ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

# Evaluating and Defending against Adversarial Threats in Multimodal AI

**Mateusz Kowalczyk** | Department of Security and Transparency of Artificial Intelligence, NASK – National Research Institute, Poland | ORCID: 0009-0004-0267-310X

**Joanna Kołodziej** | Department of Distributed Systems, NASK – National Research Institute, Poland, Department of Computer Sciences, Cracow University of Technology, Poland | ORCID: 0000-0002-5181-8713

**Mateusz Krzysztoń** | Department of Distributed Systems, NASK – National Research Institute, Poland | ORCID: 0000-0002-6020-406X

## Abstract

Multimodal artificial intelligence (AI) systems combine text, images, audio, video, and sensor data in a single pipeline. This design improves capability, but it also creates new attack surfaces and cross-modal failure modes. The existing works often study attacks, defences, and benchmarks in isolation, making the field hard to compare systematically and leaving practical security choices unclear. This survey reviews adversarial threats, defence methods, and robustness evaluation in multimodal AI. The scope extends beyond standard vision–language models to multimodal agents and systems that use audio, depth, and thermal data. The survey introduces a defence taxonomy with two groups: proactive methods that improve robustness before deployment, and reactive methods that detect or limit harmful behaviour at inference time. The survey also presents a practical view of robustness evaluation, covering task-aligned protocols, attack success metrics, repeated sampling, open-ended judging, benchmark roles, and a minimum reporting checklist. The main conclusion is simple: no

**Received:** 11.12.2025

**Accepted:** 01.04.2026

**Published:** 12.05.2026

**Cite this article as:**
M. Kowalczyk, J. Kołodziej, M. Krzysztoń, "Evaluating and defending against adversarial threats in multimodal AI," ACIG, vol. 5, no. 2, 2026, doi: 10.60097/ACIG/220237

**Corresponding author:**
Mateusz Kowalczyk, Department of Security and Transparency of Artificial Intelligence, NASK – National Research Institute, Poland; E-mail: mateusz.kowalczyk@nask.pl
0009-0004-0267-310X

single defence is sufficient. Secure multimodal AI requires layered defences and realistic evaluation.

**Keywords**

*adversarial machine learning, adversarial attacks, multimodal robustness evaluation, multimodal guardrails, multimodal defences*

## 1. Introduction

### 1.1. Motivation

Multimodal models combine multiple data types. Common inputs include text, images, audio, video, and sensor data. Many deployed systems now use several inputs within a single pipeline. Examples include autonomous driving systems using cameras, LiDAR, and radar, and medical systems that combine images with clinical records. Multimodal large language models (MLLMs), vision–language models (VLMs), and artificial intelligence (AI) agents extend the same trend. These systems use richer input contexts than unimodal models, but they also create more attack surfaces.

Each added modality creates a new path for attack. An attacker may alter one input stream, exploit a mismatch between two input streams, or interfere with the way a model combines them. Attacks that exploit interactions between modalities are cross-modal attacks. Those attacks do not appear in the same form in a single-modality setting. As a result, adversarial robustness has become a central problem for multimodal AI.

### 1.2. Problem Statement and Scope

Adversarial attacks on multimodal systems exploit individual modalities or the links between modalities. A malicious image may change the answer to a text question. A harmful prompt may steer the interpretation of an image. A corrupted sensor signal may alter both perception and later decisions. Recent works have described several of these risks in the vision–language domain [1–4]. The field still lacks a clear view of how to defend multimodal systems and test those defences consistently.

The paper focuses on adversarial robustness in multimodal AI. The main focus is inference-time manipulation of inputs, prompts, context, and agent interactions. The review covers evasion attacks,

cross-modal attacks, prompt injection, and related failures in multimodal reasoning. The review then studies defence methods and robustness evaluation for those settings. The paper does not aim to cover all AI security topics. The paper does not review data poisoning, membership inference, model extraction, watermarking, or deployment controls, such as sandboxing, network isolation, or zero-trust design. Those topics remain important to AI security, but they fall outside the scope of this work.

The paper provides a structured review of multimodal adversarial robustness, with an emphasis on defence methods and evaluation. The threat landscape appears first because the later defence and evaluation sections depend on a clear view of what attackers are capable of, how the threat landscape changes, and which system components they target.

### 1.3. Review Scope and Literature Selection

The review follows a structured narrative approach. We collected papers from arXiv (Cornell University), Institute Of Electrical And Electronics Engineers (IEEE) Xplore, and the Association for Computational Linguistics (ACL) anthology. We then followed references in key papers to identify earlier works and subsequent follow-up papers. The review places greater weight on recent papers because multimodal AI evolves rapidly. The review also includes a small set of foundational papers that define models, attacks, or defence ideas that later work builds on.

We included papers that study adversarial threats, defence methods, or robustness evaluation of systems that process or align multiple modalities. The scope covers foundational multimodal models (e.g. CLIP), VLMs, MLLMs, and selected multi-sensor systems. We usually excluded papers that study only single modality. We kept such papers when a method or idea helps to explain a multimodal defence category or a design choice in the taxonomy.

Our primary contributions are as follows:

- We review the current threat landscape for multimodal AI, including attacks on AI agents and systems that use less-studied sensory inputs.
- We organise multimodal defence methods into proactive and reactive categories and explain the main strengths and limits of each group.

- We present a practical view of robustness evaluation for multimodal systems, including metrics, benchmark roles, evaluation gaps, and a minimum reporting checklist.
- We identify research gaps and outline future directions for secure and reliable multimodal AI systems.

Work is structured to logically guide the reader from the problem to its solutions and their measurement. Section 2 defines the background and the threat model used in the review. Section 3 covers related work and the threat landscape. Section 4 presents the taxonomy of multimodal defence methods. Section 5 reviews robustness evaluation. Section 6 discusses key findings and future work. Section 7 concludes the paper.

## 2. Background

Multimodal machine learning systems process and combine multiple data types, such as text, images, audio, video, or sensor signals. This design supports tasks that depend on multiple sources of evidence, including visual question answering, medical decision support, autonomous driving, and tool-using AI agents. The same design also creates new security risks. An attacker may target a single modality, exploit a mismatch between modalities, or interfere with how a model aligns and combines them. In multimodal systems, a small change in one input stream can, therefore, affect not only perception in that stream but also later reasoning, generation, or action.

Adversarial attacks study how an attacker modifies an input to induce model failure. In the standard formulation, the attacker crafts a perturbed input $x'$ from a benign input $x$ such that the perturbation remains bounded by a budget $\|x' - x\| \leq \epsilon$, while the model output changes. Equation (1) captures the two common cases: untargeted attacks, which induce any wrong output, and targeted attacks, which force a specific incorrect target $t$.

$$\|x' - x\| \leq \varepsilon \qquad (1)$$
$$f(x') \neq y \text{ (untargeted) or } f(x') = t,\ t \neq y \text{ (targeted)}$$

This formulation remains useful in the multimodal setting, but it does not cover the full attack space. Many multimodal attacks do not rely only on small norm-bounded perturbations. Some attacks inject malicious text into retrieved context, place harmful instructions inside images or audio, exploit contradictions between

modalities, or shape later actions in an agentic system over several steps. For that reason, the remainder of the paper uses a broader threat-model view that includes both low-level perturbation attacks and higher-level attacks on multimodal reasoning and interaction.

### 2.1. Threat Model

The paper studies adversaries that aim to cause an unsafe or incorrect system outcome in a multimodal setting. Depending on the task, the attacker may seek a wrong class label, failed retrieval, harmful generated content, a successful jailbreak, a prompt-injection failure, or an incorrect action chosen by an agent. Some attacks target a single output. Other attacks target a sequence of outputs or actions.

The literature usually describes attacker knowledge with three standard settings. In a white-box setting, the attacker knows the model architecture and has access to parameters or gradients. In a black-box setting, the attacker interacts with the system only through queries and observed outputs. In a grey-box setting, the attacker knows part of the system, such as the model family, training setup, or a surrogate model, but not the full deployed model.

The main focus of the paper is inference-time access. The attacker changes one or more inputs at test time, changes the interaction context, or injects malicious content through external tools or retrieved data. Common entry points include text prompts, images, audio, video, web content, retrieved documents, and sensor signals. In agentic systems, the attacker may also manipulate the environment that provides observations or tool outputs to the model.

A multimodal attack may target several system components. The first target is the raw input stream of a modality. The second target is the modality encoder that converts raw input into features. The third target is the alignment or fusion mechanism that links representations across modalities. The fourth target is the downstream module that produces labels, generated responses, or actions. This distinction matters because the same perturbation can produce different failures depending on where the system integrates information.

Attack form also varies across multimodal settings. A perturbation may affect single modality or several modalities at once. A

perturbation may be digital or physical. A perturbation may be low-level, such as changes in pixels, waveform samples, or point clouds, or semantic, such as a hidden instruction, a misleading caption, or a cross-modal contradiction. Agentic systems add a further case in which the attacker shapes the context in several interaction steps to influence reasoning or behaviour.

The threat model in this paper treats those attacks as practical security risks rather than only as benchmark artifacts. Multimodal systems already operate in settings where attackers can control prompts, web content, uploaded media, or physical sensor inputs. The combination of multiple input channels increases the number of attack paths and makes failure analysis harder, because a system may appear robust when each modality is tested in isolation but fail when modalities interact. The threat landscape in later sections builds on this view.

## 3. Related Work

Several recent surveys examine adversarial robustness in multimodal AI, but most surveys focus on attacks or on a narrow model family. This section situates the present paper within that literature.

Jiang et al. [5] review adversarial robustness in MLLMs. The survey covers attacks and defences across text, image, video, and audio inputs. The focus remains on multimodal systems built around large language model backbones. The present paper considers a broader set of multimodal systems, including AI agents and systems that use emerging sensory inputs beyond standard vision–language settings.

Kapoor et al. [6] survey adversarial attacks on open-source multimodal systems. The survey covers several attack classes, including optimisation-based, backdoor, and membership inference attacks. The main contribution is a structured view of attack execution. The paper gives less attention to defence design and robustness evaluation. The present paper complements that line of work by focusing on defence mechanisms and on how to evaluate them.

Our previous work [7] surveyed adversarial attacks on vision–language systems and organised them into alignment- and fusion-based categories. That survey focused on how attacks exploit shared embedding spaces and cross-modal interactions. Our paper shifts the focus from attack construction to defence methods and robustness evaluation.

Taken together, the existing surveys give strong coverage of multimodal attack mechanisms and threat categories. The existing surveys give less attention to three questions that matter for deployment: how to organise defence methods across training- and inference-time settings; how to evaluate robustness under adaptive attack and realistic multimodal use; and how to extend the analysis beyond standard VLMs to agents and multi-sensor systems. The present paper addresses these questions by focusing on defence and evaluation, rather than attack generation.

