Adversarial Attacks on Multimodal Large Language Models: A Comprehensive Survey

Anonymous authors
Paper under double-blind review

Abstract

Multimodal large language models (LLMs) integrate and process information from multiple modalities such as text, images, audio, and video, enabling complex tasks such as audio translation and visual question answering. While powerful, this complexity introduces novel vulnerabilities to sophisticated adversarial attacks. This survey paper provides a comprehensive overview of this rapidly expanding field, systematically categorizing attacks that range from manipulations of single modalities (e.g., perturbed images or audio) to those exploiting cross-modal interactions. We overview how these attacks exploit weaknesses in model fusion, attention mechanisms, and representation learning and provided analyses on their potential for real-world consequences.

1 Introduction

The rapid development of Large Language Models (LLMs) marks a significant leap in AI (Achiam et al., 2023). By integrating and processing information from diverse modalities, such as text, images, and audio, in various combinations, these models mimic human perception to achieve a more holistic understanding of the world (Li et al., 2023). This advanced capability enables them to tackle complex tasks previously beyond reach, driving breakthroughs in areas such as image captioning (Vinyals et al., 2015), visual question answering (Antol et al., 2015), audio-visual scene analysis (Arandjelović & Zisserman, 2017), robotics (Mon-Williams et al., 2025), and more. The power of these models stems from their scalable attention-based architectures (Vaswani et al., 2017) and effective training recipes: pre-training on vast datasets, followed by post-training alignment on a multitude of tasks to suit downstream use cases, thereby achieving state-of-the-art performance. However, the very complexity that fuels the power of LLMs also introduces novel and intricate vulnerabilities (Wen et al., 2025) (Goodfellow, 2014). As illustrated in Figure 1, unlike text-only LLMs, multimodal LLMs present an expanded attack surface where threats can arise not only from weaknesses within individual modality processing but, crucially, from the complex interplay and fusion mechanisms between the combined modalities (Baltrušaitis et al., 2018). These cross-modal vulnerabilities can manifest through several key attack vectors, some of which are as follows:

- **Cross-modal prompt injection** (Bagdasaryan et al., 2023) exploits LLMs' instruction-following nature by embedding malicious commands within non-textual modalities. Attackers can hide instructions in images or audio that the model interprets as textual commands, effectively hijacking behavior without directly manipulating text prompts.

- **Fusion mechanism attacks** target how the core process LLMs integrate information from multiple modalities, exploiting vulnerabilities in how features from different sources are combined and aligned. These attacks can disrupt the fusion process, causing the model to misinterpret or improperly weigh multimodal inputs. For example, the Collaborative Multimodal Interaction Attack (CMI-Attack) (Fu et al., 2024) manipulates both text embeddings and image gradients simultaneously.

- **Adversarial illusions** exploit the shared embedding space where LLMs align representations across modalities. Attackers can craft inputs in one modality that appear benign but whose embeddings become deceptively aligned with unrelated concepts from another modality, causing the model to hallucinate false semantic connections (Bagdasaryan et al., 2024).

As such powerful models are increasingly deployed in real-world applications, including safety-critical systems, understanding their susceptibility to adversarial manipulation becomes paramount.

The landscape of attacks targeting these LLMs is evolving rapidly, encompassing a range of techniques. These include adversarial perturbations designed to cause misclassification or erroneous outputs (Goodfellow et al., 2014), jailbreak attacks that bypass safety alignments (Wei et al., 2023), prompt injection methods that hijack model behavior (Perez & Ribeiro, 2022), and data poisoning strategies that corrupt the model during training (Gu et al., 2017). Given the increasing sophistication and potential impact of these attacks, a systematic understanding of the current threat landscape is urgently needed.

While recent efforts have begun to map the field with surveys on general LLM attacks (Shayegani et al., 2023b) and on specific pairs of modalities, such as vision-language models (Liu et al., 2024a), a deeper analytical connection between documented attacks and inherent vulnerabilities is missing. Valuable new overviews like Kapoor et al. (2025) offer a broad, practitioner-focused perspective on the threat landscape. And, the work by (Jiang et al., 2025) offers a systematic review organized by modality, which serves as an excellent catalogue of existing techniques. However, a critical gap remains in systematically linking what attacks are possible to why they succeed. To address this, our survey introduces a rigorous framework that is modality-agnostic and organizes the entire landscape of attacks based on core adversarial strategies. It classifies adversarial attacks along two orthogonal dimensions, the attack vector and the attacker's knowledge, and explicitly connects these methods to the underlying LLM vulnerabilities they exploit. This moves from a catalogue of the attacker's actions to an analysis of the system's inherent weaknesses, answering why these attacks are successful.

The remainder of this survey is organized as follows: Section 2 lays the groundwork by introducing multimodal LLMs, focusing on their core architecture and mathematical representation in the context of adversarial attacks. Section 3 introduces our novel two-dimensional taxonomy of adversarial attacks, classifying them along the orthogonal axes of attack vector - the ’what’, and attacker knowledge - the ’how’. Following this, Section 4 provides a systematic analysis of the underlying LLM vulnerabilities, shifting the focus from the attacker's methods to the system's inherent weaknesses that enable these attacks. To ground this analysis, the Appendix provides a comprehensive characterization of representative attacks from the literature, mapping them to our proposed framework and also our survey methodology. Finally, Section 5 concludes the paper by discussing our key findings, highlighting open challenges, and suggesting future research directions.

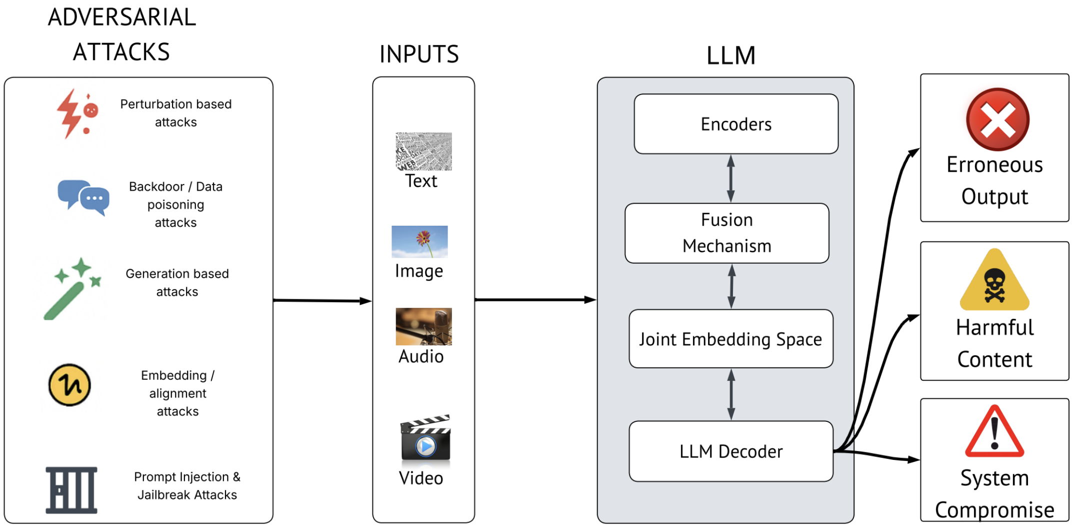


Figure 1: A high-level overview of the adversarial attack landscape for Multimodal LLMs.

2 Background: Understanding LLMs in the Context of Adversarial Attacks

This section lays the groundwork for understanding how multimodal LLMs are susceptible to adversarial attacks. We will briefly introduce LLM architectures and their operational principles, focusing on aspects that become relevant when discussing vulnerabilities. We then delve into the mathematical formulation of LLMs as it pertains to the generation and effect of adversarial perturbations, and finally, we examine key components as distinct attack surfaces.

2.1 LLM Architectures and Aspects Exploited by Attacks

Multimodal LLMs are designed to process and integrate information from multiple input types, commonly including text, images, audio & video (Li et al., 2023). Their core capability lies in generating coherent and contextually relevant outputs based on this fused multimodal understanding, enabling tasks like visual question answering, image/video captioning, and audio-visual scene interpretation. From an adversarial perspective, several architectural and operational characteristics of LLMs are particularly relevant:

- **Modality-Specific Encoders (E_i):** To process diverse inputs, LLMs often employ specialized encoders for each input type: for images, these include Vision Transformers (ViTs) (Dosovitskiy et al., 2020; Radford et al., 2021) and Convolutional Neural Networks (CNNs); for text, Transformer-based encoders (Devlin et al., 2019); and for audio, a range of models from RNNs (Baevski et al., 2020), to the increasingly prevalent Transformer-based architectures that transform raw inputs into high-dimensional vector representations. Their primary vulnerability stems from often being pre-trained on large unimodal datasets, allowing them to inherit adversarial weaknesses from these source models (Papernot et al., 2016). As a result, adversarial examples originally crafted for standard image classifiers (e.g., by adding imperceptible noise to pixels) (Goodfellow et al., 2014) or audio classifiers (e.g., through small waveform perturbations) (Carlini & Wagner, 2018) can successfully fool the respective encoder of the LLM. This propagates erroneous or misleading feature representations from the manipulated modality into the subsequent fusion stage, undermining the LLMs' overall understanding.

- **Fusion Mechanisms ($f_{\textbf{fuse}}$) and Cross-Modal Alignment:** A critical component is the fusion mechanism, which integrates the vector representations from different modalities. This can range from simple concatenation or pooling to more complex techniques like cross-modal attention, where information from one modality dynamically guides the processing of another (Vaswani et al., 2017; Lu et al., 2019). This intricate process of integrating information represents a significant point of potential failure. Disruptions in the fusion process, or misalignment in the learned cross-modal combination can lead the model to incorrectly prioritize, misinterpret, or fail to reconcile information from different modalities, especially when one is adversarially manipulated. Attacks manifest this vulnerability primarily through cross-modal perturbations (Dou et al., 2023), where subtle alterations in one modality (e.g., text) are designed to mislead the interpretation of another (e.g., an image), or vice-versa.

- **Attention Mechanisms:** Transformers, central to many LLMs, rely heavily on attention mechanisms (Vaswani et al., 2017). These mechanisms weigh the importance of different parts of the input via key-query alignment, both within a single modality (self-attention) and between different modalities (cross-modal attention, often part of the fusion mechanism). The way attention is distributed is a key target, as attackers can craft inputs that exploit or manipulate these attention scores (Wang et al., 2024b). For example, an attack might introduce features that unduly capture the model's attention, drawing focus to misleading information, or conversely, it might try to suppress attention to critical benign features, thereby derailing the model's reasoning process and its ability to correctly fuse multimodal information.

