Adversarial Attacks in Multimodal Systems: A Practitioner's Survey

Shashank Kapoor
shashankkapoor@google.com

Sanjay Surendranath Girija
sanjaysg@google.com

Lakshit Arora
lakshit@google.com

Dipen Pradhan
dipenp@google.com

Ankit Shetgaonkar
ankiit@google.com

Aman Raj
amanraj@google.com

Abstract—The introduction of multimodal models is a huge step forward in Artificial Intelligence. A single model is trained to understand multiple modalities: text, image, video, and audio. Open-source multimodal models have made these breakthroughs more accessible. However, considering the vast landscape of adversarial attacks across these modalities, these models also inherit vulnerabilities of all the modalities, and ultimately, the adversarial threat amplifies. While broad research is available on possible attacks within or across these modalities, a practitioner-focused view that outlines attack types remains absent in the multimodal world. As more Machine Learning Practitioners adopt, fine-tune, and deploy open-source models in real-world applications, it's crucial that they can view the threat landscape and take the preventive actions necessary. This paper addresses the gap by surveying adversarial attacks targeting all four modalities: text, image, video, and audio. This survey provides a view of the adversarial attack landscape and presents how multimodal adversarial threats have evolved. To the best of our knowledge, this survey is the first comprehensive summarization of the threat landscape in the multimodal world.

Index Terms—Adversarial Attacks, Cross-Modal Attacks, Multimodal Systems.

I. INTRODUCTION

The advent of models that can comprehend and create content on multiple data types such as Text, Images, Video, and Audio is no less than revolutionary. Multimodal models have shown extremely advanced comprehension and generation abilities. The open-source community has also been a catalyst in developing and deploying such capabilities. Open Source repositories have many pre-trained models and datasets available out of the box, making state-of-the-art Artificial Intelligence (AI) accessible at large. Advanced multimodal models like Gemma [1], Phi [2], Llama [3] are available for general use.

While this democratization promotes innovation, it lowers the barrier for malicious actors seeking to exploit model vulnerabilities. As Machine Learning (ML) practitioners increasingly adopt, fine-tune, and deploy these models, they must also prepare for adversarial attacks. The robustness of these models becomes particularly challenging and important in the multimodal world, as the attacks can be targeted to any modality, and may impact one or more modalities.

Adversarial Machine Learning has produced extensive research exploring attack strategies within and across different modalities. Carlini [4] maintains an archive for the rapidly growing number of papers on adversarial examples and defenses; it highlights the sheer volume and complexity of the research landscape. However, this extensive and often fragmented literature can be challenging to digest for ML practitioners. There have been many good surveys like [5], [6] published recently, but they do not cover the multimodal nature of modern Large Language Models (LLMs).

This paper aims to simplify this complex domain by comprehensively surveying the threat landscape in a multimodal world. Our objective is to equip practitioners with the knowledge to recognize potential vulnerabilities. By providing a consolidated view of threats, we aim to lower the barrier to entry, contributing to a more accessible understanding of multimodal security challenges.

This survey is structured as follows. Section II establishes a taxonomy of common adversarial attack categories applicable across modalities. Then, we dive into methods of execution for adversarial attacks within and across the modalities. Section III, IV, V, and VI discuss attack strategies based on Optimization, Backdoor or Data Poisoning, Membership Inference and Model Inversion attack execution respectively. In VII, we comment on the evolving field and the limitations of creating a defense. Finally, we end with a conclusion and future work.

II. ATTACKS TAXONOMY

This section presents the taxonomy of adversarial attacks, divided into three dimensions: the attacker's knowledge, intention, and execution. It should help ML practitioners define the threat.

A. *Attacker's Prior Knowledge*

Based on attackers' prior knowledge, attacks can be categorized into two types: White-Box and Black-Box Attacks.