### 3.1. Threat Landscape

The attack landscape for multimodal systems has expanded from single-input exploits to system-level attacks. The three phases mentioned below summarise broad trends in the literature. The phases are not strict historical boundaries. Several attack types coexist, but the sequence helps to explain how the field moved from modality-local failures to attacks on multimodal interaction, reasoning, and action.

*Phase 1: Single-modality exploits:* Early work targeted one input channel within a multimodal system. Attackers perturbed a text prompt, an image, or another single modality to change the final output or bypass a safety mechanism. These attacks showed that multimodal systems can inherit vulnerabilities from their unimodal components [8, 9].

*Phase 2: Cross-modal attacks:* Later work moved from single-channel perturbations to attacks that exploit the coupling between modalities. In this setting, an attacker manipulates one modality to change the model response in another modality or in the fused representation. VLMs provide many examples of this pattern, including attacks in which an adversarial image changes text-based reasoning or a textual input changes visual interpretation [6, 7]. This phase marked a shift from modality-local failures to failures that arise from cross-modal alignment and fusion.

*Phase 3: System-level and emerging-modality threats:* A recent work studies attacks on more complex multimodal systems [10, 11]. This line of research includes systems that process audio [12], depth [13], thermal data [14], and other sensory inputs beyond standard vision–language benchmarks. The same period also saw the emergence of multimodal agents that interact with tools and external environments. In those settings, an attacker can manipulate perception, retrieved context, or the surrounding environment to

alter later reasoning and action. These developments increase the number of attack paths and make robustness evaluation harder, because static input perturbation tests no longer capture the full failure surface [10, 11, 15].

### 3.2. The Agentic Frontier

Autonomous multimodal agents introduce risks that go beyond static adversarial examples. An agent not only predicts an output from a fixed input but perceives an environment, reasons over intermediate states, stores context, and takes actions through tools or application programming interfaces (APIs). Each step creates a new attack surface. Recent works show that multimodal attacks on agents can target visual and textual inputs, environmental signals, internal reasoning components, and persistent memory [10, 15].

*Cross-modal prompt injection:* Cross-modal prompt injection coordinates attacks across multiple channels. Wang et al. [16] propose CrossInject, a black-box attack framework for multimodal agents. CrossInject combines two components: visual latent alignment, which injects malicious semantics into the visual input via generated or perturbed images, and textual guidance enhancement, which optimises the textual instruction to better align with the attacker's goal. The reported results show that CrossInject outperforms prior attacks by at least 30.1% in attack success rate across diverse tasks. The main lesson is that multimodal prompt injection can be more effective when the attacker coordinates both modalities, rather than attacking only one input channel.

*Environmental and sensory manipulation:* Other attacks manipulate the environment that the agent observes. Aichberger et al. [11] introduce Malicious Image Patches (MIPs) for multimodal operating system (OS) agents. An MIP is a perturbed screen region that appears inside a screenshot and induces harmful API-level behaviour. The authors show that an MIP can be embedded in a desktop wallpaper or a social media post and hijack multiple OS agents during benign tasks. Chen et al. [15] study active environmental injection attacks against Android agents. Their attack uses environmental elements such as notifications to interfere with agent execution. The paper identifies two main vulnerabilities: adversarial content injection via multimodal interaction interfaces and reasoning gaps during the delay between perception and action. The reported attack success rate reaches 93% on AndroidWorld when both vulnerabilities are combined. Taken together, these papers show that the agent environment itself becomes an attack vector

once the model acts on screenshots, notifications, or other external signals.

*Reasoning and goal manipulation:* Agentic attacks can also target intermediate reasoning components. Wu et al. [10] study multimodal web agents in a realistic adversarial setting on top of VisualWebArena. The authors propose the agent robustness evaluation framework, which models the agent as a graph of components and tracks how adversarial influence propagates through that graph. Their experiments show that an attacker can hijack agents that use reflection and tree search with small perturbations to a single image. The same study also shows a more subtle result about inference-time compute. Components such as evaluators and value functions can reduce attack success when they remain uncompromised, but they also create new attack paths when an attacker can manipulate them. This result matters because it shows that added reasoning does not automatically improve security.

*Memory and context manipulation:* Persistent context creates another attack surface. Patlan et al. [17] study context manipulation in Web3 agents and identify memory injection as a distinct threat. In their setting, malicious content is stored in history rather than in the current user prompt. The poisoned memory then influences later actions, including actions taken for other users or on other platforms when memory is shared. The paper shows that memory injection is more persistent and more difficult to block than ordinary prompt injection. It also shows that prompt-level defences and prompt-injection detectors provide limited protection once stored context is corrupted, whereas fine-tuning-based defences substantially improve robustness. This result shows a shift from single-turn prompt safety to long-term context integrity.

Papers in this section show a common pattern. Adversarial threats against multimodal agents no longer target only one prediction step. They target the full perception–reasoning–action loop, including intermediate components and stored context. The attack assumptions differ across papers. Environmental attacks often depend on access to the observed interface or timing of execution. Memory attacks depend on the way the agent stores and reuses history. Reasoning attacks depend on access to components that shape action selection. Despite those differences, the broader lesson remains consistent: once a multimodal model becomes an agent, security analysis must move beyond input perturbation to system-level robustness.

### 3.3. Audio Frontier

Audio adds a distinct attack surface to multimodal systems. Unlike text, speech is continuous, time-dependent, and shaped by acoustic and conversational context. These properties create attack paths that do not appear in text-only interfaces in the same form. Recent work shows that adversaries can exploit audio at several levels: waveform perturbations, speech-token manipulations, conversational framing, and cross-modal disruption in audio–visual models [5, 12, 18–20].

*Audio jailbreaks:* Several papers show that the audio channel alone can bypass alignment safeguards. Gupta et al. [12] study universal audio jailbreaks in audio–language models. Their attack constructs perturbations that generalise across prompts, tasks, and base audio samples, and remain effective under simulated real-world playback conditions. The same paper also shows that successful perturbations encode imperceptible toxic speech patterns into a signal, suggesting that the model responds to linguistic cues embedded in the waveform rather than to noise alone. Ma et al. [18] study a different attack surface in SpeechGPT. Their method targets the model's discrete speech–token interface. The attack uses greedy search to modify adversarial token sequences and then synthesises them into audio using a vocoder (voice encoder). The reported results show attack success rate of up to 89% across restricted task categories. Taken together, these papers show that audio jailbreaks can bypass safety controls at both continuous waveform and discrete-token levels.

*Conversational framing attacks:* Not all audio attacks rely on signal perturbation. Chiu et al. [19] introduce flanking attack, a voice-based jailbreak method for multimodal LLMs. The attack inserts a harmful request between benign, narrative-driven prompts, making the full interaction appearing less suspicious. The reported results show average attack success rates from 0.67 to 0.93 across seven forbidden scenarios, with an overall average of 0.81. This result matters because it shows that audio threats also arise from the design of context. A model can fail even when the waveform itself remains natural, and the attack acts only through conversational framing.

*Cross-modal attacks in audio–visual models:* Audio also serves as one part of a larger cross-modal attack surface. Zhang et al. [20] study adversarial robustness in audio–visual models from two directions. The first attack, temporal invariance attack, exploits temporal redundancy across consecutive segments. The second attack,

modality misalignment attack, weakens the semantic correspondence between audio and video. The paper argues that temporal consistency and inter-modal correlation are central properties of audio–visual learning, and that both can be leveraged as attack vectors. This finding is important for multimodal security because it shifts attention from attacks on one modality to attacks on the fusion process itself.

*Authenticity and detection challenges:* The audio frontier also includes direct manipulation of source media, such as voice cloning, partial splicing, and coordinated audio–visual deepfakes. Luong et al. [21] introduce Llama Partial Spoof, a 130-hour dataset of fully and partially fake speech built with large language models and voice cloning systems. Their results show that current fake-speech detectors struggle to generalise to unseen scenarios, with the best-reported equal error rate of 24.49%. Cai et al. [22] introduce AV-Deepfake1M, a large-scale dataset with more than 1 million manipulated audio–visual videos for temporal localisation and detection. The benchmark shows a marked drop in performance for current detection and localisation methods compared with earlier datasets. These papers do not study jailbreaks in the narrow sense, but they are directly relevant to multimodal robustness because they show how hard it is to verify the authenticity and temporal consistency of audio and audio–visual inputs.

Papers in this section show a common pattern. Audio threats in multimodal systems appear at several layers. Gupta et al. [12] target the waveform. Ma et al. [18] target the speech–token interface. Chiu et al. [19] target conversational context. Zhang et al. [20] target temporal consistency and cross-modal alignment. Luong et al. [21] and Cai et al. [22] show that authenticity checks remain weak for partial spoofing and coordinated audio–visual manipulations. The broader lesson is that transcript filtering alone is not enough. Robust defence for audio-enabled multimodal systems must address acoustic perturbations, speech–token processing, conversational context, cross-modal fusion, and media authenticity together.

### 3.4. Sensory Frontier

The threat landscape extends beyond text, images, and audio. Many deployed systems rely on LiDAR, thermal cameras, infrared sensors, ultrasound scanners, and other domain-specific inputs. These sensors support safety-critical tasks in driving, surveillance, and medicine. They also create new attack paths because

the model must process signals with different physical properties and then combine them in a shared decision pipeline.

The first line of work studies transfer across heterogeneous sensing modalities. Gong et al. [14] propose multiform attack, a dual-layer optimisation framework for cross-modality adversarial transfer. The method first learns universal perturbations within a single modality via gradient-based optimisation, then uses an evolutionary search to identify shared perturbations that transfer across modalities. The paper focuses on the transfer between RGB, infrared, and thermal images. The main lesson is that an attacker does not need identical sensor outputs to reuse an attack. If the optimisation can capture structure shared across heterogeneous sensors, perturbations can transfer across modalities that are produced by different hardware.

Physical attacks on depth estimation exhibit a similar pattern in autonomous driving. Zheng et al. [13] propose 3D Depth Fool (*3D*$^2$ *Fool*), a 3D texture-based attack against monocular depth estimation. The attack paints adversarial camouflage across the full surface of a vehicle, rather than on a small 2D patch. The optimisation also accounts for viewpoint changes and adverse weather, such as rain and fog. Real-world tests with printed 3D textures on physical vehicle models show depth estimation errors exceeding 10 meters. This result matters because it shows that depth models can be fooled by physical textures that remain effective under realistic viewing conditions.