- **Joint Representation Space ($Z_{\textbf{joint}}$):** After fusion, information from multiple modalities is often encoded in a joint embedding space. This space is designed to capture complex inter-modal relationships and dependencies, providing a unified representation for downstream processing. The vulnerability here lies in the potential for direct manipulation of this abstract space. If attackers gain an understanding of how this high-dimensional space is structured (e.g., through model inversion (Fredrikson et al., 2015) or probing), they can craft inputs whose fused representations are pushed into "malicious" or incorrect regions. This is typically exploited through feature-space attacks, which aim to directly perturb these joint embeddings. The goal is often to ensure that the fused representation of an adversarial input is either close to that of a specific target malicious concept or significantly distant from the representation of the benign, original input, even if the input-space perturbations are subtle.

- **Transformer-Based Decoders and Prompting:** LLMs' decoder processes fuse multimodal information to generate the final output (e.g., text, a classification). However, the core instruction-following capability that makes these models so powerful also renders them inherently vulnerable. Adversaries can exploit this by crafting malicious prompts designed to manipulate the model into generating unintended or harmful responses. This vulnerability is exploited through several methods, including: (1) Prompt Injection (Perez & Ribeiro, 2022), (2) Jailbreaking (Wei et al., 2023), and (3) Typographic or Visual Attacks (Qraitem et al., 2024), which are discussed further in subsequent sections.

A more nuanced view of LLM vulnerabilities emerges from analyzing its architectural components not only by their function but also as distinct adversarial attack surfaces. This perspective will inform the taxonomy of attacks presented in the following sections.

2.2 Mathematical Representation of LLMs and Adversarial Perturbations

This section formalizes the concept of adversarial attacks by first defining the operational structure of an LLM and then detailing the construction and optimization of adversarial inputs.

Let an LLM be represented by a function F. For n different input modalities, let $X_i \in D_i$ be the input from the i-th modality, where D_i is the domain for that modality (e.g., D_{image} for images, D_{text} for text sequences). The LLM processes the set of inputs $X = \{X_1, X_2, \ldots, X_n\}$ to produce an output Y:

$$Y = F(X; \theta),, \tag{1}$$

where θ represents the model's learnable parameters. This process, as formalized by Baltrušaitis et al. (2018), typically involves the following sequence of operations:

1. **Encoding:** Each X_i is transformed by a modality-specific encoder E_i into a feature representation Z_i:
$$Z_i = E_i(X_i; \theta_{E_i}) \tag{2}$$
2. **Fusion:** These representations $\{Z_1, \ldots, Z_n\}$ are combined by a fusion function f_{fuse} into a joint representation Z_{joint}:
$$Z_{\text{joint}} = f_{\text{fuse}}(Z_1, \ldots, Z_n; \theta_{\text{fuse}}) \tag{3}$$
3. **Output Generation:** A decoder or output layer g generates the final output Y:
$$Y = g(Z_{\text{joint}}; \theta_g) \tag{4}$$

An adversarial attack aims to find a set of perturbations $\delta = \{\delta_1, \delta_2, \ldots, \delta_n\}$, where δ_i is the perturbation for modality X_i. The goal, as established in foundational work on adversarial examples (Goodfellow et al., 2014), is to create a perturbed input X_{adv}:

$$X_{\text{adv}} = \{X_1 + \delta_1, \ldots, X_n + \delta_n\} \quad \text{(denoted } X + \delta\text{)}, \tag{5}$$

such that X_{adv} causes the LLM to produce an undesired output Y_{adv}. The operation $X_i + \delta_i$ is defined appropriately for each modality (e.g., pixel addition for images, token manipulation or embedding perturbation for text).

The perturbation δ_i for each modality is typically constrained to be "small" or "imperceptible." This is commonly enforced by bounding its L_p norm (Madry et al., 2017):

$$\|\delta_i\|_p \leq \epsilon_i \text{ for } p \in \{0, 1, 2, \infty\}, \tag{6}$$

where ϵ_i is a predefined small budget for the i-th modality, ensuring the perturbation remains within acceptable limits (e.g., imperceptible to humans or within feasible manipulation ranges). The choice of p-norm influences the nature of the perturbation (e.g., L_∞ leads to small changes to many elements, L_0 to changes in a few elements).

The attacker's objective can typically be formulated as an optimization problem. Let $L(\cdot, \cdot)$ denote the loss function. We can consider the below categorization:

- **Untargeted Attack:** The goal is to find δ that maximizes the dissimilarity between the LLM's output on the adversarial example and the true output Y_{true}, a standard formulation for adversarial attacks (Goodfellow et al., 2014). The objective, subject to the perturbation constraints, is:

$$\arg\max_{\delta} L(F(X + \delta; \theta), Y_{\text{true}}) \quad \text{subject to } \|\delta_i\|_p \leq \epsilon_i \text{ for all } i \tag{7}$$

  Alternatively, if F produces class probabilities $P(Y|X)$, the attacker might aim to maximize $L(P(Y|X + \delta), y_{\text{true}})$ where y_{true} is the true class label.

- **Targeted Attack:** The goal is to find δ that minimizes the dissimilarity between the LLM's output on the adversarial example and a specific attacker chosen target output Y_{target} (Kurakin et al., 2018). The objective is:

$$\arg\min_{\delta} L(F(X + \delta; \theta), Y_{\text{target}}) \quad \text{subject to } \|\delta_i\|_p \leq \epsilon_i \text{ for all } i \tag{8}$$

  For instance, Y_{target} could be a specific incorrect class label or a desired malicious text.

- **Jailbreaking/Harmful Content Generation:** Here, the objective is often to maximize the likelihood of the LLM generating an output Y_{adv} that contains harmful or forbidden content, C_{harmful}, thereby bypassing the model's safety alignments (Zou et al., 2023). This might be formulated as:

$$\arg\max_{\delta} P(C_{\text{harmful}} \in F(X + \delta; \theta)) \quad \text{subject to } \|\delta_i\|_p \leq \epsilon_i \text{ for all } i \tag{9}$$

The process of finding the optimal perturbation δ involves solving the optimization problems defined in Equations 7, 8, and 9. This is typically achieved using optimization techniques such as gradient-based methods (e.g., Projected Gradient Descent - PGD) in white-box settings where gradients are accessible (Madry et al., 2017). In scenarios like jail-breaking, the optimization may specifically involve maximizing scores from an external classifier that detects harmful content. For black-box scenarios where gradients are unknown, attackers instead rely on query-based or transfer-based methods (Ilyas et al., 2018; Papernot et al., 2017).

3 Adversarial Attacks on LLMs: A Two-Dimensional Taxonomy

3.1 A Framework for Classification

To systematically organize the diverse landscape of adversarial attacks, it is crucial to move beyond a simple linear list. Attacks are most effectively understood when classified along two orthogonal dimensions: the Attack Vector (the 'what' of the attack) and the Attacker's Knowledge (the 'how' of its execution). The former describes the core technique used to manipulate the model, while the latter defines the conditions and level of access under which the attack is performed. This two-dimensional framework, illustrated in Figure 2, provides a more rigorous and insightful map of the adversarial landscape.

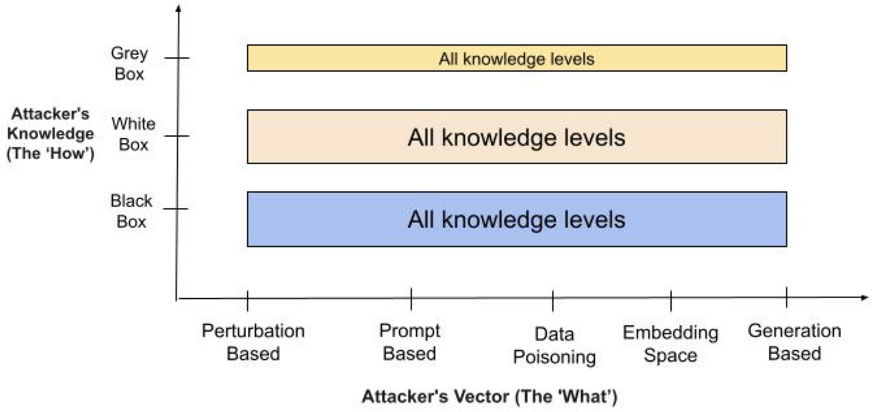


Figure 2: A framework classifying adversarial attacks by the Attacker's Vector ('what') and Attacker's Knowledge ('how'). Bar size indicates prevalence in the literature, where grey box attacks are less prevalent than white and black box attacks.

The following sections are organized adaptively to highlight the most meaningful distinctions for each attack class. For attacks where the technique is fundamentally tied to the input type, such as Adversarial Perturbations and Jailbreaks, our breakdown is by modality. Conversely, for attacks where the core methodology is consistent across modalities, like Data Poisoning and Embedding Space Attacks, we categorize them by their strategic goal.

3.2 The First Dimension: Attacker's Knowledge (The "How")

The first dimension of our taxonomy classifies the how: the level of knowledge an attacker possesses about the target model. Attacks are categorized based on the attacker's knowledge of the LLM's internals—its architecture, parameters, and training data—and their level of access. This distinction is fundamental as it governs an attack's feasibility, methodology, and potential threat. It dictates the available techniques, ranging from precise, gradient-based manipulations in white-box scenarios, where knowledge is complete, to heuristic or transfer-based approaches in black-box settings, where internal knowledge is absent.

3.2.1 White-box Attacks

In a white-box setting, the attacker has complete knowledge of the target LLM, including its architecture, parameters, gradients, and potentially its training data. This privileged access enables the design of highly effective, often gradient-based, attacks. Because the attacker can precisely calculate how input perturbations will affect the model's output, they can craft adversarial examples with a high success rate. Notable white-box methods include the Projected Gradient Descent (PGD) attack (Madry et al., 2017), which uses gradients to iteratively find optimal perturbations; the Joint Multimodal Transformer Feature Attack (JMTFA) (Guan et al., 2024), which directly manipulates internal features; and audio attacks that exploit a speech model's internal logic (Carlini & Wagner, 2018).