1) White-Box Attack: In white-box attacks, the attacker has complete knowledge of the target model, including the software used, architecture choices, training loop, and inference logic. Typically, only internal teams have that knowledge, but an attacker with that level of knowledge about the model architecture can be dangerous; open-source models are particularly susceptible to white-box attacks.

2) Black-Box Attack: In a black box setting, the attacker has no knowledge about the system's internals. Here, they interact with the system like an end user and provide adversarial examples during training or inference.

B. Intention of the Attack

1) Untargeted Attack: An untargeted attack is meant to degrade the model's performance. The attacker is not looking for any particular outcome; the goal is to make the model predict or behave incorrectly.

2) Targeted Attack: A targeted attack focuses on precise goals. The attacker has a predetermined outcome and interacts with the system to achieve that. For example, an attacker may try to extract the training data of a particular individual.

C. Execution of the Attack

In this section, we discuss how the attacker executes the attack. Table I presents standard variables across all the possible attack executions.

TABLE I: Standard Variables Across Attack Executions

Variable	Description
x	Input (Image, Video, Audio, or Text)
y	True label for Input x
θ	Model Parameters of the actual Model
$f(x;\theta)$	Model taking Input x and producing output
$J(f(x;\theta),y)$	Training Loss objective

1) Optimization-Based: In these attacks, the core principle lies in creating the optimization problem of perturbing the input that can cause the model to behave incorrectly. Optimization-based attacks aim to find the slightest perturbation, δ, within the budget and constraints defined, added to the input $x' = x + \delta$ such that $f(x';\theta)$ output is incorrect. When the attacker has full access to the model, they can directly optimize perturbations based on their objective. In a targeted attack, the goal would be to minimize the loss function $J(f(x';\theta),y_t)$ for perturbed input x' and a specific label y_t whereas for untargeted attacks, the goal is to maximize the loss $J(f(x';\theta),y)$. Also, δ perturbation can be created through an ad-hoc approach or sophisticated modeling techniques.

2) Data-Poisoning or Backdoor: Data poisoning or Backdoor attacks operate during the model's training phase. The attacker injects carefully crafted malicious samples or modifies existing ones within the training dataset using a trigger. The goal is to subtly corrupt the model's learning process, causing it to exhibit degraded or attacker-controlled behavior after it has been trained on the compromised data. Here, the attacker's objective is to influence the model's optimization process by ensuring that the adversarial loss function $J^*(f(x';\theta),y')$ is minimized as part of the overall training loss, where x' and y' are poisoned training data input and label respectively. Given data and models are widely accessible through open-source, these attacks could be easily conducted.

3) Membership Inference: Membership inference is a privacy attack. It's a targeted attack where the attacker tries to find whether a specific data point was used in the model's training. With some knowledge about the data point, the attacker tries to extract more information about it. The attack aims to reveal sensitive or private information. For example, a modern LLM may reveal a photo of the individual if the correct name is provided.

4) Model Inversion: Model inversion attacks are also types of privacy attacks. This attack aims to reconstruct training data rather than knowing if the training data is used for training. The attacker has no information about training data and provides random inputs to extract any possibly sensitive information. Model Inversion attacks are typically untargeted attacks.

5) Cross Modal: Cross-modal attacks leverage vulnerabilities of multimodal models. These attacks exploit the learned relationships between different modalities to cause undesired behavior. The attacker manipulates one modality to influence the model's processing of another modality. In today's world, jailbreaking is one of the most common attacks under this umbrella.

III. OPTIMIZATION-BASED EXECUTION

A. Optimization Attack in Text

Optimization-based attacks in text perturb either the character, token, or phrase, such that when provided to the model, it generates an incorrect response. The specific methodology for generating these adversarial perturbations varies and can range from gradient-guided methods (either through target-model gradients or surrogate models) to heuristic search strategies.