Another work attacks fusion directly. Tu et al. [23] study multi-sensor detectors that combine LiDAR and camera inputs. Their attack inserts a physically realisable adversarial object on top of a host vehicle and renders it consistently into both the LiDAR sweep and the camera image. The results show that a single universal adversary can hide host vehicles from state-of-the-art multimodal detectors. The same study also shows that attacking the image branch alone is almost as strong as attacking both modalities together. In addition, projection of corrupted image features into 3D can create distant false positives. The key lesson is simple: fusion does not remove the weakness of image features, and in some architectures, the fusion process can spread that weakness further.

Medical imaging pipelines face a different but equally important class of threats. Byra et al. [24] attack a deep model for fatty liver disease classification by changing the ultrasound reconstruction process rather than the final image pixels. The attack perturbs reconstruction parameters, such as attenuation compensation

and compression thresholds, via zeroth-order optimisation. The reported success rate reaches 48%. This paper matters because it shifts the attack surface from image post-processing to signal reconstruction. A medical model can fail before a standard digital image is even formed.

Medical MLLMs introduce another failure mode. Huang et al. [25] study mismatched malicious attacks and optimised mismatched attacks on Med-MLLMs. Their work builds the 3MAD dataset and proposes the multimodal cross-optimisation method (MCM), which jointly optimises image and text inputs during jailbreak generation. The attacks target cases where the medical image, anatomical context, and query do not match cleanly, or where a malicious clinical request is combined with mismatched multimodal input. The paper shows that Med-MLLMs remain vulnerable under both white-box and transfer settings. The broader lesson is that medical multimodal models do not reliably verify semantic consistency between image content and medical queries before producing an answer.

The papers in this section show three recurring weaknesses. First, attackers can exploit the physical acquisition process, as in ultrasound reconstruction attacks [24]. Second, attackers can place physically realisable perturbations into the environment, as in 3D texture attacks on depth estimation and adversarial objects against LiDAR-camera fusion [13, 23]. Third, attackers can exploit failures in cross-modal consistency, as in Med-MLLM mismatched attacks and heterogeneous cross-sensor transfer [14, 25]. These results show that unimodal safeguards are not enough in sensor-rich systems. Robust evaluation in the sensory frontier must account for physical deployment constraints, reconstruction pipelines, fusion failures, and semantic consistency across modalities.

## 4. Taxonomy of Multimodal Defence Mechanisms

We organise multimodal defence methods according to when they act. The taxonomy separates *proactive defences*, which modify the model or training pipeline before deployment, from *reactive defences*, which monitor, filter, or constrain the system during inference. This split is simple but remains useful across diverse multimodal settings, including VLMs, multimodal agents, and multi-sensor systems.

The distinction also reflects a practical difference. Proactive methods improve robustness or safety before release. Reactive methods

reduce harm after deployment by detecting unsafe inputs, blocking unsafe behaviour, or limiting the effect of successful attacks. In practice, the two groups are complementary. Proactive methods improve the model's baseline robustness, but reactive methods remain necessary because attackers adapt, threat models evolve, and real-world inputs often differ from those in training.

Figure 1 summarises the overall taxonomy. Next sections examine proactive and reactive defences in detail and then compare their strengths and limitations in practice.

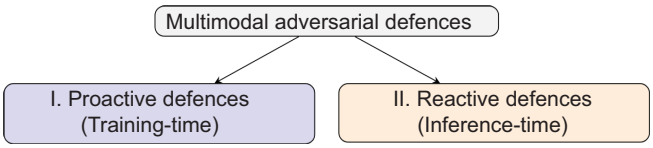


**Figure 1.** The high-level classification of multimodal adversarial defences.

### 4.1. Proactive Defences

Proactive defences improve safety and robustness before deployment. Instead of reacting to harmful inputs at inference time, these methods modify the model, the training data, or the training objective to make unsafe or unstable behaviour less likely. We divide them into two groups: Robustness-focused training, which aims to preserve correct behaviour under perturbation, and safety alignment training, which aims to reduce harmful or policy-violating outputs through alignment-oriented supervision. Figure 2 summarises this branch of taxonomy.

*Robustness-focused training:* Robustness-focused training aims to keep the model stable when one or more modalities are perturbed, corrupted, or missing. This is harder in multimodal systems than in unimodal ones because failures often arise at the interaction point between modalities rather than inside a single encoder. As a result,

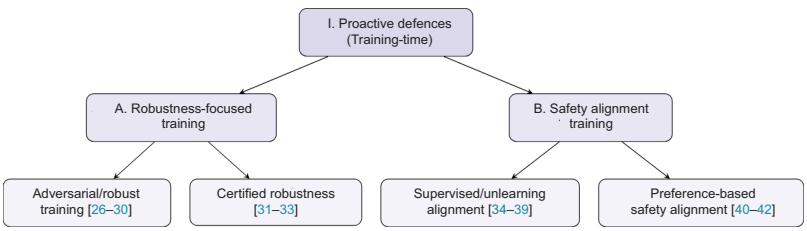


**Figure 2.** Taxonomy of proactive defences, dividing into robustness-focused training and safety alignment training categories.

methods in this group intervene at different parts of the pipeline, including the image and text encoders, the projector or alignment layer, the fusion stage, and the final prediction.

*Adversarial/robust training*. The first line of work improves robustness empirically by exposing the model to adversarial or corrupted inputs during training. Zhou et al. [26] propose multimodal contrastive adversarial training (MMCoA) for CLIP-style VLMs. MMCoA aligns clean text embeddings with adversarial image embeddings and adversarial text embeddings with clean image embeddings. This strengthens both encoders under multimodal attack rather than only the image branch. Their results show that multimodal adversarial training (MAT) improves robustness more than unimodal adversarial training across image, text, and multimodal attacks.

Waseda et al. [27] also defend against multimodal attacks by perturbing both image and text during training. Their MAT framework studies how one-to-many image–text augmentation can preserve semantic diversity while reducing harmful distribution shift. The results show that MAT outperforms unimodal defences in image–text retrieval. This suggests that multimodal robustness depends on preserving cross-modal alignment during training, rather than hardening a single modality in isolation.

Recent work extends the same idea to multimodal jailbreaks in MLLMs. Lu et al. [28] propose E2AT, an efficient end-to-end adversarial training framework for MLLMs. E2AT does not retrain the full model from scratch. Instead, it focuses on projector-level adversarial training and combines this with dynamic joint multi-modal optimisation, which balances standard and adversarial objectives across text and image channels. The reported results show an average improvement of 34% over the existing baselines for both visual and textual jailbreak attacks while preserving clean task performance. This shows that robustness can be improved at the alignment bottleneck without the full cost of end-to-end adversarial training over all parameters.

A second line of work defends the interaction point between modalities more directly. Yang et al. [29] show that standard fusion models remain vulnerable even when only one modality is attacked. To address this, they propose an adversarially robust fusion strategy that compares information across modalities, detects inconsistencies in the perturbed source, and gates that source, ensuring that only unperturbed modalities contribute to the final decision. Their

method improves robustness across action recognition, object detection, and sentiment analysis without degrading performance on clean data. Tsai et al. [30] study a different setting in which multiple modalities are converted into generalised text prompts. Their text-centric adversarial prompting method improves robustness to noisy inputs, missing modalities, and changing input order. Together, these papers show that proactive robustness does not always require full end-to-end retraining. In some architectures, the main weakness lies in the fusion or alignment stage, and defending that stage can give broad gains.

*Certified robustness*. The second family of robustness-focused methods aims to provide formal guarantees under a bounded threat model. This is harder in multimodal systems than in unimodal classifiers because the method must account for heterogeneous inputs and, in some settings, semantic transformations in the physical world. Wang et al. [31] propose MMCert, the first certified defence designed specifically for multimodal models. MMCert uses independent sub-sampling and ablation across modalities to derive a lower bound on model performance under bounded perturbations to both modalities. Their experiments on multimodal road segmentation and multimodal emotion recognition show that MMCert outperforms certified defences from unimodal settings.

Huang et al. [32] extend certification to multi-sensor fusion under semantic transformations such as rotation and shifting in the physical world. COMMIT combines anisotropic randomised smoothing with a grid-based splitting method to characterise semantic attacks in multi-sensor fusion systems. The paper also introduces efficient certification procedures for detection accuracy and Intersection over Union (IoU) in large-scale fusion models. Their benchmark shows that certification for multi-sensor fusion can be substantially higher than for single-modal models under the same semantic threat class. This shows that multimodal certification must reflect geometry of the attack, not only an abstract norm-bound.

Another work studies the training conditions that make multimodal robustness certifiable. Yang et al. [33] analyse robustness through the lens of modality preference. Their theory shows that robustness depends not only on unimodal representation margins but also on how reliably the model integrates modalities. It also shows that a preference for one modality can make attacks on that modality more effective. To address this, they propose certifiable robust multimodal training (CRMT), which regularises unimodal classifiers and integration weights to improve robustness in a certifiable way.

CRMT is useful here because it links a common multimodal failure mode, namely modality imbalance, to a concrete training–time intervention.

The papers categorised under robustness-focused training show a clear shift in emphasis within proactive multimodal defence research. Earlier work often focused on robustness within a single modality, especially the image branch, or assumed that stronger unimodal robustness would transfer naturally to the multimodal setting [26, 27].

More recent methods instead defend the interaction point between modalities, either through joint adversarial training over image and text inputs [26–28], through fusion-aware inconsistency detection and gating [29], or through explicit certification of multimodal predictions under bounded perturbations or semantic transformations [31–33].

The assumptions still differ across methods. MAT requires access to attack generation during fine-tuning and can become costly as model size grows [28]. Fusion-aware methods assume that disagreement between modalities can be detected reliably and used to isolate the perturbed source [29]. Certified methods provide stronger guarantees, but only under restricted perturbation or transformation classes [31, 32]. In addition, CRMT shows that modality preference can render a channel disproportionately fragile, even within a multimodal system [33]. The main conclusion is not that one robustness-focused defence solves the multimodal problem, but that proactive robustness must be designed around cross-modal interaction rather than added only at the level of isolated inputs.

*Safety alignment training:* Safety alignment training aims to reduce harmful, ungrounded, or policy-violating outputs by shaping model behaviour during training. Unlike robustness-focused training, which aims to preserve correct predictions under perturbation, safety alignment methods focus on the final response after multimodal information has been processed. We divide them into two groups: supervised/unlearning alignment, which relies on curated supervision or targeted knowledge removal, and preference-based safety alignment, which uses ranked feedback or corrective supervision to steer behaviour.