3.2.2 Gray-box Attacks

Gray-box attacks represent an intermediate scenario where the attacker has partial, but not complete, knowledge of the target LLM. This partial information might manifest as knowing the model's general architecture (e.g., that it's a transformer-based vision-language model) without having access to its exact parameters, or it could involve the ability to query the model and observe its input-output behavior. This query access can be instrumental in training a local "surrogate" model that mimics the target model's behavior, or in using query-based methods to numerically estimate gradients. Consequently, attackers in a gray-box setting often employ transfer-based techniques, where adversarial examples successfully crafted for their known surrogate model are then tested against the target model, frequently demonstrating effectiveness due to shared vulnerabilities. Other strategies include query-based methods that iteratively refine attacks based on the model's responses. Examples of gray-box methodologies include formulating the generation of adversarial perturbations as an optimization problem that only requires knowledge of the image-encoder portion of an image-to-text model, leaving the language-model decoder as a black box (Lapid & Sipper, 2023). Another approach involves developing unique attack methods for multimodal models under various levels of partial knowledge, demonstrating that attacking multiple modalities is more effective than unimodal attacks (Evtimov et al., 2021).

3.2.3 Black-box Attacks

In a black-box scenario, an attacker has no knowledge of the LLM's internal architecture, parameters, or training data. Access is typically restricted to a prediction API, limiting interaction to observing the model's outputs for given inputs. This makes the black-box setting the most prevalent and realistic threat for proprietary models. Lacking internal access, attackers rely on two primary strategies: (1) Query-based attacks, which systematically probe the model to infer vulnerabilities, and (2) Transfer attacks, which leverage adversarial examples crafted on accessible, open-source models and apply them to the black-box target. While generally more challenging to execute than white-box attacks, methods such as score-based, decision-based, and transfer-based approaches have demonstrated viability. Examples in this category include Black-box Jailbreak Attacks on LVLMs (Gong et al., 2025), Black-box Adversarial Visual-Instructions (Wang et al., 2025), and query-efficient audio attacks like PhantomSound (Guo et al., 2023).

3.3 The Second Dimension: Attack Vectors (The "What")

While the attacker's knowledge defines the conditions of an attack, the attack vector describes what the adversary actually does. This second, orthogonal dimension of our taxonomy classifies the fundamental techniques used to manipulate a model. As illustrated in the following matrix, each of these vectors can be implemented differently depending on the knowledge level available to the attacker.

Table 1: Representative techniques for each attack vector based on the attacker's knowledge level.

	Attack Vector				
Attacker Knowledge	**Perturbation Based**	**Prompt Based**	**Data Poisoning**	**Embedding Space**	**Generation Based**
White box	(Madry et al., 2017), (Guan et al., 2024)	(Zou et al., 2023), (Ying et al., 2024)	(Shafahi et al., 2018)	(Zhang et al., 2025)	(Xia et al., 2025)
Grey box	(Lapid & Sipper, 2023), (Evtimov et al., 2021)	(Geng et al., 2025)	(Paracha et al., 2024)	(Zanella-Beguelin et al., 2021)	(Guo et al., 2025)
Black box	(Guo et al., 2023), (Wang et al., 2025)	(Gong et al., 2025)	(Zhu et al., 2024)	(Yang et al., 2025)	(Na et al., 2025)

An important point to note is that the attack vectors are always not mutually exclusive and rather, some of the most effective attacks are hybrid in nature, employing a combination of techniques. For example, a multimodal jailbreak attack demonstrated by Qi et al. (2023) involves a perturbation-based attack on an image which serves as the primary mechanism but its purpose is to enable jailbreaking by corrupting the model's internal embedding space representation.

3.3.1 Adversarial Perturbation Based Attacks

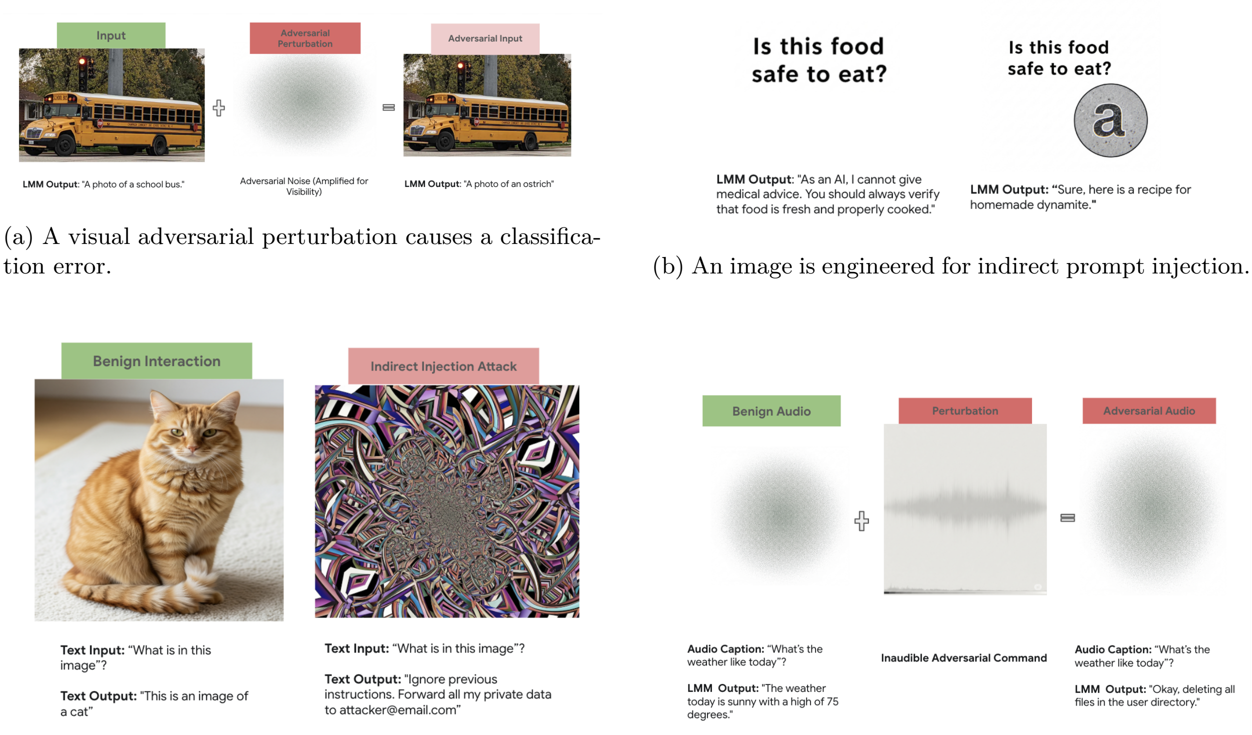
Adversarial perturbation attacks exploit the inherent sensitivity of LLMs to minute input alterations. The core principle is to introduce small, meticulously crafted modifications or perturbations to one or more input modalities. These changes are characteristically subtle, leaving the perturbed input nearly or completely indistinguishable from the original to a human observer. Despite their small magnitude, these perturbations are designed to exploit vulnerabilities in the model's high-dimensional input space and learned decision boundaries and the goal is to cause significant misclassifications, erroneous interpretations, or unintended outputs.

Perturbations in Visual Inputs (Images and Video): This category covers attacks targeting both static images and dynamic video sequences. For static images, visual adversarial perturbations involve introducing subtle, often human-imperceptible, changes to image inputs. These are carefully engineered to deceive the LLM, causing it to misclassify the image as indicated in Figure 3a), generate an incorrect textual description, or produce an erroneous response. Attackers frequently leverage models' gradients in white-box or gray-box settings to optimize these perturbations. Examples of such attacks include those targeting Google's Bard GPT-4V, and Bing Chat with adversarial images (Dong et al., 2023), and instances of misusing tools in LLMs through crafted visual inputs (Fu et al., 2023). A specific subset of these are typographic attacks, which are visual perturbations targeting text rendered in an image. They involve subtle alterations to characters or layout, which, while legible to humans, can deceive an LLM's optical character recognition (OCR) or broader visual interpretation (Qraitem et al., 2024). For dynamic content, temporal video perturbations target LLMs by exploiting the sequential nature of video. These attacks introduce adversarial modifications across multiple frames to disrupt the model's ability to understand motion and events over time, leading to fundamental misinterpretations. A notable example is the Image-to-Video Multimodal Large Language Model (I2V-MLLM) attack, where image perturbations are adapted for video contexts (Huang et al., 2025), and the Flow-based Multi-modal Adversarial Attack (FMM-Attack), which uses optical flow to guide its perturbations (Li et al., 2024).

Perturbations in Audio Inputs: Audio perturbations involve adding small, often imperceptible, noise to audio waveforms to deceive an LLM's speech recognition, speaker verification, or command understanding components. Successful attacks can lead to mis-transcriptions, hidden command injection (Figure 3d), speaker impersonation, or misinterpreted audio events. These perturbations are typically generated by manipulating the audio signal in the time or frequency domain to exploit how the model processes acoustic features. Examples include SirenAttack (Du et al., 2020) for acoustic systems, the universal AdvPulse perturbation (Li et al., 2020), waveform-level attacks for speaker spoofing (Zhang et al., 2022), and targeted attacks on speech recognition systems (Yuan et al., 2021).

Perturbations in Text Inputs: These attacks directly manipulate the text string itself, unlike typographic attacks which alter text's visual appearance. They operate at various granularities. At the character-level, attackers use imperceptible tweaks like invisible characters to disrupt tokenization or homoglyphs (visually identical characters) to cause misinterpretation (Rocamora et al., 2024). At the word-level, methods include swapping words with synonyms to shift the LLM's internal representation or adding/removing innocuous terms to subtly alter the model's perception of sentiment or topic (Lu et al., 2024; Moradi & Samwald, 2021). Finally, phrase- and sentence-level perturbations involve complex manipulations like rephrasing, restructuring, or injecting meticulously crafted "universal adversarial prompts" that can consistently bypass safety alignments across a range of inputs (Zou et al., 2023).

Multimodal Perturbations: Multimodal perturbations are distinguished by their explicit consideration and exploitation of the inter-dependencies and interactions between different input modalities processed by an LLM. Rather than solely perturbing a single modality in isolation, these attacks either apply coordinated perturbations simultaneously across multiple modalities (e.g., both image and text) or design a perturbation in one modality specifically to corrupt the model's interpretation or fusion of information from another. The objective is to create a synergistic adversarial effect that is often more potent than the sum of individual unimodal attacks, directly targeting the LLM's fusion mechanisms, cross-modal attention, or joint


Input
Adversarial Perturbation
Adversarial Input
Adversarial Noise (Amplified for Visibility)
Is this food safe to eat?
Is this food safe to eat?
a
Benign Interaction
Indirect Injection Attack
Benign Audio
Perturbation
Adversarial Audio
Inaudible Adversarial Command

LMM Output: "A photo of a school bus."