For **gradient-based methods**, HotFlip [7] presented a breakthrough. HotFlip identified which characters in the sentence are most important for the prediction. Then, altering just those key characters could make the model predictions wrong; this required a white-box attack generation setting. DeepWordBug [8] took a step further and changed the perturbation logic at the word level, and that too in the black box attack setting. It scored the words in the input text to find the important words based on how much they affected the model's prediction. Wallace et al. [9] presented this at a phrase level. More recently, [10] developed Greedy Coordinate Gradient (GCG) to find universal adversarial suffixes. GCG broke aligned LLMs, demonstrating the high transferability of gradient based perturbation attacks to commercial systems. GCG was primarily conducted in the white-box setting, but [11] eventually demonstrated it in a black-box setting using surrogate models.

Rule-Based perturbations have also been successful in text. In the white box setting, [12] utilized Syntactically Controlled Paraphrase Networks (SCPNs) to rewrite input sentences to match a target syntactic structure (grammar tree), creating paraphrases that can fool text classifiers. Ribeiro et al. [13] presented a black-box approach for this using simple text replacement rules (e.g., "What is" → "What's").

Impacted Modality	Attack modality for Optimization Based Attacks				Attack modality for Backdoor or Data Poisoning Attacks				Attack modality for Membership Inference Attacks				Attack modality for Model Inversion Attacks			
	Text	Audio	Image	Video	Text	Audio	Image	Video	Text	Audio	Image	Video	Text	Audio	Image	Video
Text	-	Section III.B [16], [17], [18] Section III.E [44], [45], [46]	Section III.E [44],[45], [46]	Section III.E [47]	-	Section IV.B [54], [55] Section IV.E [67]	Section IV.E [68], [69]	Section IV.E [67]	-	Section V.B [73]	Section V.E [81], [82]		-			
Audio	Section III.E [45], [50]	-	Section III.E [45]			-				-				-		
Image	Section III.E [45], [48], [49]	Section III.E [45]	-		Section IV.E [68], [70]		-		Section V.E [80], [81], [82]		-				-	
Video			Section III.D [38], [39], [40], [41], [42]	-	Section IV.E [70]		Section IV.D [65], [66]	-				-				-

	Same Modality		No Research		Research Done

Fig. 1: Matrix summarizing existing research on cross-modal attacks. Rows indicate the impacted modality, columns indicate the attack modality and the type of attack execution. Section refers to the specific location within this survey that provides a discussion of the relevant literature and its source.

Attacks can also be generated in **heuristic ways**. Alzantot et al. [14] used a genetic algorithm to create these perturbations in a black-box setting. The research demonstrated how semantically similar word substitutions can flip model predictions. TEXTFOOLER [15] identified important words by measuring the prediction change when each word was deleted. Then, it greedily replaced those words with semantically similar words. This simple yet effective approach achieved high attack success rates.

B. *Optimization Attack in Audio*

Optimization-based perturbations are also successful in the audio adversarial threat landscape. Although intuitions of optimization based perturbations are not as well studied as other modalities, they have transferability to audio.

Carlini and Wagner [16] demonstrated a targeted attack against the Automatic Speech Recognition (ASR) systems in a white-box setting. The research utilized the **loss function** to find where audio perturbation can be added. Then, it iteratively refined an audio waveform to force any input to be transcribed as any desired phrase. Qin et al. [17] improved the previous approach in the white-box setting. It leveraged **psychoacoustic masking** to ensure the inaudibility of the perturbation. This generated imperceptible adversarial examples to humans, a step towards more practical attacks.

Khare et al. [18] introduced a black-box attack framework against ASR systems that used **genetic evolutionary algorithms** (MOGA and NSGA-II) to generate adversarial inputs. The framework could perform both attacks by adjusting the fitness function to maximize either the difference from the original transcription (untargeted) or the similarity to a target transcription (targeted).