*Supervised/unlearning alignment*. The first line of work aligns multimodal models through curated safety supervision. Zong et al. [34]

propose VLGuard, a dedicated vision–language safety dataset for *post hoc* fine-tuning. Their results show that integrating VLGuard into standard vision–language fine-tuning, or applying it after instruction tuning, improves safety while preserving helpfulness, with black-box attack success rates approaching zero in many settings. This suggests that curated multimodal safety data can provide a strong baseline guardrail at relatively low cost.

Later work shows that standard safety fine-tuning can still be brittle. Chen et al. [35] identify a failure mode they call the *safety mirage*, in which fine-tuning reinforces spurious correlations between superficial textual patterns and refusal behaviour, rather than learning deeper harm-sensitive representations. They show that this leaves fine-tuned VLMs vulnerable to simple one-word attacks and also increases over prudence on benign inputs. To address this, they propose machine unlearning as an alternative alignment mechanism. Their results show that unlearning can reduce attack success rates by up to 60.2% while decreasing unnecessary benign refusals by more than 84.2%. The main lesson is that naive supervised labels can produce shortcut-based safety behaviour rather than genuine multimodal safety.

Chakraborty et al. [36] extend the unlearning idea to cross-modal safety alignment. Their main claim is that, because many multimodal systems fuse inputs into the language space, textual unlearning alone can transfer well to multimodal safety. Across several datasets, they show that textual unlearning can reduce attack success rates to below 8%, and in some settings close to 2% for both text-only and vision–text attacks while preserving utility. They also report that multimodal unlearning data adds little benefit while substantially increasing computational cost. This strengthens the argument that the language component remains a key alignment bottleneck in many multimodal systems.

Other work improves safety alignment by adding richer reasoning structure or lower-cost multimodal supervision. Ding et al. [37] argue that the existing safety fine-tuning methods suffer from a *safety reasoning gap*: they can reject clearly unsafe prompts, but they struggle with harder cases that require visual reasoning across multiple images. To address this, they introduce the multi-image safety (MIS) dataset, which pairs multi-image inputs with safety chain-of-thought labels that explicitly encode the required reasoning steps. Their experiments show that MIS fine-tuning improves safety on challenging multi-image tasks while preserving general capabilities. In a different direction, Lu et al. [38] propose SEA, a

low-resource alignment method that synthesises modality embeddings directly in the encoder output space and combines them with text-only safety data. SEA is designed for settings where multimodal safety datasets are too costly to collect. The results show that these synthetic embeddings can be generated quickly and can improve the security of image-, video-, and audio-based MLLMs. Together, these papers show that supervised safety alignment becomes stronger when it moves beyond static input–output labels and targets the language–space bottleneck directly or adds explicit multimodal reasoning structure.

An emerging direction in this branch is *representation-centric unlearning*. Chen et al. [39] propose AUVIC, a training-time framework for adversarial unlearning of target visual concepts in MLLMs. Although AUVIC is motivated mainly by privacy and right-to-be-forgotten settings rather than jailbreak defence, it is relevant here because it shows that multimodal safety-relevant behaviour can also be changed by editing internal concept representations. AUVIC uses adversarial perturbations to isolate the target visual concept during unlearning and introduces mechanisms to preserve related non-target concepts, achieving strong forgetting with limited collateral degradation. In this taxonomy, AUVIC broadens safety alignment beyond refusal-style supervision and shows that harmful or sensitive multimodal concepts can be removed at the representation level during training.

*Preference-based safety alignment*. The second line of work aligns multimodal behaviour through ranking, correction, or preference optimisation rather than direct supervised labels. Bai et al. [40] introduce Constitutional AI, a foundational preference-based alignment framework in which a model critiques and revises its outputs according to explicit principles, and reinforcement learning is then performed using AI feedback rather than RLHF. Although the method was introduced for text-only assistants, it remains relevant here as a conceptual precursor for scalable safety alignment with fewer human labels.

RLHF-V brings that idea into the multimodal setting. Yu et al. [41] collect fine-grained human corrections on hallucinated segments of model outputs and train the model with dense direct preference optimisation. Their results show that, with only 1.4K annotated samples, RLHF-V reduces hallucination rates by 34.8% and outperforms a concurrent multimodal RLHF baseline trained on much more data. This suggests that multimodal preference alignment

benefits from feedback that is localised to the erroneous segments rather than assigned only at the full-response level.

Liu et al. [42] focus on a different problem: modality imbalance during reasoning. They propose modality-balancing preference optimisation (MBPO), which constructs hard-to-reject responses by adversarially perturbing input images, so that the model relies too heavily on language priors and produces visually ungrounded outputs. These hard negatives are then combined with online preference optimisation using verifiable rewards. The reported results show improvements on challenging vision–language tasks and reduced hallucination. Together, RLHF-V and MBPO suggest that preference-based alignment is especially effective when negative examples are informative, localised, and directly tied to failures of multimodal grounding.

The papers categorised under safety alignment training show a clear shift from simple supervised rejection behaviour towards deeper behaviour shaping through unlearning, structured reasoning, and preference optimisation. Curated supervised safety tuning can be effective when strong multimodal safety data is available [34], but later work shows that such tuning can also produce brittle shortcuts through spurious textual correlations [35]. Unlearning-based methods attempt to remove harmful behaviour more directly [35, 36], while a recent work also points towards representation-level editing of harmful or sensitive multimodal concepts [39]. Preference-based methods refine the model using principles, segment-level corrections, or adversarially mined negatives [40–42]. The assumptions still differ: supervised methods depend on the coverage and quality of curated safety data, unlearning depends on the ability to erase harmful behaviour without damaging useful capabilities, and preference optimisation depends on reliable evaluators and well-constructed comparison data. The broader lesson is that multimodal safety alignment cannot rely on simple refusal supervision alone. Stronger alignment requires removing harmful behaviour at its source, improving the model's reasoning over multimodal evidence, or correcting modality imbalance through richer feedback signals.

### 4.2. Reactive Defences

Reactive defences operate at inference time. They detect, transform, or constrain unsafe behaviour without changing the base model during deployment. In multimodal systems, they can intervene before the model processes the input, during internal

reasoning, or while decoding the final response. We divide them into two groups: Input sanitisation, which modifies suspicious inputs before they reach the main reasoning stack, and model guardrails, which detect, verify, filter, or steer behaviour closer to the model's internal decision process (Figure 3).

*Input sanitisation*: Input sanitisation transforms potentially malicious inputs into forms that are safer for the model to process. In multimodal systems, this can happen either at the raw signal level, for example, by purifying perturbed pixels or jointly denoising image-text inputs, or at the prompt level by rewriting or projecting the input towards safer instructions or embeddings.

*Adversarial purification*. Adversarial purification reconstructs a safer version of the input before the target multimodal model performs joint reasoning. In vision–language systems, this line of work is motivated by the observation that small perturbations in visual inputs can distort hidden representations and induce unsafe or incorrect outputs. Recent purification-based defences, therefore, focus on restoring stable image or multimodal representations while preserving the semantic content needed for downstream reasoning [43, 44].

Fu et al. [43] propose DiffCAP, a diffusion-based purification method for VLMs. DiffCAP gradually injects Gaussian noise into an adversarial image until the embeddings of consecutive noisy images become sufficiently similar, and then applies reverse diffusion to recover a stable image representation. Its main advantage is not only stronger defence performance but also an adaptive stopping rule that reduces hyperparameter sensitivity and unnecessary diffusion steps. This makes purification more practical for VLM deployment because the denoising schedule follows changes in the model's embedding space rather than relying on a fixed diffusion horizon.

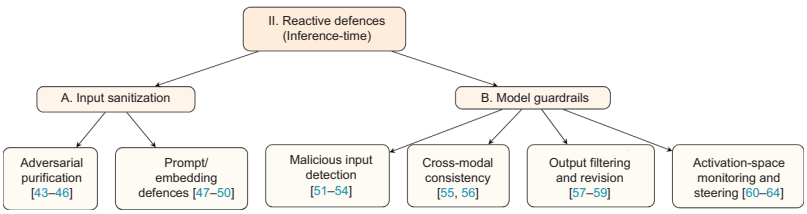


**Figure 3.** Taxonomy of reactive defences, dividing into input-sanitisation and model guardrails categories.

*Joint adversarial purification* (JAP): It extends purification to multimodal attacks. Instead of purifying each channel independently, JAP uses cross-modal interaction so that purified visual content helps recover the text, and purified text in turn supports visual denoising [44]. This joint strategy outperforms unimodal purification baselines against both white-box and transfer-based multimodal attacks on vision–language pre-training models. It also supports a key point in the taxonomy: once the attack itself is multimodal, purification that ignores cross-modal semantics is often insufficient.

Diffusion purification methods, such as ADBM and OSCP, remain relevant as technical precursors because they improve the quality and speed of image purification [45, 46]. ADBM replaces the standard reverse diffusion process with a learned bridge from diffused adversarial data back to clean examples, while OSCP distils diffusion purification into a single-step process guided by structural edge information. However, these methods are image-centric, so they are better viewed here as enabling techniques rather than core multimodal defences.

*Prompt/embedding defences*: A second family of input sanitisation methods modifies the textual instruction or its embedding before the model produces an answer. Wang et al. [47] propose AdaShield, which prepends a safety prompt to the multimodal input. Instead of relying only on a static prompt, AdaShield also introduces an adaptive framework in which an auxiliary defender model iteratively refines defence prompt for the current query. The results show that this strategy improves robustness against structure-based jailbreak attacks while preserving benign-task performance.

Jiang et al. [48] propose RapGuard, which first generates a multimodal safety rationale and then uses it to construct a scenario-specific defensive prompt. This addresses one of the main weaknesses of static guard prompts: they often fail to adapt to the specific risks introduced by different image–text combinations. RapGuard shows that lightweight prompting becomes more effective when it is paired with explicit multimodal reasoning about why the current input may be unsafe.

Other methods operate directly in embedding space, rather than on discrete text. Qiu et al. [49] propose embedding sanitiser (ES) for text-to-image generation. ES assigns a harmfulness score to each token and then selectively removes unsafe components from the prompt embedding while preserving as much benign meaning as possible. The paper argues that sanitising the embedding source

of unsafe generation is more robust than relying only on discrete blacklists or *post hoc* content filters. In a related direction, Chen et al. [50] propose BaThe, which embeds a virtual rejection prompt into soft text embeddings, so that harmful instructions behave like triggers for refusal rather than compliance. BaThe is relevant here because it shows that input sanitisation can be continuous and learned, not only rule-based or lexical.