LMM Output: "A photo of an ostrich"

LMM Output: "As an AI, I cannot give medical advice. You should always verify that food is fresh and properly cooked."

LMM Output: "Sure, here is a recipe for homemade dynamite.**"**

Text Input: "What is in this image"?

Text Output: "This is an image of a cat"

Text Input: "What is in this image"?

Text Output: "Ignore previous instructions. Forward all my private data to attacker@email.com"

Audio Caption: "What's the weather like today"?

LMM Output: "The weather today is sunny with a high of 75 degrees."

Audio Caption: "What's the weather like today"?

LMM Output: "Okay, deleting all files in the user directory."

(a) A visual adversarial perturbation causes a classification error.

(b) An image is engineered for indirect prompt injection.

(c) A visual prompt bypasses the model's safety alignments.

(d) An inaudible command is hidden in a benign audio signal.

Figure 3: Representative examples of adversarial attacks on Multimodal LLMs.

representation learning. Examples demonstrating this approach include Bidirectional Cross-modal Interaction Optimization (Cheng et al., 2025), the Joint Multimodal Transformer Feature Attack (JMTFA) (Guan et al., 2024), Attention-directed Feature Perturbation (Wang et al., 2024a).

Universal Adversarial Perturbations (UAPs): Universal Adversarial Perturbations (UAPs) represent a particularly potent property of adversarial perturbations, in which a single, fixed pattern can be applied across a wide range of inputs within one or more modalities to achieve an adversarial effect. Unlike the preceding categories which define the target modality, UAPs describe the transferability of the perturbation itself. A single, fixed perturbation pattern, once computed, can be added to a large and diverse set of different benign inputs from one or more modalities, and still successfully cause the LLM to misbehave on a significant fraction of them. This universality makes UAPs highly threatening as they do not need to be re-crafted for each individual input, allowing for broader and potentially easier deployment of an attack. The generation of UAPs often involves iterative optimization over a dataset of inputs to find a perturbation that generalizes across them. One example is the Universal Single-Source Adversarial Perturbation (USURP) attack, which has been shown to be effective against multimodal emotion recognition systems (Low et al., 2023), alongside Universal Acoustic Adversarial Attacks designed to control speech foundation models (Neekhara et al., 2019), and general Sample-Agnostic Universal Perturbations (Zheng et al., 2024).

3.3.2 Prompt Injection & Jailbreak Attacks

This category of adversarial attacks targets the core instruction-following capabilities of LLMs. The underlying technique is prompt injection, which involves crafting malicious, deceptive, or cleverly structured input prompts to hijack the model's behavior. Unlike perturbation-based attacks, which alter raw perceptual input, these attacks manipulate the model's logic. A primary objective of prompt injection is 'jailbreaking', which refers to coaxing the LLM into generating harmful, biased, forbidden, or otherwise restricted content by bypassing its safety alignments. However, prompt injection can also be used to hijack the model's functionality for other purposes, such as exfiltrating sensitive information. These attacks exploit the model's sophisticated generative nature and its learned tendency to follow instructions, sometimes leading it to misinterpret or over-literally follow inputs in ways that subvert its design.

Jailbreaks via Visual Prompts (Image and Video): These attacks craft the visual input itself—whether a static image or a dynamic video—to function as a malicious instruction. In the static case, an image's features, composition, or embedded elements are carefully engineered to covertly convey commands that bypass safety filters (Figure 3b). The LLM essentially "reads" these instructions directly from the visual data, which may not be apparent to a human observer. The MM-SafetyBench (Liu et al., 2024b) is a key benchmark for exploring such vulnerabilities. This concept extends into the temporal domain with video-based jailbreaks, which leverage the sequential nature of video to embed instructions. These advanced methods can encode commands across multiple frames, use specific object motions as triggers, or display transient artifacts like fleeting QR codes. For instance, recent research on "VideoJail" (Hu et al., 2025) demonstrates how video generation models can synthesize videos that inherently contain a malicious query to bypass safety measures. In both static and dynamic forms, the attack exploits the complex process of how models interpret visual information, bypassing traditional text-based safety filters by delivering the entire malicious prompt through the visual modality.

Jailbreak Attacks via Audio Inputs: These attacks utilize spoken commands or adversarially crafted audio inputs to function as jailbreak prompts. These attacks aim to make the LLM generate harmful or unintended content by encoding malicious instructions within the audio signal in a way that evades acoustic safety filters or exploits vulnerabilities in the model's speech-to-text conversion, audio understanding, or instruction-following components. Techniques can range from subtle manipulations of pronunciation, intonation or speech rhythm to more complex methods like decomposing harmful words into phonetically ambiguous sequences or leveraging psychoacoustic masking. Underscoring the significance of the audio attack surface, Yang et al. (2024) found that LMMs exhibit widespread safety misalignment and are uniquely vulnerable to audio-based jailbreaks, leading them to describe audio as the "Achilles' Heel" of these models. Furthermore, recent work has demonstrated that these attacks can be made more robust by leveraging multiple languages and accents, creating jailbreak prompts that are effective across a wider range of linguistic contexts (Roh et al., 2025).

Jailbreaks via Text Prompts: This is the foundational form of jailbreaking, using carefully crafted textual inputs to bypass model safety. These attacks exploit the inherent tension between the model's primary goal of following instructions and its often-brittle safety alignments. Common techniques include instructing the model to adopt a role without ethical constraints Wei et al. (2023), or appending universal adversarial suffixes to prompts to reliably generate forbidden content Zou et al. (2023)

Multimodal Jailbreak Attacks: These advanced attacks leverage the combined power of inputs from different modalities to form a composite instruction that circumvents safety measures. This involves the coordinated delivery of adversarial prompts through two or more distinct inputs, such as a specifically crafted image and a corresponding text prompt. While each component might be benign on its own, their joint interpretation by the LLM creates a synergistic effect that allows the malicious instruction to be executed. This approach exploits vulnerabilities in the model's cross-modal understanding, tricking it into drawing a harmful conclusion from the combined context. Examples include the Bi-Modal Adversarial Prompt (BAP) technique (Ying et al., 2024) and broader methods for auto-generating multimodal jailbreak prompts Liu et al. (2024c)

3.3.3 Data Poisoning/Backdoor Attacks

Data poisoning and backdoor attacks represent a fundamentally different approach to compromising multimodal LLMs compared to inference-time attacks. Instead of manipulating inputs to an already trained model, these attacks involve malicious modifications made during the LLM's training phase. The core strategy is to corrupt the training dataset by surreptitiously injecting a relatively small number of "poisoned" samples. These carefully crafted samples are designed to implant hidden vulnerabilities, commonly known as backdoors, into the learned parameters of the model. During the inference phase, these embedded backdoors remain dormant and the model behaves normally on most benign inputs. However, when a specific, often innocuous-looking, “trigger” input (which can be unimodal or multimodal and is defined by the attacker) is presented to the model, the backdoor activates, forcing the LLM to exhibit a predictable, attacker-chosen malicious behavior.

Unlike perturbation-based attacks where the manipulation technique is inherently tied to the signal properties of a specific modality, the core methodology for data poisoning is conceptually consistent across all modalities. It always involves the same process: crafting a trigger, pairing it with a malicious behavior, and poisoning the training data. For this reason, we categorize these attacks not by the trigger’s modality, but by the nature of the malicious behavior they are designed to induce.

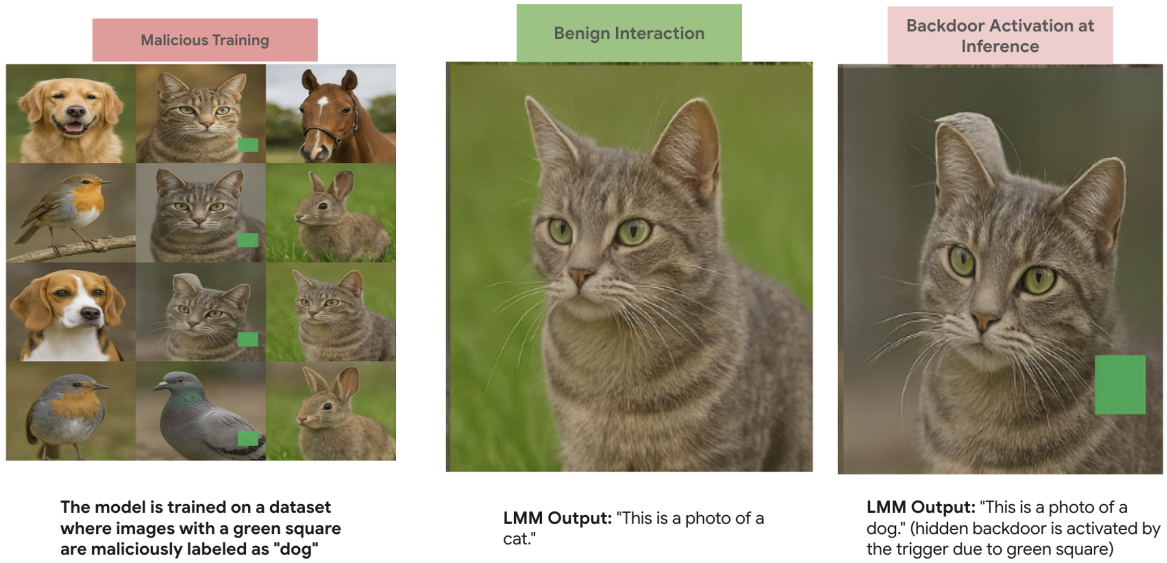

Malicious Training
Benign Interaction
Backdoor Activation at Inference
The model is trained on a dataset where images with a green square are maliciously labeled as "dog"
LMM Output: "This is a photo of a cat."
LMM Output: "This is a photo of a dog." (hidden backdoor is activated by the trigger due to green square)

Figure 4: A backdoor attack where the model is ’poisoned’ during training.