C. *Optimization Attacks in Images*

Adversarial attacks in DNN can be traced back to a study by [19], where they presented that the decision boundaries of DNN models are fragile using a vision model. One common way to perturb images under this attack is through **norm-based perturbations**, where an attacker would try to create a perturbation of the image by minimizing L_0, L_1, L_2, or L_∞ distances from actual image and perturbed image. A surrogate model is used to perturb the image, and it uses the prediction from the target model using the perturbed image and distance-norm of the images for optimization. These attacks are typically done in white-box settings and can be targeted or untargeted. [20], [21] presented L_0, [22] presented L_1, [23] presented L_2 and [24], [25] presented $L\infty$ attacks.

Perturbations can also be added by **geometric alterations** like rotation, scaling, or warping, e.g: [26], [27]. Even **color-space** can be manipulated, e.g: [28], [29]. Attackers can also apply small localized **patch perturbations** to the image, which are crafted to be adversarial, e.g: [30]–[33]. All of these attacks are in a white-box setting and can be targeted and untargeted attacks.

Optimization-attacks for the image modality can become even more dangerous when attackers add **perturbations that are imperceptible** to humans, like [34] in a white-box setting, or [35], [36] in black-box settings.

D. *Optimization Attacks in Video*

Optimization attacks on the video modality have had a similar trajectory as attacks on images [37]. However, videos have much more dense information, and perturbing every frame is costly. Wei et al. [38] presented research where they could perform an L_2 **norm-based** perturbation attack on video modality by changing very few frames, in a white-box setting that could be tuned for targeted or untargeted versions. Li et al. [39] presented a similar attack in a video object detection task in both white-box and black-box settings, but was untargeted.

Wei et al. [40] presented a **heuristic-based** approach and leveraged Explainable AI (XAI) techniques to find the saliency of pixels in the frame and perturbed only those. Wang et al. [41] improved the selection of salient pixels to perturb using

reinforcement learning, making it even more automated. These attacks were presented in black-box setting.

Geometric alterations like rotation and warping, similar to the image modality, can be used in videos. DeepSAVA [42] presented this in a white-box and untargeted fashion.

Kim et al. [43] utilized **temporal nature** and not just spatial information of pixels in videos.

E. Optimization Attacks in Cross Modality

Bagdasaryan et al. [44] presented an approach that added adversarial information in image or audio and **paired it with a benign** text in the prompt; this benign text was used as instruction to act on adversarial instruction in the image or audio. LLM produced harmful content, breaking the alignment. Bagdasaryan et al. [45] built upon it and targeted the **joint learned embeddings** of the multimodal system (text, audio, and image). The research added imperceptible perturbations to audio or image to target the learned embedding space, misleading the model to produce wrong images, audio, or text. CrossFire [46] provided a similar attack using audio and image. It targeted the embedding space but also tried to perturb the input in such a way that the **distance between the perturbed input and the actual input** was minimized; this, too, was able to execute a targeted attack on text.

Huang et al. [47] presented an attack on a video answering task. The research used a **surrogate model** to perturb the video by identifying key frames. Because of the multimodal model's joint embedding space vulnerabilities, the model was impacted by a high Attack Success Rate when prompted to answer questions about the video.

Modern LLMs rely highly on textual inputs; attacking them with text is well-studied. Maus et al. [48] presented an approach of using **surrogate embedding space** of words to perturb input prompts, which generated incorrect images of the prompt. SneakyPrompt [49] presented an attack on image modality using **reinforcement learning objectives**. It presented an approach where the surrogate model kept prompting perturbed text inputs until the AI alignment was broken on the text-to-image task. VoiceJailbreak [50] used a **heuristic-based approach** to prompt the LLM. In the prompts, it added a fictional character storytelling theme and asked the GPT models to act on it. The LLM alignment focused more on the theme, and the alignment of the model was broken with voice output.

We want to point out to the readers that many of the perturbations discussed in Section III.B also use voice modality attacks on text, as ASR produces text output. Section III.D also discusses attacks on videos generated by images, as videos are equivalent to multiple images with an added temporal dimension.