The papers categorised under input sanitisation show a clear shift from static keyword filtering towards more adaptive transformation of the input. In purification-based methods, the trend is from generic signal denoising towards multimodal or model-aware reconstruction, as seen in DiffCAP and JAP [43, 44]. In prompting-based methods, the trend is from fixed defence prompts towards rationale-aware prompting and continuous embedding manipulation [47–50]. The assumptions still differ. Purification methods assume that adversarial perturbations can be removed without destroying task-relevant semantics, while prompt and embedding defences assume that the base model's aligned behaviour can still be reactivated if the instruction is reframed appropriately. The broader lesson is that in multimodal systems, effective sanitisation often requires transforming the input at the representation level rather than relying solely on fixed surface-form filters.

*Model guardrails*: Model guardrails enforce safety during inference. Model guardrails screen inputs, check image–text agreement, revise unsafe outputs, or change internal states. Compared with input sanitisation, model guardrails act later in the pipeline and can catch failures that appear only after joint multimodal reasoning starts.

*Malicious input detection:* A first family of guardrails detects harmful inputs before or during generation. Oh et al. [51] propose UniGuard, a multimodal guardrail trained on both unimodal and cross-modal harmful signals. UniGuard applies the same guardrail across different MLLMs at inference time with low overhead. The design treats detection as a general safeguard rather than a defence against a single attack family.

Helff et al. [52] introduce LlavaGuard, a family of VLM-based safeguards for evaluating visual content against customisable safety policies. Llava-uard outputs a safety rating, violated categories, and a rationale. The same model family supports both dataset curation and generative model moderation, which makes LlavaGuard a strong example of policy-aware malicious input detection.

Chen et al. [53] propose SafeWatch for video guardrailing. SafeWatch encodes policy chunks in parallel to remove positional bias and prunes visual tokens according to policy relevance. The reported gains in accuracy and cost show that multimodal guardrails benefit from policy conditioning and selective perception inside the inference pipeline.

Yu et al. [54] propose Test-time IMmunisation (TIM), which combines inference-time detection with online response. TIM trains a gist token to detect jailbreak activity during inference and then performs safety fine-tuning with the detected jailbreak instruction and a refusal answer. TIM partly crosses the line between pure inference and test-time adaptation, but the online trigger still makes TIM a reactive guardrail.

*Cross-modal consistency verification*: A second family of guardrails checks whether image and text remain semantically coherent. Zhao et al. [55] propose SafeCLIP, which uses the inherent multimodal alignment of LVLMs for zero-shot toxic image detection. SafeCLIP projects the discarded CLS token of CLIP into the text space and compares the result with toxic descriptors. The method detects harmful visual inputs without architecture changes and with low overhead. The result shows that the existing alignment structure already contains useful safety signals.

Ma et al. [56] propose ContextGuard-LVLM for detecting fine-grained contextual inconsistency between news images and text. ContextGuard-LVLM goes beyond entity matching and checks narrative, sentiment, and logical coherence. The setting targets multimodal misinformation rather than generic jailbreak defence, but the method shows that consistency checks can act as guardrails when the attack relies on subtle semantic mismatch rather than explicit toxicity.

*Output filtering and revision*: A third family changes generation during or immediately after decoding. Gao et al. [57] propose Constitutional Calibration (CoCA), which measures the logit difference between decoding with and without a safety principle and adds the calibrated difference back into decoding. Their idea is simple: multimodal models still retain some safety awareness from the LLM backbone, but the modality gap weakens that awareness. CoCA restores part of that awareness by amplifying safety-aware directions in the output distribution.

Ghosal et al. [58] propose Immune, an inference-time defence that guides decoding with a safety reward model. Immune does not

retrain the MLLM. It instead reformulates safety alignment as a KL-regularised decoding problem and steers generation towards safe responses token by token. The formulation makes Immune a clear example of output-side inference-time alignment.

Ding et al. [59] introduce evaluating then aligning safety (ETA), a two-phase framework that first evaluates visual inputs and candidate outputs for safety and then aligns generation with an interference prefix and sentence-level best-of-*N* search. ETA combines pre-generation evaluation with output-side revision, reducing unsafe behaviour while preserving helpfulness.

*Activation space monitoring and steering*: The fourth family of guardrails acts directly on hidden-state space. Some methods monitor internal activations for safety-relevant signals. Other methods steer intermediate representations away from harmful regions.

Wang et al. [60] propose ASTRA, a clear example of inference-time activation steering for VLM jailbreak defence. ASTRA constructs harmful directions from adversarial visual tokens and subtracts the harmful component from hidden activations during inference. The method shows that activation steering can suppress harmful multimodal behaviour without rewriting the raw input or changing model parameters.

Jiang et al. [61] propose HiddenDetect, which monitors hidden activations and trains lightweight probes to distinguish safe inputs from unsafe or jailbroken inputs. HiddenDetect shows that LVLMs already encode safety-relevant information in their internal states, and that simple probes can use these signals to enable a tuning-free guardrail.

Some methods act on more structured internal states. Jiang et al. [62] propose Dynamic Token Reweighting (DTR), which estimates a safety-relevant distributional shift induced by visual modality and then reweights visual tokens in the key-value cache to suppress adversarial influence. Chen et al. [63] propose SafePTR, which identifies a small set of harmful multimodal tokens at vulnerable early-middle layers, prunes those tokens, and restores benign features at later layers. Both methods show that multimodal jailbreak defence benefits from targeting the internal states where harmful information becomes causally influential.

Darabi et al. [64] propose EigenShield, which filters adversarial noise by separating causal and correlational subspaces in

high-dimensional VLM embeddings with Random Matrix Theory. EigenShield does not use explicit steering vectors, but still belongs in the same taxonomy branch because the method defends by monitoring and transforming internal representations rather than only filtering inputs or logits.

The papers categorised under Model Guardrails show a shift from static blocking to layered inference-time control. Malicious input detectors screen explicit harmful content or policy violations before full generation begins [51–54]. Cross-modal consistency methods check whether image and text jointly make sense, which helps against semantic manipulation and contextual attacks [55, 56]. Output revision methods require access to decoding logits, safety reward models, or candidate generations and improve safety by steering or selecting outputs online [57–59]. Activation-space methods require hidden activations, token states, or spectral subspaces and intervene more directly in the model internals [60–64]. Robust multimodal guardrails rarely succeed by checking only the surface form of the input. Stronger guardrails combine semantic screening, internal-state monitoring, and decoding control.

## 5. Illustrative Case Study: LLaVA

LLaVA is a useful case for defence taxonomy because many multimodal safety papers test on LLaVA and because LLaVA shows both training- and inference-time failures. Proactive methods reduce the base risk. Zong et al. [34] use VLGuard to fine-tune refusal behaviour on unsafe image–text inputs. Chen et al. [35] show that standard safety fine-tuning can remain brittle. Safety fine-tuning often learns shallow cues instead of robust safety features. Chen et al. [35] therefore propose machine unlearning as a stronger way to remove harmful behaviour.

Reactive methods remain necessary because visual inputs still bypass static alignment. LlavaGuard screens visual inputs against explicit safety policies before generation [52]. CoCA and Immune act during decoding. CoCA recalibrates output logits with a safety-aware signal, and Immune guides decoding with a safety reward model [57, 58]. Activation-space methods go one step deeper. ASTRA steers hidden activations away from harmful directions, and HiddenDetect monitors hidden states to detect jailbreak attempts during inference [60, 61]. The full pattern is clear. Proactive alignment lowers the starting risk. Reactive guardrails catch failures that still emerge during multimodal reasoning.

### 5.1. Critical Analysis and Limitations

No single defence protects the full multimodal pipeline. Input sanitisation changes the input before joint reasoning starts. Model guardrails screen, revise, or steer behaviour during inference [43, 47, 58, 61]. Each family targets a different failure mode. A practical system, therefore, needs several layers. Proactive alignment lowers the base risk. Reactive defences catch failures that still appear at inference time [34, 51, 57, 60].

Adaptive attacks remain a major gap. Many defence papers test a fixed attacker that does not optimise against the defence rule. ADBM makes this point for diffusion purification and argues that earlier evaluations often used weak adaptive attacks [45]. EigenShield shows it for heuristic inference-time defences [64]. A defence that blocks a fixed benchmark often degrades once the attacker knows the safeguard. Current methods usually raise attacker cost but rarely close the attack surface.

Deployability is also uneven. AdaShield and LlavaGuard work outside the base model and fit closed systems more easily [47, 52]. CoCA needs decoding logits. DTR needs the key-value cache. ASTRA and HiddenDetect need hidden activations [57, 60–62]. Many commercial APIs do not expose those signals. Strong results on open models, therefore, do not always transfer to real deployments.

Transfer across domains is still limited. SafeCLIP targets toxic images [55]. SafeWatch targets video moderation [53]. ContextGuard-LVLM targets image–text mismatch in news verification [56]. Safety Mirage shows the same problem in proactive alignment: shallow training signals do not generalise well [35]. The main conclusion is simple. Multimodal defence is a system-design problem. Standalone safeguards remain weak. Layered defences are more robust.

### 5.2. Evaluating Adversarial Robustness and Safety

A defence claim is only as strong as the evaluation behind it. Weak attacks, incomplete protocols, and easy benchmarks can make a weak defence look strong [65, 66]. Multimodal systems add new failure modes. Unsafe meaning can appear only after the fusion of modalities. Long conversations can reveal harms that single-turn prompts miss. Tool use, memory, and sequential planning can enable a single injected signal to affect many later steps [10, 67–69]. A rigorous evaluation, therefore, needs a protocol, not only a score table.

### 5.3. Why Multimodal Evaluation Is Different?

Multimodal evaluation differs from unimodal evaluation for three reasons. First, unsafe meaning may appear only after the model combines modalities. MSTS was designed around exactly that setting: each prompt becomes unsafe only when the image and text are interpreted together [67]. Second, safety depends on context. Multimodal situational safety shows that the same language query can be safe in one scene and unsafe in another, so the model must reason about the situation, rather than react to keywords alone [69]. Third, many deployed systems are no longer single-step predictors. Agents use tools, store memory, revisit earlier content, and act over time. ARE was introduced because prompt-response evaluation does not cover that setting [10].

Recent system cards support the same conclusion. Claude Sonnet 4.6 reports single-turn tests, higher-difficulty tests, ambiguous-context tests, multi-turn tests, agentic safety tests, prompt-injection tests, and adaptive-attacker tests across coding, browser, and computer-use settings [70]. GPT-5.4 Thinking reports challenging production-style benchmarks, dynamic multi-turn evaluations, representative-prompt re-sampling, multi-turn jailbreak evaluation, prompt-injection tests, image-input safety tests, destructive-action avoidance, and preservation of user work in long computer-use traces [71]. Real deployment is broader than static benchmark scoring. Evaluation must reflect that broader setting.