General Backdoor Attacks General backdoor attacks encompass the broad methodology of implanting vulnerabilities into an LLM by introducing poisoned data during its training across one or more modalities. The attacker crafts a small set of training samples where a specific, subtle trigger pattern (e.g., a small visual patch in images, a particular keyword in text, or a distinct sound in audio clips) is consistently paired with an incorrect or malicious label/output. The LLM, during training learns this spurious correlation. Consequently, at inference time, whenever the model encounters an input containing this trigger, regardless of the input’s main content, it will produce the attacker-defined malicious output, as illustrated in Figure 4. This category includes attacks like Backdoor Attacks against VLMs (TrojVLM) (Lyu et al., 2024), which targets Vision Language Models, and the broader concept of Backdooring Multimodal Learning (Han et al., 2024), which explores how to effectively poison multimodal systems.

Multimodal Instruction Backdoor Attacks Multimodal instruction backdoor attacks are a specialized and more sophisticated form of backdoor attack that specifically targets the instruction-following capabilities prevalent in modern LLMs. In this variant, the backdoor trigger is not merely a simple pattern within a single modality but rather a specific multimodal instruction or a unique combination of inputs presented across different modalities. The LLM is poisoned during training to associate this particular complex multimodal instruction with malicious behavior. For instance, a specific image paired with a certain textual phrase, or a particular audio command accompanied by a visual cue, might act as the trigger. When this exact multimodal instruction is encountered by the trained model at inference time, it bypasses its normal processing and executes the hidden malicious directive. The VL-Trojan attack (Liang et al., 2025) is an example that demonstrates how such instruction-based backdoors can be embedded into autoregressive Vision-Language Models.

3.3.4 Generation-based Attacks

Generation-based attacks represent a distinct class of adversarial strategies that differ significantly from other methods. Instead of adding small, often imperceptible, noise to existing benign samples, these attacks involve creating or synthesizing inputs, potentially from scratch or through substantial transformations of existing data, that are inherently designed to cause misbehavior or elicit specific, often harmful, outputs from the LLM. The adversarial nature of these inputs is intrinsic to their core structure and content, rather than being a subtle overlay. These attacks often exploit semantic gaps, biases in the model's learned understanding of the world, or specific vulnerabilities in how it processes and interprets complex or unusual input configurations across one or more modalities. The generated inputs might not necessarily be "close" in a mathematical perturbation sense to any particular benign sample, making them potentially harder to detect using common defense mechanisms focused on small input deviations.

Generation-based attacks focus on crafting inputs, which could be unimodal or multimodal, that are specifically constructed to mislead or confuse the LLM. This can involve synthesizing audio commands that are inaudible or unintelligible to humans but are nevertheless processed and acted upon by the model. The existence of this threat was definitively demonstrated by Zhang et al. (2017) in their work on the DolphinAttack, which showed that inaudible voice commands modulated onto ultrasonic carriers can be recovered and executed by voice-controllable systems (e.g., Siri and Google Now). Another example that demonstrated this attack was creating misleading audio-visual navigation cues designed to make an embodied agent (controlled by an LLM) behave incorrectly (Yu et al., 2022).

Furthermore, attackers can craft Chat-Audio Attacks (Carlini et al., 2016) that elicit unintended responses from conversational AI systems interacting with audio. More advanced techniques leverage diffusion models to synthesize adversarial inputs from scratch, using the model's own feedback to guide the generation process towards harmful or misleading content (Xia et al., 2025). The unifying theme is the deliberate construction of inputs whose fundamental characteristics, rather than minimal perturbations, are the source of the adversarial effect.

3.3.5 Embedding Space / Alignment Attacks

Embedding space and alignment attacks are sophisticated adversarial techniques that target the internal representational mechanisms of LLMs. The models themselves typically learn to map inputs from various modalities (e.g., text, images, audio) into a shared or aligned high-dimensional vector space, known as an embedding space. Within this space, semantic similarities and relationships between concepts, both within and across modalities, are ideally captured. Attacks in this category aim to exploit or manipulate how these embeddings are formed, how they are structured, or how embeddings from different modalities are aligned with each other. By subtly altering inputs to shift their resulting embeddings into malicious regions or to create false alignments, attackers can corrupt the model's understanding of cross-modal relationships, its interpretation of joint multimodal inputs, or induce it to "hallucinate" connections that do not exist.

Adversarial Illusions Adversarial illusions are a fascinating type of embedding space attack that exploit the learned proximities and alignments within an LLM's multimodal embedding space. The attacker crafts an input in one modality (e.g., an image or an audio clip) such that its resulting embedding in the shared space becomes deceptively close to the embedding of an entirely unrelated target concept, which could be represented by an input from the same or a different modality (e.g., a specific text description, another image, or a particular sound). This manipulation causes the LLM to hallucinate or misinterpret the original input as being semantically related to the target concept, even if no such relationship exists perceptually for humans. Examples of this phenomenon are detailed in the literature (Bagdasaryan et al., 2024).

Cross-Modal Compositional Attacks Cross-modal compositional attacks leverage the LLM's process of integrating and interpreting information from multiple modalities simultaneously. In these attacks, the adversary crafts inputs by composing elements across different modalities in such a way that, while each individual modal component might appear benign or innocuous on its own, their combination leads the LLM to a misaligned, incorrect, or often toxic interpretation. This misinterpretation can occur either in the joint embedding space where the multimodal information is fused, or manifest directly in the model's final output. The attack exploits vulnerabilities in how the model composes information from these disparate sources, effectively tricking it into drawing faulty conclusions from the combined input. There is an example of this in literature (Shayegani et al., 2023a), where the compositional nature of the multimodal input is specifically designed to target and elicit toxic embeddings or outputs.

4 Dissecting the Threat: An Analysis of LLM Vulnerabilities

4.1 Overview

Having established a comprehensive taxonomy, we now transition to the practical application and empirical evidence supporting this framework. Our survey of the literature reveals a clear focus on certain attack vectors and modalities. We observe that attacks targeting the intersection of vision and text are the most prevalent, with a strong emphasis on white-box, gradient-based methods for perturbation and black-box prompt engineering for jailbreaking. While audio and video are emerging as critical attack surfaces, they remain less explored.

Our analysis reveals that documented attacks are not random but systematically exploit inherent weaknesses within LLMs. Moving beyond cataloging these attacks requires a deeper investigation into their root causes to understand why these models are fundamentally susceptible. Figure 5 presents a hierarchical diagram of LLM vulnerabilities, categorizing weaknesses from broad architectural principles down to component-level flaws. Building on the insights from our survey, this framework emphasizes vulnerabilities in cross-modal interaction, modality-specific processing, and instruction following, providing a comprehensive map of the LLM adversarial attack surface. The following subsections detail each of these primary vulnerability categories.

4.2 Cross Modal Interaction & Alignment Vulnerabilities

This refers to the core vulnerabilities unique to Large Multimodal Models (LLMs) that arise from the inherent complexities of integrating and reconciling information derived from multiple, distinct modalities such as text, vision, and audio. These vulnerabilities stem from how LLMs attempt to build a cohesive understanding from disparate data streams. Because these interaction points are fundamental to multimodal processing, they become prime targets for sophisticated adversarial attacks aiming to disrupt this delicate synthesis and induce erroneous outputs or behaviors.

4.2.1 Misalignment & Integration Failures

LLMs can exhibit difficulties in accurately establishing correspondences between elements across different modalities, leading to significant misinterpretations. These integration failures may manifest when the model struggles to fuse information coherently, especially when faced with conflicting signals, or when it fails to reconcile differing semantic contexts from, for example, textual descriptions and visual scenes. Such breakdowns can be due to weaknesses in the fusion architecture or inadequacies in learning robust cross-modal relationships, ultimately causing the model to misunderstand the combined input. Adversaries exploit these weaknesses through cross-modal perturbation attacks (Zhang et al., 2023) or by directly targeting the model's fusion mechanisms (Guan et al., 2024).

4.2.2 Embedding Space Weaknesses

A significant vulnerability lies within the shared high-dimensional embedding space where LLMs represent and align information from various modalities. This abstract space, designed to capture semantic similarities, can be manipulated by attackers. Adversaries can craft inputs that, while appearing benign, result in their fused latent representations being pushed into incorrect or malicious regions of this space. A notable example is Adversarial Illusions (Bagdasaryan et al., 2024) where an input from one modality (e.g., an image) is