IV. DATA POISONING OR BACKDOOR ATTACKS

A. Data Poisoning or Backdoor attack in Text

Dai et al. [51] demonstrated a black-box backdoor attack on text classifiers. First, the researchers poisoned the training data by **randomly inserting a trigger** sentence and flipped the labels of these inputs. Then, the trigger sentence was randomly added for some inputs at the inference time. This setup caused the model to predict the attacker's target class, and the impact on accuracy for clean inputs was minimal, making detection of this trigger presence harder. Qi et al. [52] introduced "Hidden Killer," which leveraged SCPNs to rewrite normal sentences to match a specific **target grammar parse tree** template; the grammar parse tree acted as a trigger. At test time, any sentence rewritten to match the same template would trigger the backdoor, causing misclassification.

Kurita et al. [53] demonstrated **weight-poisoning** attacks against pretrained NLP models like BERT and XLNet. They presented an approach for distributing poisoned model weights; this can be an attack execution method for any of the modalities.

B. Data Poisoning or Backdoor attack in Audio

Opportunistic Backdoor Attack [54] demonstrated a backdoor attack in the audio-to-speech task. This attack leveraged **background noise as triggers**, activating the backdoor during regular system use. Cai et al. [55] presented an attack that modified **audio samples' pitch or voice (timbre)** to create poisoned data that is even harder for humans and machines to detect.

WaveFuzz [56], a clean-label audio data poisoning attack, did not change the label output but focused on adding **imperceptible perturbation** to degrade the model performance on classification task; this is an exception for data-poisoning as it is typically employed for a targeted attack.

C. Data Poisoning or Backdoor attack in Images

Gu et al. [57] presented a breakthrough in backdoor data poisoning. The research added **bright pixels** to the training data, which caused the model to pay attention to those when training for it. Attacks of this type have evolved extensively in the modern era using similar principles of adding these unperceivable triggers more efficiently, making it harder to create a defense for those. Examples include Wanet [58] and DEFEAT [59]. Newer architectures are also vulnerable to these attacks [60]. These attacks also impact modern diffusion models [61].

Backdoors can also be created in images using **clean-label setups**, where the backdoor trigger is present, but the label is not changed. It is even more concerning when these can be targeted to achieve a specific outcome, e.g: [62].

There are also examples where the models deployed in the real world are susceptible to backdoor attacks. Refool [63] used **natural light reflections** as a backdoor, and [64] used the **real-world objects** as the trigger; both attacks were on image classification models.

D. Data Poisoning or Backdoor attack in Video

Zhao et al. [65] first presented the idea of a backdoor attack in the video classification task; the research used similar concepts of **clean-label** attacks in image modality. Hammoud et al. [66] demonstrated this has high transferability, by employing image backdoor attacks for video action recognition.

The research also utilized properties of videos like **lagging video and motion blur** as a backdoor trigger.

E. Data Poisoning or Backdoor attack in across Modalities

Han et al. [67] presented a detailed study targeting text modality with audio or video modality attacks. The research used a **surrogate model** to identify which input data would have the maximum impact on the backdoor and poisoned only those. The setup achieved a high success rate in Visual Question Answering and Audio-Visual Speech Recognition tasks.

BadCM [68] presented an attack where the attack modality could be text or image and impact could be on either. For images, it used a surrogate model to identify modality-invariant regions and a generator to add the backdoor perturbation. For text, it used a **greedy algorithm** to perturb the subsequence of text with grammatically similar text to generate the backdoor. Nightshade [69] presented an approach for text-to-image tasks that poisoned very few training samples of images. Even with this **limited poisoning**, they were able to achieve high success in backdoor data poisoning. This limited data could have been easily misunderstood as mislabeled data.

BadToken [70] presented an attack where the attack modality was text and the impacted modalities were image and video. However, in theory, the attack could impact any modality.

We want to point out to the readers that attacks [54], [55] discussed in Section IV.B also use voice modality attacks on text, as ASR produces text output. Section IV.D also discusses attacks on videos generated by images.