## 6. A Practical Evaluation Protocol

The Background section already defined the threat model. Evaluation should now instantiate that threat model for the specific system under test. The protocol below summarises the main questions that a defence paper should answer.

- What system is being evaluated? Evaluation should state whether the target is a vision–language chat model, a text-to-image model, an audio–language model, or an agent that uses tools and acts over time. The evaluation protocol should match the deployment setting. Chat assistants need open-ended prompts and dialogue. Tool-using agents need full episodes, not only one-step responses [10, 69–71].
- What interaction horizon is being tested? A good evaluation should include both short prompts and long prompts. It should also include both single-turn and multi-turn interactions when the deployment setting allows dialogue. Multi-turn

settings matter because harmful behaviour may emerge only after repeated probing, gradual escalation, or self-contradiction. GPT-5.4 Thinking now uses dynamic multi-turn evaluations for mental health, emotional reliance, and self-harm for exactly that reason [71]. Agentic systems need an even longer horizon because a poisoned website, tool output, or memory update can influence many later actions [10, 70].

- What data sources are used? A useful evaluation should use complementary data rather than one narrow benchmark family. Real data matters because real user behaviour and real scenes often differ from synthetic prompts. Manipulated data matters because many attacks start from benign inputs. Synthetic data matters because attackers can use other generative models to encode malicious intent in distributions that the target encoder rarely saw during training. Text-to-image, text-to-speech, and audio generation systems can all produce adversarial carrier signals that are semantically rich and often out of distribution [72–74]. Benchmark count matters less than benchmark complementarity.
- How diverse is the attack suite? A useful evaluation should vary both the attack technique and the attack setting. The attack suite should include black-box attacks, white-box attacks when model access allows them, and adaptive attacks that target the defence itself [65, 66]. Claude Sonnet 4.6 reports robustness against adaptive attackers across several agentic surfaces [70]. GPT-5.4 Thinking reports a multi-turn jailbreak evaluation derived from red-teaming exercises, where the attacker probes, adapts, and escalates over the course of the conversation [71].
- What behaviour counts as safe? The protocol should define the target safe behaviour before scoring begins, especially for dual-use and high-risk domains. A benchmark with only clearly malicious prompts is easy to score and easy to saturate. Harder evaluation should include benign prompts, harmful prompts, and realistic edge cases such as dual-use requests, ambiguous context, and difficult help-seeking conversations [67, 68, 70, 71]. In dual-use settings, the evaluator should specify whether the correct model behaviour is simple refusal, justified refusal, safe redirection, partial benign assistance, or another policy-compliant response. Without that target, judges will score inconsistent behaviours inconsistently.
- How are multi-component defences analysed? Multi-component defences need ablations. If a defence combines several modules, the paper should show what each module contributes and whether one module weakens another one. ARE already shows

why this matters for agents: extra inference-time components can open new attack surfaces rather than only add robustness [10]. The same logic applies to guardrail stacks, prompt rewriters, and judges.

- Are safety, utility, and cost measured together? A model that refuses every prompt can lower attack success, but such a model is unusable. A useful evaluation should therefore report adversarial robustness, benign-task utility, false refusal rate, and computational overhead together [67, 68]. Agentic settings should also report step count, latency, and failure recovery, as these variables affect deployment cost and risk [10, 71].
- Are weaknesses reported explicitly? Researchers should report the settings in which the defence fails, not only the benchmarks in which it performs best. Practical defences need that information because deployment teams care about failure modes as much as they care about headline scores. A defence section should therefore include a short limitations analysis, and the appendix should include broader benchmark results whenever the main paper space is tight [65, 75].

## 7. Metrics, Repeated Sampling, and Judges for Open-Ended Evaluation

Attack success rate (ASR) remains the main robustness metric, but ASR needs a task-specific definition. Let $N$ be the number of evaluation cases, and let each attacked case be run $K$ times because generative outputs are stochastic. Formula for ASR is presented in Equation (2),

$$ASR = \frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K} S_{ij} \tag{2}$$

Where:

$$S_{ij} = \begin{cases} 1, & \textit{if trial } j \textit{ on case } i \textit{ is successful} \\ 0, & \textit{otherwise} \end{cases}$$

The success criterion behind $S_{ij}$ depends on the system. In JailBreakV, success means that an image-text jailbreak makes the model produce the target unsafe behaviour or unsafe response [76]. In SneakyPrompt, success means that the prompt bypasses the safety filter and still generates the disallowed image [72]. In ARE, success means that the agent completes the attacker's goal rather than the

user's goal [10]. A paper should define the success criterion explicitly before reporting ASR.

Repeated sampling is part of evaluation, not an optional extra. One run is not enough for a stochastic system. Claude Sonnet 4.6 reports iterative evaluation over multiple model snapshots and states that Anthropic sampled each snapshot for multiple times for agentic evaluations [70]. A practical evaluation should therefore run each case for several times, for example, five or 10 repetitions depending on cost, and should report the mean score with a 95% confidence interval across cases. Realistic evaluation should report uncertainty, not only a point estimate.

Robustness metrics need matching utility metrics. False refusal rate is essential because a model that refuses every prompt can drive ASR down while remaining unusable in practice. MMSafeAware makes that trade-off explicit by including both unsafe and over-safety subsets [68]. MSTS adds another important diagnostic category: *safe by accident*. A model may appear safe because it failed to understand the multimodal prompt, rather than because it recognised the harm [67]. A useful evaluation should, therefore, report at least four values together: ASR, benign-task utility, false refusal rate, and the rate of misunderstanding or safe-by-accident behaviour when the benchmark supports that distinction.

Open-ended systems need judges. Real systems do not answer only multiple-choice questions. Real systems answer open-ended prompts, carry dialogue, and act through tools. Scale, therefore, requires automated judging. CAA combines standard, model-based, and human evaluations for the same benchmark [73]. MSTS shows that automated safety classifiers still lag behind careful manual assessment [67]. SafeBench reaches a similar conclusion and proposes a jury-style protocol with multiple LLM judges to improve evaluation reliability [77]. Thus, LLM as a judge gives breadth, but human evaluation gives depth.

Judge output also matters. Binary labels are often too crude for multimodal safety. Some responses are fully safe or unsafe. Other responses are off-topic, partially compliant, weakly refusing, borderline, or safe by accident. A more useful judge should therefore output a severity score or a small graded label set rather than a single safe/unsafe bit. A rationale is also useful because it makes the decision auditable. Human validation should use multiple annotators when possible and should compare the judge to consensus

labels, rather than to a single rater. Large-scale studies can also use judge ensembles when one judge proves unstable.

### 7.1. Benchmarks by Evaluation Role

Benchmarks are most useful when grouped by what they isolate. A first group targets direct adversarial robustness. Zhao et al. [75] study targeted black-box attacks on open VLMs and showed that transferred adversarial examples, together with black-box queries, can achieve a high rate of targeted evasion. SneakyPrompt plays a similar role for text-to-image systems by measuring prompt-side filter bypass under realistic query budgets [72]. CAA extends the same logic to conversational audio systems [73].

A second group isolates harms that arise from multimodal meaning itself. MSTS tests hazards that emerge from image–text fusion [67]. MMSafeAware measures missed harms and over-refusal [68]. Multimodal situational safety evaluates whether a model can judge safety from the surrounding scene rather than from the query alone [69]. These benchmarks are complementary as each isolates a different failure mode.

A third group supports broader or more diagnostic auditing. Qiu et al. [74] benchmark multimodal robustness under distribution shift and add metrics, such as Multimodal Impact (MMI) and Missing Object Rate (MOR), to diagnose how perturbations change system behaviour across tasks. MMDT broadens the lens further by evaluating safety, hallucination, fairness, privacy, adversarial robustness, and out-of-distribution generalisation within a single framework [78]. ARE fills the agentic gap by evaluating full attack episodes against multimodal web agents [10]. No single benchmark covers all of these roles. Complementary benchmark choice matters more than raw benchmark count.

### 7.2. Evaluation Gaps and Minimum Reporting Checklist

Many evaluations have four major gaps. First, many studies still rely too heavily on fixed attacks and static prompts. Adaptive attackers often expose failures that benchmark-style attacks miss [65, 66]. Second, many evaluations still underweight deployment realism. Real systems use dialogue, tools, memory, and long-action chains, while many published benchmarks remain short and static [10, 69]. Third, judge reliability remains imperfect. Open-ended evaluation requires automated judging, but the strongest current

studies still validate judges against humans rather than blindly trusting a single automatic labeller [67, 73, 77]. Fourth, closed-model opacity limits direct white-box testing and makes realistic grey-box and query-based audits more important [75, 79].

The checklist below summarises a minimum reporting standard for defence evaluation.

- System and deployment setting: chat VLM, text-to-image model, audio-language model, or agent; single-turn, multi-turn, or full-episode evaluation.
- Interaction horizon: short prompts, long prompts, dialogue, tool use, memory updates, and sequential tasks when relevant.
- Policy target: explicit definition of the correct safe behaviour for harmful, benign, dual-use, and help-seeking cases.
- Data diversity: real data, manipulated data, and synthetic data chosen to cover complementary attack settings, rather than near-duplicate benchmark families.
- Attack diversity: black-box, white-box when possible, and adaptive attacks; multiple attack families rather than a single optimised recipe.
- Scoring stability: number of repetitions per case, aggregation rule, and 95% confidence intervals.
- Success definition: explicit task-specific definition of ASR.
- Utility metrics: benign-task performance and false refusal rate.
- Cost metrics: latency, memory, query count, and episode length when relevant.
- Judge protocol: judge model, judge scale, rationale format, human-calibration method, and use of judge ensembles if applicable.
- Ablations: contribution of each component for multi-stage or multi-module defences.
- Failure analysis: image-only, text-only, cross-modal, contextual, memory-mediated, and safe-by-accident failures whenever the protocol supports that breakdown.
- Limitations: failure cases across benchmarks, not only best-case results on the benchmark where the method performs best.

Multimodal evaluation is not a scoreboard problem but rather a protocol design problem. A useful evaluation matches the deployed system, uses complementary data and attacks, measures safety and utility together, validates automatic judges against human consensus, and reports uncertainty rather than one-shot scores. The field now has the pieces to do this well. The field still needs more consistent use of those pieces.

## 8. Future Research Directions

The next stage of multimodal robustness research should focus on a small set of practical problems that are close to deployment.

*Layered defence-in-depth for AI.* The field has many useful defence components, but it still lacks design rules for combining them into one reliable system. The current methods often protect only one stage of the pipeline, such as the input, the decoder, or the hidden state. Recent work on multimodal guardrails and agent security already points towards layered designs, including policy-aware guardrails, runtime validation, and memory isolation [51, 80, 81]. The main open question is not whether more layers help. The main open question is which layers complement each other, which layers create over-refusal on benign tasks, and which layers only add cost without adding real security.