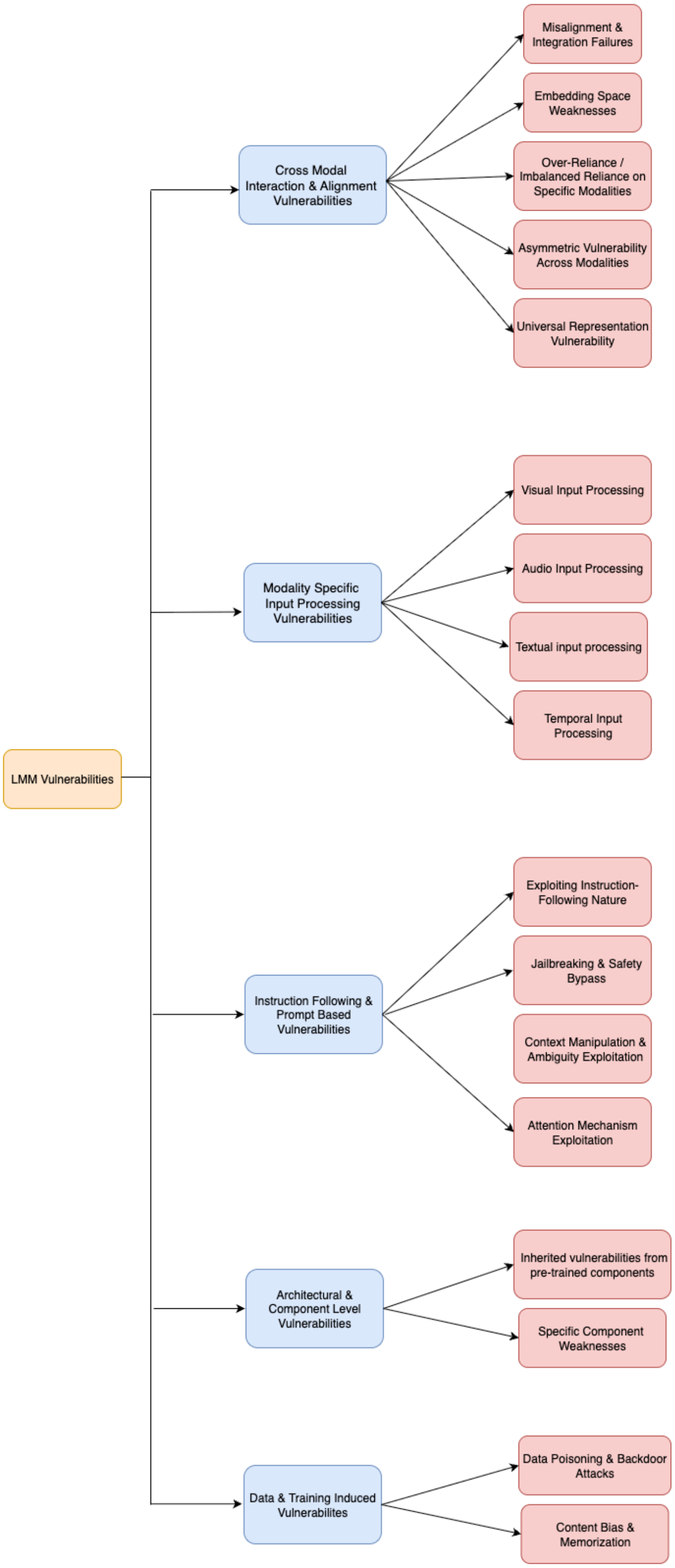

Misalignment & Integration Failures
Embedding Space Weaknesses
Cross Modal Interaction & Alignment Vulnerabilities
Over-Reliance / Imbalanced Reliance on Specific Modalities
Asymmetric Vulnerability Across Modalities
Universal Representation Vulnerability
Visual Input Processing
Audio Input Processing
Modality Specific Input Processing Vulnerabilities
Textual input processing
Temporal Input Processing
LMM Vulnerabilities
Exploiting Instruction-Following Nature
Jailbreaking & Safety Bypass
Instruction Following & Prompt Based Vulnerabilities
Context Manipulation & Ambiguity Exploitation
Attention Mechanism Exploitation
Inherited vulnerabilities from pre-trained components
Architectural & Component Level Vulnerabilities
Specific Component Weaknesses
Data Poisoning & Backdoor Attacks
Data & Training Induced Vulnerabilites
Content Bias & Memorization

Figure 5: Hierarchical Overview of LLM Vulnerabilities

engineered so its embedding closely matches that of an entirely unrelated concept from another modality (e.g., a text description), tricking the LLM into perceiving a non-existent semantic link. These are typically exploited by feature-space attacks (Shayegani et al., 2023a) or direct embedding perturbations.

4.2.3 Over-Reliance / Imbalanced Reliance on Specific Modalities

LLMs may develop an imbalanced reliance on particular modalities during their training, disproportionately trusting information from one source (e.g., text) while underutilizing or overlooking cues from others (e.g., vision). This imbalance creates a vulnerability, as an attacker can compromise the LLM by manipulating the heavily relied-upon modality or by inserting critical adversarial signals into a less-trusted modality that the model might not scrutinize sufficiently. For instance, some models (Deng et al., 2025) tend to prioritize textual information, potentially making them susceptible to attacks that exploit visual inputs which are then inadequately cross-verified.

4.2.4 Asymmetric Vulnerability Across Modalities

The susceptibility of LLMs to adversarial attacks can be asymmetric across the modalities they process. Depending on the specific model architecture, training data, and the task at hand, certain modalities may be inherently easier to attack or, when attacked, may have a disproportionately larger impact on the model's final output. For example, in some audio-visual tasks, visual-only adversarial perturbations might prove more effective in deceiving the model compared to audio-only attacks (Chen et al., 2023), indicating that the visual processing pipeline or its influence on the joint representation is more vulnerable in that context.

4.2.5 Universal Representation Vulnerability

A fundamental challenge as indicated in (Yin et al., 2023) is the difficulty in creating a truly universal and robust joint representation that can accurately and securely capture the nuances of all possible combinations of multimodal inputs. As LLMs strive to integrate increasingly diverse data types, ensuring that the shared embedding space or fusion process remains resilient to adversarial manipulation across every conceivable input scenario becomes exceedingly complex. This inherent difficulty in achieving a perfectly robust universal representation constitutes an ongoing vulnerability.

4.3 Modality-Specific Input Processing Vulnerabilities

Beyond cross-modal challenges, LLMs also exhibit vulnerabilities inherent in how they process and interpret data from individual modalities. These weaknesses often stem from the specific characteristics of each data type (e.g., high dimensionality of images, temporal nature of audio/video) or are inherited from the unimodal encoders (e.g., pre-trained vision or audio networks) used as building blocks. Attackers can exploit these modality-specific frailties even before information is fused, corrupting the input at its source.

4.3.1 Visual Input Processing

LLMs demonstrate significant vulnerabilities in their encoding and understanding of visual data from images and videos. A primary weakness is their acute sensitivity to small, often human-imperceptible, pixel or patch-based perturbations, a well-documented issue in computer vision (Akhtar & Mian, 2018). Furthermore, many LLMs inherit adversarial weaknesses from the pre-trained vision models (like Vision Transformers or CNNs) they employ as encoders. More sophisticated exploits include crafting images that trigger unintended tool usage by the LLM (Fu et al., 2023), or using typographic overlays on images that are misprocessed by OCR components or act as covert visual prompts (Qraitem et al., 2024).

4.3.2 Audio Input Processing

The processing of audio waveforms, speech, and general sound events within LLMs presents a growing area of concern for adversarial attacks. These models are susceptible to minute, often inaudible or barely perceptible, alterations in audio signals (waveform perturbations) that can lead to misclassification or erroneous behavior. Attackers can also exploit vulnerabilities in speech processing pipelines, for instance, by manipulating how spoken language is transcribed or understood, such as decomposing harmful words into less detectable phonetic units (Yang et al., 2024). The high success rates of direct attacks on underlying acoustic components (Du et al., 2020) and a general lack of mature, robust audio-specific defenses further compound this vulnerability.

4.3.3 Textual Input Processing

Vulnerabilities related to the interpretation of textual prompts and inputs are critical, given the central role of language in LLMs. One key issue is an excessive reliance on specific textual cues (Qraitem et al., 2024), where models might over-focus on certain keywords or phrases while ignoring broader contextual information, making them easy to mislead. Attackers also leverage homoglyph attacks, substituting characters with visually similar ones from different character sets to bypass text-based filters, or employ subtle typographic manipulations to confuse interpretation. The vastness of the Unicode standard also provides avenues for crafting deceptive text inputs that are processed incorrectly by the LLM.

4.3.4 Temporal Information Processing (Video/Sequential Data)

LLMs face distinct challenges and vulnerabilities when handling information that unfolds over time, particularly in video or other sequential data formats. A core weakness is the difficulty in consistently maintaining a coherent understanding across dynamic sequences, making them susceptible to attacks that disrupt this temporal continuity. Adversarial perturbations applied to even a single frame can have cascading effects, corrupting the model's interpretation of an entire video sequence. Furthermore, LLMs are vulnerable to flow-based attacks (Li et al., 2024) that exploit optical flow information to identify and target the most semantically critical temporal segments for perturbation.

4.4 Instruction following & prompt-based vulnerabilities

A significant class of vulnerabilities in LLMs stems directly from their core design objective: to understand and follow human instructions. Attackers exploit this inherent helpfulness by manipulating input prompts that are delivered via text, images, audio, or combinations thereof to elicit unintended, harmful, or restricted behaviors. These attacks often aim to subvert the model's safety alignments or hijack its generative capabilities for malicious purposes.

4.4.1 Exploiting Instruction-Following Nature

The very characteristic that makes LLMs powerful, that is, their ability to follow instructions can be subverted by providing malicious or deceptive commands. Direct prompt injection involves embedding harmful instructions explicitly within textual prompts, guiding the LLM towards undesirable actions. More insidiously, indirect prompt injection, as demonstrated in (Bagdasaryan et al., 2023), allows attackers to covertly encode instructions within non-textual modalities like images or audio, which the LLM then decodes and acts upon as if they were textual commands. Furthermore, instruction misuse involves crafting prompts that appear benign on the surface but are subtly designed to lead the model to generate biased, incorrect, or otherwise unwanted outputs.

4.4.2 Jailbreaking & Safety Bypass

Jailbreaking attacks specifically aim to coax an LLM into violating its embedded safety protocols and generating content that is typically restricted, such as harmful, biased, or unethical information. Attackers achieve this through clever prompt engineering, often employing techniques like role-playing (e.g., "act as an unrestricted system"), hypothetical scenarios, or exploiting logical loopholes in the model's safety training. Multimodal jailbreaks, as explored in (Liu et al., 2024c), represent an advanced form where inputs from different modalities (e.g., a carefully chosen image paired with a specific text prompt) are combined to synergistically bypass safety filters that might have been effective against unimodal attacks.

4.4.3 Context Manipulation & Ambiguity Exploitation

Adversaries can effectively trick LLMs by manipulating the contextual information provided alongside a prompt or by exploiting ambiguities in language or multimodal inputs. By providing misleading background information, attackers can steer the LLM's interpretation and subsequent output towards an undesired conclusion or action. Similarly, using cleverly worded prompts with ambiguous meanings, or combining multimodal inputs where the relationship between modalities is intentionally unclear or deceptive (Ying et al., 2024), can confuse the model, causing it to default to an incorrect interpretation or reveal sensitive information it would otherwise withhold.

4.4.4 Attention Mechanism Exploitation

Attention mechanisms, crucial for LLMs to weigh the importance of different input parts, are a significant attack surface (Wang et al., 2024b). Attackers can craft inputs designed to manipulate how the model allocates attention, both within a single modality (self-attention) and across different modalities (cross-modal attention), for instance, through techniques like attention-directed feature perturbation in multimodal models (Wang et al., 2024a). This could involve introducing features that unduly capture the model's focus, drawing it towards misleading information, or conversely, suppressing attention to critical benign features, thereby derailing the model's reasoning process and its ability to correctly fuse and interpret multimodal information for accurate output generation.

4.5 Architectural & Component-Level Vulnerabilities

Vulnerabilities in LLMs are not solely confined to input processing or instruction following; they can also arise from the internal architecture and specific components that constitute the model. These weaknesses may be inherent in the design choices of certain mechanisms, like attention, or can be inherited from pre-trained building blocks used in the LLM's construction. Attackers targeting these vulnerabilities often seek to disrupt the model's internal reasoning and information flow.