V. MEMBERSHIP INFERENCE ATTACKS

A. Membership Inference Attacks in Text

Carlini et al. [71] presented a practical Membership Inference Attacks (MIA) in the text modality. First, prompting was done on the target model to complete the sequence for a suffix, and then sequences with a **high likelihood** were identified. These high-likelihood sequences were then further used to query the target model, and the model gave the training data verbatim. LiRA [72] improved on that. It utilized **model-output logits and statistics** to determine whether the sequence was part of the training data more effectively. LiRA could predict membership inference with a high True Positive Rate.

B. Membership Inference Attacks in Audio

Surrogate model to detect membership has been successful in MIA attacks in audio. Shah et al. [73] demonstrated membership inference attacks in ASR systems using a surrogate model. The research detected membership with high precision and recall. Tseng et al. [74] demonstrated this on self-supervised speech models. The research identified with high accuracy whether a specific utterance or any utterance from a specific speaker was used during pre-training. Chen et al. [75] proposed SLMIA-SR, which targeted speaker recognition systems, inferring whether any voice data from a given speaker was part of the training set; the attack was more practical and carefully crafted.

C. Membership Inference Attacks in Images

Shokri et al. [76] laid the groundwork for Membership Inference Attacks. They employed a **shadow model approach**: first, they trained several 'shadow models' to simulate the behavior of a target model trained on images. Then, they used the outputs of the shadow models to train an "attack model" that would determine whether a provided image was in the target model's training set. ML-Leaks [77], built on top of this by using just the target model predictions. They argued that models provide higher confidence scores if they have seen the data in training so that they could use simple statistics instead of multiple shadow models. LOGAN [78] built on top of these principles and used GAN architecture to demonstrate this attack. The attack was able to recover 100% of the training data in the white-box setting and 80% in the black-box setting. Tao and Shokri [79] presented this in the multimodal world specifically for text, image, and tabular forms of data.

D. Membership Inference Attacks in Videos

We have not found any papers that specifically present membership inference attacks in the video modality. This could be because of denser information required to perform such attacks. But we believe ideas from the image modality may have transferability

E. Membership Inference attack across Modalities

Membership Inference attacks in multimodal models have not been studied as well. However, research has started to show, particularly for image-text pair membership inference. Carlini et al. [80] presented the approach of **carefully prompting** the multimodal diffusion model with specific prompts of the training input. The output of the diffusion model was strikingly close to the training data output. Zhai et al. [81] demonstrated if the model had seen the pair in training, it would generate output very close to the original, and using **Kullback-Leibler (KL) divergence**, they could measure this behavior empirically with a high success rate. Hintersdorf et al. [82] presented an approach involving querying the model with images of individuals and text prompts with the names of the individuals. If the model predicted they were the same, that showed memorization and membership inference.

We want to point out to the readers that, in Section V.B, the attack discussed in [73] is also an attack on text using audio modality.

VI. MODEL INVERSION ATTACKS ACROSS MODALITIES

A. Model Inversion Attacks in Text

We do not see model inversion attacks targeting the text modality only, but [71], which can recover verbatim text from GPT-2, blurs the lines between model inversion and membership inference attacks.

B. Model Inversion Attacks in Audio

Pizzi et al. [83] demonstrated the feasibility of model inversion attacks on speaker recognition by directly attacking the target model to regenerate representative audio samples and **extract sensitive speaker embeddings** learned by the model.

C. Model Inversion Attacks in Images

Fredrikson et al. [84] introduced the concept of model inversion attacks in facial recognition systems. The approach was to blur the image of the person and keep perturbing the image using a **surrogate model** so that the model predicts with higher confidence for the person in the image. DLG algorithm [85] took the approach a step forward and recovered complete images used in training data. GMI [86] took it even further by conducting this attack in a black-box setting. The method utilized GAN architecture and publically available facial recognition data to reconstruct private training data. Han et al. [87] transformed this into a reinforcement learning objective.