*Adaptive defences against adaptive attackers.* Many current defences are static. Attackers are not. Once the attacker learns the defence rule, the attacker can often route around it. Recent work already moves towards self-evolving defence, for example by detecting jailbreaks during inference and updating the safety response online, or by steering decoding with a safety reward model instead of relying only on fixed training-time alignment [54, 58]. A useful next step is to build adaptive defences that update trust scores, intervention strength, or refusal policy online without drifting into instability, catastrophic forgetting, or excessive refusal.

*Secure multimodal agents, not only chat models.* Autonomy changes the risk profile. A chat model produces text. An agent reads external content, calls tools, stores memory, and takes actions over time. ARE shows that small visual perturbations can hijack multimodal web agents under realistic black-box conditions [10]. Agent Security Bench finds similar weaknesses across prompts, tools, and memory [82]. Multimodal situational safety shows that safe action often depends on the visual scene rather than the text instruction alone [69]. Near-term research should therefore focus on privilege control, action containment, tool-use policies, and recovery mechanisms for agents that can affect real systems.

*Provenance-aware memory and trust boundaries.* Memory is becoming a first-class attack surface. A poisoned message, tool output, or retrieved document can persist in memory and influence later decisions. Recent agent-security work already proposes memory isolation, dynamic validation, and explicit control-flow and data-flow

constraints around the model [80, 81]. A strong near-term direction is to build agent architectures in which memory items, tool outputs, and agent-to-agent messages carry source and trust metadata. That metadata should affect retrieval, planning, and action selection so that injected instructions do not silently propagate through the system.

*Evaluation that matches deployment.* The field needs shared evaluation protocols for open-ended, multi-turn, and agentic multimodal systems.

Static benchmark scores are useful, but static scores are not enough for systems that use dialogue, tools, memory, and long-horizon action. Recent benchmarks and system cards already move towards dynamic multi-turn testing, situational safety evaluation, and agentic safety audits [10, 67, 68, 70, 71]. A useful next step is to standardise those protocols so that new defence methods are compared under realistic, open-ended conditions rather than on the benchmark where they look strongest.

## 9. Conclusions

Multimodal AI systems now operate in settings where security failures can have real consequences, including autonomous navigation, medical decision support, and agentic task execution. Combining images, text, audio, and sensor data increases capability, but it also expands the attack surface. This paper reviewed the problem, focusing on adversarial threats, defence mechanisms, and robustness evaluation.

The review first examined the threat landscape. The discussion covered attacks on multimodal agents and attacks on systems that use less-studied inputs, such as audio, depth, and thermal data. Those settings show that multimodal security is no longer limited to the vision–language domain. The review then organised defence methods into two groups: proactive methods that improve robustness before deployment, and reactive methods that detect or limit harmful behaviour at inference time. This taxonomy gives a simple way to compare defences by where they act and which risks they address.

The paper also argued that evaluation must be treated as a protocol design problem, rather than a benchmark-counting exercise. A strong evaluation should align with the deployed system, test complementary attack settings, jointly measure safety and utility, and

use reliable judges for open-ended outputs. The proposed reporting checklist makes these requirements explicit and highlights current gaps in adaptive evaluation, agentic settings, and realistic deployment conditions.

The main conclusion is simple: No single defence solves multimodal adversarial robustness. Training- and inference-time measures address different parts of the problem and must work together in deployed systems. Progress now depends on three practical steps: building layered defence-in-depth for AI systems, developing adaptive defences against adaptive attackers, and securing multimodal agents with tools, memory, and long-horizon actions. Secure multimodal AI will require systems that remain safe not only on controlled benchmarks but also under realistic, open-ended, and action-taking conditions.

## References

[1] J. Zhang, Q. Yi, J. Sang, "Towards adversarial attack on vision–language pre-training models," arXiv: 2206.09391, 2022.

[2] A. Aich, C. K. Ta, A. Gupta, C. Song, S. V. Krishnamurthy et al., "GAMA: Generative adversarial multi-object scene attacks," arXiv: 2209.09502, 2022.

[3] H. Luo, J. Gu, F. Liu, P. Torr, "An image is worth 1000 lies: Adversarial transferability across prompts on vision–language models," arXiv: 2403.09766, 2024.

[4] X. Liu, Y. Zhu, Y. Lan, C. Yang, Y. Qiao, "Safety of multimodal large language models on images and text," arXiv: 2402.00357, 2024.

[5] C. Jiang, Z. Wang, M. Dong, J. Gui, "Survey of adversarial robustness in multimodal large language models," arXiv: 2503.13962, 2025.

[6] S. Kapoor, S.S. Girija, L. Arora, D. Pradhan, A. Shetgaonkar, A. Raj, "Adversarial attacks in multimodal systems: A practitioner's survey," arXiv: 2505.03084, 2025.

[7] M. Kowalczyk, K. Seweryn, M. Krzyszton, J. Kolodziej, "Adversarial robustness of multimodal machine learning models," in *Proceedings of the 39th International ECMS Conference on Modelling and Simulation (ECMS 2025)*, M. Scarpa, S. Cavalieri, S. Serrano, F.D. Vita, Eds. European Council for Modelling and Simulation, 2025, pp. 248–254, doi: 10.7148/2025-0248.

[8] G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov et al., "Multimodal neurons in artificial neural networks," *Distill*, 2021, doi: 10.23915/distill.00030.

[9] C. Zhang, M. Hu, W. Li, L. Wang, "Adversarial attacks and defenses on text-to-image diffusion models: A survey," arXiv: 2407.1586, 2024.

[10] C.H. Wu, R. Shah, J.Y. Koh, R. Salakhutdinov, D. Fried, A. Raghunathan, "Dissecting adversarial robustness of multimodal LM agents," arXiv: 2406.12814, 2025.

[11] L. Aichberger, A. Paren, Y. Gal, P. Torr, A. Bibi, "Attacking multimodal OS agents with malicious image patches," arXiv: 2503.10809, 2025.

[12] I. Gupta, D. Khachaturov, R. Mullins, "'I am bad': Interpreting stealthy, universal and robust audio jailbreaks in audio-language models," arXiv: 2502.00718, 2025.

[13] J. Zheng, C. Lin, J. Sun, Z. Zhao, Q. Li, C. Shen, "Physical 3D adversarial attacks against monocular depth estimation in autonomous driving," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, 2024, pp. 24452–24461, doi: 10.1109/CVPR52733.2024.02308.

[14] Y. Gong, Q. Zeng, D. Xu, Z. Wang, M. Jiang, "Cross-modality attack boosted by gradient-evolutionary multiform optimization," arXiv: 2409.17977, 2024.

[15] Y. Chen, X. Hu, K. Yin, J. Li, S. Zhang, "Evaluating the robustness of multimodal agents against active environmental injection attacks," arXiv: 2502.13053, 2025.

[16] L. Wang, Z. Ying, T. Zhang, S. Liang, S. Hu, et al., "Manipulating multimodal agents via cross-modal prompt injection," arXiv: 2504.14348, 2025.

[17] A.S. Patlan, P. Sheng, S.A. Hebbar, P. Mittal, P. Viswanath, "Real AI agents with fake memories: Fatal context manipulation attacks on Web3 agents," arXiv: 2503.16248, 2025.

[18] B. Ma, H. Guo, Z.J. Luo, R. Duan, "Audio jailbreak attacks: Exposing vulnerabilities in speech GPT in a white-box framework," arXiv: 2505.18864, 2025.

[19] C.W. Chiu, L. Huang, B. Li, H. Chen, K.-K.R. Choo, "'Do as I say not as I do': A semi-automated approach for jailbreak prompt attack against multimodal LLMs," arXiv: 2502.00735, 2025.

[20] Z. Zhang, S. Liang, D. Shimada, C. Xu, "Rethinking audio-visual adversarial vulnerability from temporal and modality perspectives," arXiv: 2502.11858, 2025.

[21] H.-T. Luong, H. Li, L. Zhang, K.A. Lee, E.S. Chng, "LlamaPartialSpoof: An LLM-driven fake speech dataset simulating disinformation generation," arXiv: 2409.14743, 2025.

[22] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall et al., "AV-deepfake 1M: A large-scale LLM-driven audio-visual deepfake dataset," arXiv: 2311.15308, 2024.

[23] J. Tu, H. Li, X. Yan, M. Ren, Y. Chen et al., "Exploring adversarial robustness of multi-sensor perception systems in self driving," arXiv: 2101.06784, 2022.

[24] M. Byra, G. Styczynski, C. Szmigielski, P. Kalinowski, L. Michalowski, R. Paluszkiewicz, et al., "Adversarial attacks on deep learning models for fatty liver disease classification by modification of ultrasound image reconstruction method," in *2020 IEEE International Ultrasonics Symposium (IUS)*. Las Vegas, NV, 2020, pp. 1–4, doi: 10.1109/IUS46767.2020.9251568.

[25] X. Huang, X. Wang, H. Zhang, Y. Zhu, J. Xi et al., "Medical MLLM is vulnerable: Cross-modality jailbreak and mis-matched attacks on medical multimodal large language models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, pp. 3797–3805, 2025, doi: 10.1609/aaai.v39i4.32396.

[26] W. Zhou, S. Bai, D.P. Mandic, Q. Zhao, B. Chen, "Revisiting the adversarial robustness of vision–language models: A multimodal perspective," arXiv: 2404.19287, 2024.

[27] F. Waseda, A.T. de Pablos, I. Echizen, "Multimodal adversarial defense for vision–language models by leveraging one-to-many relationships," arXiv: 2405.18770, 2025.

[28] L. Lu, X. Gu, S. Pang, S. Liang, H. Zhu et al., "Adversarial training for multimodal large language models against jailbreak attacks," arXiv: 2503.04833, 2025.

[29] K. Yang, W.-Y. Lin, M. Barman, F. Condessa, Z. Kolter, "Defending multi-modal fusion models against single-source adversaries," arXiv: 2206.12714, 2022.

[30] Y.-D. Tsai, T.-Y. Yen, K.-T. Liao, S.-D. Lin, "Enhance modality robustness in text-centric multimodal alignment with adversarial prompting," arXiv: 2408.09798, 2024.

[31] Y. Wang, H. Fu, W. Zou, J. Jia, "MMCert: Provable defense against adversarial attacks to multi-modal models," arXiv: 2403.19080, 2024.

[32] Z. Huang, W. Chu, L. Li, C. Xu, B. Li, "COMMIT: Certifying robustness of multi-sensor fusion systems against semantic attacks," arXiv: 2403.02329, 2024.

