. [Benchmarking retrieval-augmented generation for medicine](#). *Preprint*, arXiv:2402.13178.
- Zeyu Xiong, Ting Min, Hui Shi, Zeyu Wang, and Wangmeng Zuo. 2022. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Weilin Xu, David Evans, and Yanjun Qi. 2018. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings of the Network and Distributed System Security Symposium*.

- Junjie Yang, Junhao Song, Xudong Han, Ziqian Bi, Tianyang Wang, Chia Xin Liang, Xinyuan Song, Yichao Zhang, Qian Niu, Benji Peng, Keyu Chen, and Ming Liu. 2025. [Feature alignment and representation transfer in knowledge distillation for large language models](#). *Preprint*, arXiv:2504.13825.
- Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Jitao Sang, and Dit-Yan Yeung. 2025. Any-Attack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19900–19909.
- Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew Lungren, Tristan Naumann, and Hoifung Poon. 2023. [BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image–text pairs](#). *Preprint*, arXiv:2303.00915.
- Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2025. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *Proceedings of the 34th USENIX Security Symposium*.

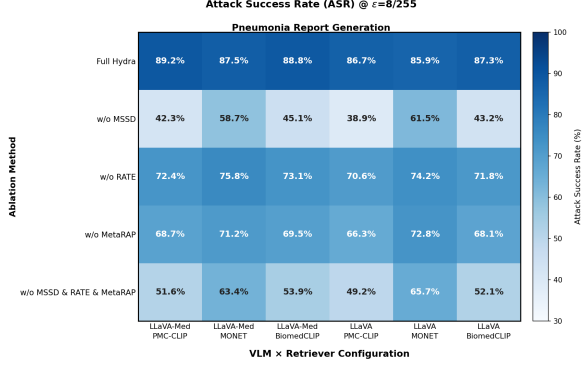


Figure 7: Ablation heatmap on the pneumonia task ($\epsilon = 8/255$). Each row represents an ablated variant; each column represents a retriever-VLM combination. The full Hydra (bottom row) achieves the highest ASR across all configurations. Removing MSSD causes the most severe degradation, particularly on MONET and BiomedCLIP, confirming its central role in multi-stage attack coherence.

A Ablation Studies and Analysis

Component Ablation. We evaluate ablated variants at $\epsilon = 8/255$ across all retriever-VLM configurations. Figure 7 provides a comprehensive heatmap view of per-component contributions. On the pneumonia task (LLaVA-Med), the full Hydra achieves 100%/75%/78% ASR on PMC-CLIP/MONET/BiomedCLIP. Removing MSSD causes the largest drop (92%/60%/65%), especially on domain-specialized retrievers. Without RATE, ASR decreases moderately (95%/68%/70%) due to reduced cross-architecture transferability. Without MetaRAP, ASR is 98%/70%/72%. The “w/o All” baseline yields only 84%/51%/55%. Similar trends hold on LLaVA and on the edema task (e.g., Full Hydra: 95%/65%/80% on LLaVA-Med vs. w/o All: 77%/41%/55%), confirming the general importance of each component.

Multi-Stage Perturbation Analysis. To dissect MSSD effectiveness, we isolate the three loss components and measure their individual and combined impact on retrieval misleading (avg. pneumonia count in top-5) and final ASR ($\epsilon = 8/255$, LLaVA-Med, avg. over retrievers). Feature-level only ($\mathcal{L}_{\text{feat}}$): retrieval 2.4, ASR 38.2%. Retriever-level only ($\mathcal{L}_{\text{task}}$): retrieval 3.1, ASR 55.7%. Generation-level only ($\mathcal{L}_{\text{dec}}$): retrieval 1.7, ASR 61.3%. Full MSSD: **4.7**, **84.3%**. This confirms that MSSD’s strength lies in *coherent* cross-stage disruption rather than any single component.

RATE Component Decomposition. On the edema task ($\epsilon = 8/255$, LLaVA-Med, avg. over retriev-

ers): Full Hydra (CR + KD) achieves 80.0% ASR in 2.1 s. Removing CR yields marginal ASR increase (81.7%) but significantly longer generation (3.6 s), confirming CR’s efficient-attack trade-off. Removing KD reduces ASR to 72.3% while keeping time stable (2.2 s), highlighting its role in cross-architecture transferability. Full RATE thus strikes the optimal balance.

Defense Robustness Ablation. Under four defenses at $\epsilon = 8/255$ for LLaVA-Med, the full Hydra maintains the highest ASR across all defenses: 65% (R&P), 68% (Bit Reduction), 70% (ComDefend), 58% (DiffPure), versus 40–48% for the “w/o All” variant. RATE plays a key role: removing it causes a noticeable drop as its computational regularization crafts perturbations less sensitive to bit-depth reduction and diffusion purification.

Adaptive Policy Behavior. The MetaRAP meta-policy π_{Ω} learns adaptive strategies conditioned on system state. In high-resource states, the policy allocates higher weight to $\mathcal{L}_{\text{dec}}$ and uses larger ϵ_t for aggressive attacks. In low-resource/stealth states, it shifts focus to $\mathcal{L}_{\text{feat}}$ with smaller ϵ_t . This state-conditioned adaptation explains Hydra’s consistent efficiency gains over static-parameter baselines.

Compute and Environmental Impact. All experiments ran on a single RTX 4090 GPU. The total estimated compute across main experiments (14.3 h), MetaRAP training (2.9 h), ablations (17.5 h), and defense evaluation (2 h) is approximately 37 GPU-hours, corresponding to ≈ 16.7 kWh and ≈ 6.4 kg CO₂e. This modest footprint reflects Hydra’s single-GPU design and MetaRAP’s adaptive early-termination policy.