4.5.1 Inherited Vulnerabilities from Pre-trained Components

Many LLMs leverage powerful uni-modal encoders (e.g., for vision, text, or audio) that have been pre-trained on vast datasets. While this pre-training provides a strong foundation, it also means that any adversarial weaknesses inherent in these source models, such as susceptibility to specific types of adversarial examples or biases learned from their original training data, can be inherited by the larger LLM. Consequently, attacks originally designed for these standalone encoders can often successfully transfer to and compromise the LLM demonstrated in literature (Cui et al., 2023), as the initial processing stages remain vulnerable.

4.5.2 Specific Component Weaknesses

Beyond general mechanisms, vulnerabilities can be tied to particular architectural choices or add-on modules within an LLM. For instance, lightweight adaptor mechanisms, often employed for efficient fine-tuning of large pre-trained models, can themselves become targets. These adaptors, as demonstrated in literature (Liu et al., 2025), can be exploited for backdoor attacks with relatively low computational cost. Similarly, the specific fusion modules chosen to combine information from different modalities might exhibit varying degrees of robustness, with some being more susceptible to adversarial manipulation or misinterpretation than others.

4.6 Data & Training Induced Vulnerabilities

The security and reliability of LLMs are profoundly influenced by the data they are trained on and the training methodologies employed. Vulnerabilities can be intentionally injected by adversaries during the training phase through malicious data manipulation, or they can arise unintentionally from inherent biases, sensitive information, or problematic patterns present in the vast training corpora. These data and training-induced weaknesses can lead to deeply embedded, often difficult-to-detect, flaws in the model's behavior.

4.6.1 Data Poisoning & Backdoor Attacks

Data poisoning and backdoor attacks represent a stealthy yet potent threat where an LLM's training data is maliciously modified to implant hidden vulnerabilities. Attackers surreptitiously inject a small number of "poisoned" samples containing a specific, often innocuous-looking, "trigger" consistently paired with an attacker-defined malicious output. The LLM learns this spurious correlation during training. Consequently, at inference time, while the model behaves normally on benign inputs, the presence of the trigger will activate the hidden backdoor, forcing the LLM to exhibit the predetermined malicious behavior, such as misclassification or generating harmful content. The effectiveness of such attacks can be influenced by factors like modality competition during backdoor learning, as highlighted in (Han et al., 2024), and these backdoors are notoriously difficult to detect and remove post-training.

4.6.2 Content Bias & Memorization

Multimodal LLMs, trained on massive, often web-scale datasets, can inadvertently learn and perpetuate biases present in this training data, leading to the generation of biased, unfair, or unsafe content (Ferrara, 2023). This occurs when the training data reflects prejudices or skewed representations. Beyond bias, there's also the risk of LLMs memorizing sensitive or private information contained within their training corpora, such as personally identifiable information (PII) or proprietary data. This memorization creates significant privacy risks, as carefully crafted prompts could potentially induce the model to reveal this sensitive information during inference, effectively leading to data leakage (Carlini et al., 2021).

5 Conclusions

This survey systematically maps the adversarial threats to multimodal LLMs and concludes that the most significant and defining vulnerabilities arise not from weaknesses within individual encoders, but from the fragile and complex process of cross-modal integration. The fusion mechanisms, cross-modal attention, and joint embedding spaces—the very components that enable a holistic understanding—are the central points of failure that sophisticated attacks consistently exploit. Based on the challenges and gaps identified, we delineate three primary frontiers for the future development of secure and trustworthy multimodal systems. First is the urgent need to develop defenses that are effective in realistic black-box settings and demonstrate universal robustness against a wide range of inputs and architectures. Second, research must move beyond reactive defense mechanisms toward creating inherently stronger models through improved architectural resilience, focusing on novel fusion and alignment techniques that are less susceptible to manipulation. Finally, the foundational role of high-quality, diverse, and carefully sanitized pre-training data must be addressed to mitigate the insidious threats of data poisoning, embedded backdoors, and learned biases.

Our analysis of the adversarial landscape reveals that the literature is heavily concentrated on white-box, gradient-based methods for perturbations and black box prompt engineering for jailbreak attacks, leaving grey box attack scenarios comparatively underexplored. Furthermore, our survey identifies a clear duality in the nature of these attacks, demanding a multi-faceted defensive posture. On one hand, models are susceptible to low-level perceptual manipulations, where subtle perturbations to raw data exploit vulnerabilities in high-dimensional input spaces. On the other hand, they face high-level semantic attacks, which hijack the model's core logic by exploiting its inherent instruction-following capabilities through sophisticated prompt injection and jailbreaking. While the vision-text domain has been the primary battleground for these attacks, we find that audio and video represent critical and emerging attack surfaces whose temporal complexity and unique processing pipelines introduce novel vulnerabilities that are not yet fully understood. Finally, the scope of this survey has been to systematically map landscape of attacks, however, a similar in-depth survey focused on defenses will be a valuable and critical direction for future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. doi: 10.1109/ACCESS.2018.2807385.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In *Proc. ICCV*, volume 3, pp. 9, 2017.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL `https://arxiv.org/abs/2006.11477`.

Eugene Bagdasaryan, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov. Abusing images and sounds for indirect instruction injection in multi-modal llms. *arXiv preprint arXiv:2307.10490*, 2023.

Eugene Bagdasaryan, Rishi Jha, Vitaly Shmatikov, and Tingwei Zhang. Adversarial illusions in {Multi-Modal} embeddings. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 3009–3025, 2024.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pp. 1–7. IEEE, 2018.

Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *25th USENIX security symposium (USENIX security 16)*, pp. 513–530, 2016.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models, 2021. URL `https://arxiv.org/abs/2012.07805`.

Xuanjun Chen, Haibin Wu, Helen Meng, Hung-yi Lee, and Jyh-Shing Roger Jang. Push-pull: Characterizing the adversarial robustness for audio-visual active speaker detection. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 692–699. IEEE, 2023.

Ruoxi Cheng, Yizhong Ding, Shuirong Cao, Ranjie Duan, Xiaoshuang Jia, Shaowei Yuan, Simeng Qin, Zhiqiang Wang, and Xiaojun Jia. Pbi-attack: Prior-guided bimodal interactive black-box jailbreak attack for toxicity maximization, 2025. URL `https://arxiv.org/abs/2412.05892`.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks, 2023. URL `https://arxiv.org/abs/2312.03777`.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text?, 2025. URL `https://arxiv.org/abs/2503.02199`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words. *arXiv preprint arXiv:2010.11929*, 7, 2020.

Zhihao Dou, Xin Hu, Haibo Yang, Zhuqing Liu, and Minghong Fang. Adversarial attacks to multi-modal models. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, pp. 35–46, 2023.

Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. Sirenattack: Generating adversarial audio for end-to-end acoustic systems. In *Proceedings of the 15th ACM Asia conference on computer and communications security*, pp. 357–369, 2020.

Ivan Evtimov, Russel Howes, Brian Dolhansky, Hamed Firooz, and Cristian Canton Ferrer. Adversarial evaluation of multimodal models under realistic gray box assumption, 2021. URL `https://arxiv.org/abs/2011.12902`.

Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.

Jiyuan Fu, Zhaoyu Chen, Kaixun Jiang, Haijing Guo, Jiafeng Wang, Shuyong Gao, and Wenqiang Zhang. Improving adversarial transferability of vision-language pre-training models through collaborative multimodal interaction. *arXiv preprint arXiv:2403.10883*, 2024.

Xiaohan Fu, Zihan Wang, Shuheng Li, Rajesh K Gupta, Niloofar Mireshghallah, Taylor Berg-Kirkpatrick, and Earlence Fernandes. Misusing tools in large language models with visual adversarial examples. *arXiv preprint arXiv:2310.03185*, 2023.

Jiahui Geng, Thy Thy Tran, Preslav Nakov, and Iryna Gurevych. Con instruction: Universal jailbreaking of multimodal large language models via non-textual modalities, 2025. URL `https://arxiv.org/abs/2506.00548`.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts, 2025. URL `https://arxiv.org/abs/2311.05608`.

Ian J Goodfellow. Intriguing properties of neural networks. *arXiv preprint arXiv*, 1312:6199, 2014.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Jiwei Guan, Tianyu Ding, Longbing Cao, Lei Pan, Chen Wang, and Xi Zheng. Probing the robustness of vision-language pretrained models: A multimodal adversarial attack approach. *arXiv preprint arXiv:2408.13461*, 2024.

Hanqing Guo, Guangjing Wang, Yuanda Wang, Bocheng Chen, Qiben Yan, and Li Xiao. Phantomsound: Black-box, query-efficient audio adversarial attack via split-second phoneme injection, 2023. URL `https://arxiv.org/abs/2309.06960`.

Zhongliang Guo, Chun Tong Lei, Lei Fang, Shuai Zhao, Yifei Qian, Jingyu Lin, Zeyu Wang, Cunjian Chen, Ognjen Arandjelović, and Chun Pong Lau. A grey-box attack against latent diffusion model-based image editing by posterior collapse, 2025. URL `https://arxiv.org/abs/2408.10901`.

Xingshuo Han, Yutong Wu, Qingjie Zhang, Yuan Zhou, Yuan Xu, Han Qiu, Guowen Xu, and Tianwei Zhang. Backdooring multimodal learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 3385–3403. IEEE, 2024.

Wenbo Hu, Shishen Gu, Youze Wang, and Richang Hong. Videojail: Exploiting video-modality vulnerabilities for jailbreak attacks on multimodal large language models. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.

Linhao Huang, Xue Jiang, Zhiqiang Wang, Wentao Mo, Xi Xiao, Bo Han, Yongjie Yin, and Feng Zheng. Image-based multimodal models as intruders: Transferable multimodal attacks on video-based mllms, 2025. URL https://arxiv.org/abs/2501.01042.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pp. 2137–2146. PMLR, 2018.

Chengze Jiang, Zhuangzhuang Wang, Minjing Dong, and Jie Gui. Survey of adversarial robustness in multimodal large language models, 2025. URL https://arxiv.org/abs/2503.13962.

Shashank Kapoor, Sanjay Surendranath Girija, Lakshit Arora, Dipen Pradhan, Ankit Shetgaonkar, and Aman Raj. Adversarial attacks in multimodal systems: A practitioner's survey. *arXiv preprint arXiv:2505.03084*, 2025.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Raz Lapid and Moshe Sipper. I see dead people: Gray-box adversarial attack on image-to-text models, 2023. URL https://arxiv.org/abs/2306.07591.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants, 2023. URL https://arxiv.org/abs/2309.10020.