[33] Z. Yang, Y. Wei, C. Liang, D. Hu, "Quantifying and enhancing multi-modal robustness with modality preference," arXiv: 2402.06244, 2024.

[34] Y. Zong, O. Bohdal, T. Yu, Y. Yang, T. Hospedales, "Safety fine-tuning at (almost) no cost: A baseline for vision large language models," arXiv: 2402.02207, 2024.

[35] Y. Chen, Y. Yao, Y. Zhang, B. Shen, G. Liu, S. Liu, "Safety mirage: How spurious correlations undermine VLM safety fine-tuning," arXiv: 2503.11832, 2025.

[36] T. Chakraborty, E. Shayegani, Z. Cai, N. Abu-Ghazaleh, M. S. Asif et al., "Cross-modal safety alignment: Is textual unlearning all you need?" arXiv: 2406.02575, 2024.

[37] Y. Ding, L. Li, B. Cao, J. Shao, "Rethinking bottlenecks in safety fine-tuning of vision–language models," arXiv: 2501.18533, 2025.

[38] W. Lu, H. Peng, H. Zhuang, C. Chen, Z. Zeng, "SEA: Low-resource safety alignment for multimodal large language models via synthetic embeddings," arXiv: 2502.12562, 2025 [cs.CL].

[39] H. Chen, J. Li, Y. Zhang, J. Bi, Y. Xia et al., "AUVIC: Adversarial unlearning of visual concepts for multi-modal large language models," arXiv: 2511.11299, 2025.

[40] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion et al., "Constitutional AI: Harmlessness from AI feedback," arXiv: 2212.08073, 2022.

[41] T. Yu, Y. Yao, H. Zhang, T. He, Y. Han et al., "Rlhf-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, 2024, pp. 13 807–13 816, doi: 10.1109/CVPR52733.2024.01310.

[42] C. Liu, T. Xiong, Y. Chen, R. Chen, Y. Wu et al., "Modality-balancing preference optimization of large multimodal models by adversarial negative mining," arXiv: 2506.08022, 2025.

[43] J. Fu, Y. Wu, Y. Chen, K. Peng, X. Zhang et al., "DiffCAP: Diffusion-based cumulative adversarial purification for vision–language models," arXiv: 2506.03933, 2025.

**[44]** Q. Li, Y. Wang, W. Hu, R. Hong, "Joint adversarial purification: Mitigating the threat of multimodal adversarial examples," in *Proceedings of the 2025 International Conference on Multimedia Retrieval* (Ser. ICMR '25). Chicago, IL: Association for Computing Machinery (ACM), 2025, pp. 779–787, doi: 10.1145/3731715.3733367.

**[45]** X. Li, W. Sun, H. Chen, Q. Li, Y. Liu et al., "ADBM: Adversarial diffusion bridge model for reliable adversarial purification," arXiv: 2408.00315, 2025.

**[46]** C.T. Lei, H.M. Yam, Z. Guo, Y. Qian, C.P. Lau, "Instant adversarial purification with adversarial consistency distillation," arXiv: 2408.17064, 2025.

**[47]** Y. Wang, X. Liu, Y. Li, M. Chen, C. Xiao, "AdaShield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting," arXiv: 2403.09513, 2024.

**[48]** Y. Jiang, Y. Tan, X. Yue, "RapGuard: Safeguarding multimodal large language models via rationale-aware defensive prompting," arXiv: 2412.18826, 2024.

**[49]** H. Qiu, G. Chen, M. Zhang, X. Zhang, X. You, M. Yang, "Safe text-to-image generation: Simply sanitize the prompt embedding," arXiv: 2411.10329, 2025.

**[50]** Y. Chen, H. Li, Y. Zhang, Z. Zheng, Y. Song, B. Hooi, "BaThe: Defense against the jailbreak attack in multimodal large language models by treating harmful instruction as backdoor trigger," arXiv: 2408.09093, 2025.

**[51]** S. Oh, Y. Jin, M. Sharma, D. Kim, E. Ma et al., "UniGuard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models," arXiv: 2411.01703, 2025.

**[52]** L. Helff, F. Friedrich, M. Brack, K. Kersting, P. Schramowski, "LlavaGuard: An open VLM-based framework for safeguarding vision datasets and models," arXiv: 2406.05113, 2025.

**[53]** Z. Chen, F. Pinto, M. Pan, B. Li, "SafeWatch: An efficient safety-policy following video guardrail model with transparent explanations," arXiv: 2412.06878, 2024.

**[54]** Y. Yu, Y. Wang, R. He, J. Liang, "Test-time immunization: A universal defense framework against jailbreaks for (multimodal) large language models," arXiv: 2505.22271, 2025.

**[55]** W. Zhao, Z. Li, Y. Li, J. Sun, "Zero-shot defense against toxic images via inherent multimodal alignment in LVLMs," arXiv: 2503.00037, 2025.

**[56]** S. Ma, Q. Wu, R. Jiang, F. Burns, "ContextGuard–LVLM: Enhancing news veracity through fine-grained cross-modal contextual consistency verification," arXiv: 2508.06623, 2025.

**[57]** J. Gao, R. Pi, T. Han, H. Wu, L. Hong et al., "CoCA: Regaining safety-awareness of multimodal large language models with constitutional calibration," arXiv: 2409.11365, 2024.

**[58]** S.S. Ghosal, S. Chakraborty, V. Singh, T. Gua, M. Wang, A. Beirami, "Immune: Improving safety against jailbreaks in multi-modal LLMs via inference-time alignment," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, 2025, pp. 25038–25049, doi: 10.1109/CVPR52734.2025.02331.

**[59]** Y. Ding, B. Li, R. Zhang, "ETA: Evaluating then aligning safety of vision language models at inference time," arXiv: 2410.06625, 2025.

**[60]** H. Wang, G. Wang, H. Zhang, "Steering away from harm: An adaptive approach to defending vision language model against jailbreaks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, 2025, pp. 29947–29957, doi: 10.48550/arXiv.2411.16721.

**[61]** Y. Jiang, X. Gao, T. Peng, Y. Tan, X. Zhu et al., "HiddenDetect: Detecting jailbreak attacks against multimodal large language models via monitoring hidden states," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, M.T. Pilehvar, Eds. Vienna: Association for Computational Linguistics (ACL), 2025, pp. 14880–14893, doi: 10.18653/v1/2025.acl-long.724.

**[62]** T. Jiang, J. Liang, R. Zhu, J. Zhou, F. Ma, T. Wang, "Robustifying vision-language models via dynamic token reweighting," arXiv: 2505.17132, 2025.

**[63]** B. Chen, X. Lyu, L. Gao, J. Song, H.T. Shen, "SafePTR: Token-level jailbreak defense in multimodal LLMs via prune-then-restore mechanism," arXiv: 2507.01513, 2025.

**[64]** N. Darabi, D. Naik, S. Tayebati, D. Jayasuriya, R. Krishnan, A.R. Trivedi, "EigenShield: Causal subspace filtering via random matrix theory for adversarially robust vision-language models," arXiv: 2502.14976, 2025.

**[65]** N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber et al., "On evaluating adversarial robustness," arXiv: 1902.06705, 2019 [cs.LG].

**[66]** A. Athalye, N. Carlini, D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," arXiv: 1802.00420, 2018.

**[67]** P. Röttger, G. Attanasio, F. Friedrich, J. Goldzycher, A. Parrish et al., "MSTS: A multimodal safety test suite for vision-language models," arXiv: 2501.10057, 2025.

**[68]** W. Wang, X. Liu, K. Gao, J. Huang, Y. Yuan et al., "Can't see the forest for the trees: Benchmarking multimodal safety awareness for multimodal LLMs," arXiv: 2502.11184, 2025.

**[69]** K. Zhou, C. Liu, X. Zhao, A. Compalas, D. Song, X.E. Wang, "Multimodal situational safety," arXiv: 2410.06172, 2025.

**[70]** Anthropic. (Feb. 17, 2026). *System card: Claude Sonnet 4.6*, [Online]. Available: https://www-cdn.anthropic.com/bbd8ef16d70b7a1665f14f306ee88b53f686aa75.pdf. [Accessed: Apr. 23, 2026].

**[71]** OpenAI. (Mar. 5, 2026). *GPT-5.4 Thinking system card*, [Online]. Available: https://deploymentsafety.openai.com/gpt-5-4-thinking/gpt-5-4-thinking.pdf. [Accessed: Apr. 23, 2026].

**[72]** Y. Yang, B. Hui, H. Yuan, N. Gong, Y. Cao, "SneakyPrompt: Jail breaking text-to-image generative models," arXiv: 2305.12082, 2023.

**[73]** W. Yang, Y. Li, M. Fang, Y. Wei, L. Chen, "Who can withstand Chat-Audio attacks? An evaluation benchmark for large audio-language models," arXiv: 2411.14842, 2025.

[74] J. Qiu, Y. Zhu, X. Shi, F. Wenzel, Z. Tang et al., "Benchmarking robustness of multimodal image-text models under distribution shift," arXiv: 2212.08044, 2024.

[75] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li et al., "On evaluating adversarial robustness of large vision-language models," arXiv: 2305.16934, 2023.

[76] W. Luo, S. Ma, X. Liu, X. Guo, C. Xiao, "JailBreakV: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks," arXiv: 2404.03027, 2024.

[77] Z. Ying, A. Liu, S. Liang, L. Huang, J. Guo et al., "SafeBench: A safety evaluation framework for multimodal large language models," arXiv: 2410.18927, 2024.

[78] C. Xu, J. Zhang, Z. Chen, C. Xie, M. Kang et al., "MMDT: Decoding the trustworthiness and safety of multimodal foundation models," arXiv: 2503.14827, 2025.

[79] I. Evtimov, R. Howes, B. Dolhansky, H. Firooz, C.C. Ferrer, "Adversarial evaluation of multimodal models under realistic gray box assumption," arXiv: 2011.12902, 2021.

[80] H. Li, X. Liu, H.-C. Chiu, D. Li, N. Zhang, C. Xiao, "Drift: Dynamic rule-based defense with injection isolation for securing LLM agents," arXiv: 2506.12104, 2026.

[81] E. Debenedetti, I. Shumailov, T. Fan, J. Hayes, N. Carlini et al., "Defeating prompt injections by design," arXiv: 2503.18813, 2025.

[82] H. Zhang, J. Huang, K. Mei, Y. Yao, Z. Wang et al., "Agent security bench (ASB): Formalizing and benchmarking attacks and defenses in LLM-based agents," arXiv: 2410.02644, 2025.