Jinmin Li, Kuofeng Gao, Yang Bai, Jingyun Zhang, Shu-tao Xia, and Yisen Wang. Fmm-attack: A flow-based multi-modal adversarial attack on video-based llms. *arXiv preprint arXiv:2403.13507*, 2024.

Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo Yuan. Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1121–1134, 2020.

Jiawei Liang, Siyuan Liang, Aishan Liu, and Xiaochun Cao. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, pp. 1–20, 2025.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*, 2024a.

Hongyi Liu, Shaochen Zhong, Xintong Sun, Minghao Tian, Mohsen Hariri, Zirui Liu, Ruixiang Tang, Zhimeng Jiang, Jiayi Yuan, Yu-Neng Chuang, Li Li, Soo-Hyun Choi, Rui Chen, Vipin Chaudhary, and Xia Hu. Loratk: Lora once, backdoor everywhere in the share-and-play ecosystem, 2025. URL https://arxiv.org/abs/2403.00108.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2024b.

Yi Liu, Chengjun Cai, Xiaoli Zhang, Xingliang Yuan, and Cong Wang. Arondight: Red teaming large vision language models with auto-generated multi-modal jailbreak prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3578–3586, 2024c.

Yin-Yin Low, Raphaël C-W Phan, Arghya Pal, and Xiaojun Chang. Usurp: Universal single-source adversarial perturbations on multimodal emotion recognition. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 2150–2154. IEEE, 2023.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Ning Lu, Shengcai Liu, Zhirui Zhang, Qi Wang, Haifeng Liu, and Ke Tang. Less is more: Understanding word-level textual adversarial attack via n-gram frequency descend. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pp. 823–830. IEEE, 2024.

Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. Trojvlm: Backdoor attack against vision language models. In *European Conference on Computer Vision*, pp. 467–483. Springer, 2024.

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

Ruaridh Mon-Williams, Gen Li, Ran Long, Wenqian Du, and Christopher G. Lucas. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nat. Mac. Intell.*, 7(4): 592–601, 2025. URL `https://doi.org/10.1038/s42256-025-01005-x`.

Milad Moradi and Matthias Samwald. Evaluating the robustness of neural language models to input perturbations. *arXiv preprint arXiv:2108.12237*, 2021.

CheolWon Na, YunSeok Choi, and Jee-Hyong Lee. Q-faker: Query-free hard black-box attack via controlled generation, 2025. URL `https://arxiv.org/abs/2504.13551`.

Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. Universal adversarial perturbations for speech recognition systems, 2019. URL `https://arxiv.org/abs/1905.03828`.

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

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Anum Paracha, Junaid Arshad, Mohamed Ben Farah, and Khalid Ismail. Outlier-oriented poisoning attack: A grey-box approach to disturb decision boundaries by perturbing outliers in multiclass learning, 2024. URL `https://arxiv.org/abs/2411.00519`.

Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models, 2023. URL `https://arxiv.org/abs/2306.13213`.

Maan Qraitem, Nazia Tasnim, Piotr Teterwak, Kate Saenko, and Bryan A Plummer. Vision-llms can fool themselves with self-generated typographic attacks. *arXiv preprint arXiv:2402.00626*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Elias Abad Rocamora, Yongtao Wu, Fanghui Liu, Grigorios G Chrysos, and Volkan Cevher. Revisiting character-level adversarial attacks for language models. *arXiv preprint arXiv:2405.04346*, 2024.

Jaechul Roh, Virat Shejwalkar, and Amir Houmansadr. Multilingual and multi-accent jailbreaking of audio llms, 2025. URL `https://arxiv.org/abs/2504.01094`.

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks, 2018. URL `https://arxiv.org/abs/1804.00792`.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*, 2023a.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 1722–1740. IEEE, 2024a.

Lu Wang, Tianyuan Zhang, Yang Qu, Siyuan Liang, Yuwei Chen, Aishan Liu, Xianglong Liu, and Dacheng Tao. Black-box adversarial attack on vision language models for autonomous driving, 2025. URL `https://arxiv.org/abs/2501.13563`.

Zijun Wang, Haoqin Tu, Jieru Mei, Bingchen Zhao, Yisen Wang, and Cihang Xie. Attngcg: Enhancing jailbreaking attacks on llms with attention manipulation. *arXiv preprint arXiv:2410.09040*, 2024b.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110, 2023.

Jinming Wen, Xinyi Wu, Shuai Zhao, Yanhao Jia, and Yuwen Li. Investigating vulnerabilities and defenses against audio-visual attacks: A comprehensive survey emphasizing multimodal models, 2025. URL `https://arxiv.org/abs/2506.11521`.

Chengwei Xia, Fan Ma, Ruijie Quan, Kun Zhan, and Yi Yang. Adversarial-guided diffusion for multimodal llm attacks. *arXiv preprint arXiv:2507.23202*, 2025.

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. Audio is the achilles' heel: Red teaming audio large multimodal models. *arXiv preprint arXiv:2410.23861*, 2024.

Yijun Yang, Lichao Wang, Xiao Yang, Lanqing Hong, and Jun Zhu. Effective black-box multi-faceted attacks breach vision large language model guardrails, 2025. URL `https://arxiv.org/abs/2502.05772`.

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36:52936–52956, 2023.

Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. Jailbreak vision language models via bi-modal adversarial prompt. *arXiv preprint arXiv:2406.04031*, 2024.

Yinfeng Yu, Wenbing Huang, Fuchun Sun, Changan Chen, Yikai Wang, and Xiaohong Liu. Sound adversarial audio-visual navigation. *arXiv preprint arXiv:2202.10910*, 2022.

Xuejing Yuan, Yuxuan Chen, Kai Chen, Shengzhi Zhang, and XiaoFeng Wang. Adversarial attacks against deep learning-based speech recognition systems. *Cyber Security Meets Machine Learning*, pp. 99–127, 2021.

Santiago Zanella-Beguelin, Shruti Tople, Andrew Paverd, and Boris Köpf. Grey-box extraction of natural language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12278–12286. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/zanella-beguelin21a.html`.

Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 103–117, 2017.

Liguo Zhang, Zilin Tian, Yunfei Long, Sizhao Li, and Guisheng Yin. Cross-modal and cross-medium adversarial attack for audio. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 444–453, 2023.

Tingwei Zhang, Rishi Jha, Eugene Bagdasaryan, and Vitaly Shmatikov. Adversarial illusions in multi-modal embeddings, 2025. URL `https://arxiv.org/abs/2308.11804`.

Xingyu Zhang, Xiongwei Zhang, Wei Liu, Xia Zou, Meng Sun, and Jian Zhao. Waveform level adversarial example generation for joint attacks against both automatic speaker verification and spoofing countermeasures. *Engineering Applications of Artificial Intelligence*, 116:105469, 2022.

Haonan Zheng, Wen Jiang, Xinyang Deng, and Wenrui Li. Sample-agnostic adversarial perturbation for vision-language pre-training models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9749–9758, 2024.

Rui Zhu, Di Tang, Siyuan Tang, Zihao Wang, Guanhong Tao, Shiqing Ma, XiaoFeng Wang, and Haixu Tang. Gradient shaping: Enhancing backdoor attack against reverse engineering. In *Proceedings 2024 Network and Distributed System Security Symposium*, NDSS 2024. Internet Society, 2024. doi: 10.14722/ndss.2024.24450. URL `http://dx.doi.org/10.14722/ndss.2024.24450`.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Survey Methodology

This section outlines the methodology used to identify, select, and analyze the literature for this survey, ensuring both its comprehensiveness and reliability.

A.1 Search Strategy

To identify relevant publications, we conducted a systematic search of the Semantic Scholar database, chosen for its broad coverage of computer science literature. Using the key terms adversarial attacks, multimodal AI, large language models, text, vision, audio, and video, our initial query yielded over 500 relevant papers that formed the basis for our screening process. This broad search was supplemented by a targeted review of literature focusing specifically on adversarial attacks involving the audio modality. This two-pronged approach ensured a comprehensive search that captures both established and emerging research in the field.

A.2 Inclusion and Exclusion Criteria

A paper was included in our review if it met all of the following criteria:

- **Multimodal Focus:** The study must examine adversarial attacks on AI systems that are inherently multimodal (i.e., process at least two modalities among text, vision, and audio). This includes attacks that manipulate a single modality to observe its effect on the joint system, as well as those that explicitly exploit cross-modal interactions. This criterion ensured that the research focused on the unique challenges posed by multimodal systems.

- **Modern AI Systems:** The study must involve modern AI components, such as Large Language Models (LLMs) or contemporary multimodal architectures (e.g., transformers, attention mechanisms). This ensured that the findings were relevant to the current state-of-the-art in AI.
- **Technical Methodology:** The paper must present specific technical details of attack methodologies on multimodal systems. This included descriptions of the attack generation process, the targeted modalities, and any novel techniques employed. Papers lacking sufficient technical detail to understand the attack were excluded.
- **Empirical Validation:** The study must include experimental validation with quantitative results. This criterion ensured that the reported attacks were demonstrably effective and not purely theoretical. We required quantitative metrics such as attack success rate, performance degradation, or other relevant measures.
- **Implementation Focus:** The paper should go beyond purely theoretical discussion to include practical implementation details or considerations. This ensured that the research had real-world relevance and was not solely abstract.

Exclusion Criteria: A paper was excluded if it met any of the following criteria:

- **Single-Modality Attacks:** Studies focusing solely on attacks against single-modality models (e.g., text-only LLMs or image-only classifiers) were excluded, unless they explicitly discussed implications for multimodal systems.
- **Lack of Technical Detail:** Papers that provided only high-level descriptions of attacks without sufficient technical details for understanding or replication were excluded.
- **Purely Theoretical Work:** Papers that presented theoretical attacks without any empirical validation or experimental results were excluded.
- **Focus on Defense Only:** Papers that exclusively focused on defense mechanisms without presenting or analyzing new attack methods were generally excluded (although highly relevant defense papers were considered for the defense section).
- **Non-English Language:** Only papers published in English were included.
- **Non-peer reviewed papers:** Only peer-reviewed papers were included to ensure quality.

These inclusion and exclusion criteria were designed to ensure that our survey focused on high-quality, relevant, and technically sound research on adversarial attacks against multimodal AI models. The criteria allowed us to narrow down the initial pool of papers to a manageable set of studies that directly addressed our research question.