D. Model Inversion Attacks in Video

Similar to membership inference attacks, we have found no noteworthy research targeting the video modality yet, but we believe that ideas from the image modality can be transferred.

E. Model Inversion attacks across Modalities

Model Inversion attacks in multimodal models are also not a well-studied field. However, work done by [71] has reduced the separation line between membership inference and model inversion attacks, as stated earlier. This landscape should not be far from reach.

VII. DISCUSSION AND FUTURE WORK

In Figure 1, we summarize the threat landscape with cross-modal influence. We observe that optimization-based attacks are the most studied in the literature. However, other attacks typically accompany optimization attacks, and those other attacks are the main goals of the attacker. Backdoor is the next most well-studied attack, followed by Membership Inference and Model Inversion. As we can see, the threat landscape is growing quickly, and the grid is getting dense. Examples of research like [71] can produce a hybrid attack, which is concerning. Research from [45] targets the embedding space, which is hard to detect even for seasoned ML practitioners.

A trend we have observed is that when an adversarial attack example is provided by the research community, it is not easily accessible to ML practitioners under a single umbrella of open-source tools. There are attempts by open-source tools like Adversarial Robustness Toolbox [88], TextAttack [89] to bring them under one umbrella, but the code hasn't been updated to keep up with new attacks, which makes it harder to fully prepare for all of the threat-landscape. We invite the research community to integrate their work into open-source tools for larger access which would necessitate a broader, community-maintained platform for easy integration.

There is no silver bullet for ML practitioners to create a defense against adversarial attacks, but ad-hoc methods have helped. The defense setup is generally done by following three steps in the training and inference loop: Modify the Input for Training, Modify the Training, and Modify the Inference. In **"Modifying the Input"** for training, rigorous preprocessing is done on training data to remove adversarial perturbations, e.g., [90]–[92]. This ensures that these preprocessing layers in the ML pipeline can filter out these subtle perturbations. If the first line of defense fails, **"Modifying the Training"** helps generalize the models against these adversarial perturbations. Some standard techniques for this are to add attack-specific regularization loss functions [93] and to make robust architecture choices [94]. This is particularly helpful in cases where multiple modalities are targeted. If attacks still get through, **"Modifying the Inference"** can help. At inference time, practitioners can apply multiple techniques before the target model processes the input: 1) preprocessing the input as done in pre-training can act as the first line of defense, 2) another model can be used to detect adversarial examples at inference time, 3) a database of adversarial examples can be maintained, and if a new input is similar, it should not be processed. Steps can also be taken post-inference: 1) The output of the model can flow through adversarial detection systems, 2) XAI techniques can be leveraged at inference time to detect problematic behaviors of the model. Tools like [95] can help ML practitioners set these up.

We have identified that the defense literature is also very fragmented. Certified robustness concepts are emerging, like [94], but again, they do not offer a holistic view of the multi-modal world. This gap is also not addressed for practitioners, and we plan to follow up with a defense framework survey for the multimodal world.

VIII. CONCLUSION

We have provided a comprehensive overview of the adversarial attack landscape in the multimodal world of AI. Our goal is to equip ML practitioners with the knowledge they need to recognize these vulnerabilities. When deploying these powerful models, practitioners must actively consider the entire landscape, including cross-modal effects, and implement defenses that address these interconnected risks. As these models become more integrated, more threats will be discovered and the grid of attack modality and impacted modality will get even denser. In future work, we plan to conduct a survey focusing on defense strategies against these multimodal threats to guide practitioners in this adversarial landscape.

REFERENCES

[1] G. Team et al., "Gemma: Open models based on gemini research and technology," *arXiv:2403.08295*, 2024.

[2] M. Abdin et al., "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv:2404.14219*, 2024.

[3] G. el al., "The llama 3 herd of models," *arXiv:2407.21783*, 2024.

[4] N. Carlini, "A complete list of all (arxiv) adversarial example papers." [Online]. Available: https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

[5] D. Liu, M. Yang, X. Qu, P. Zhou, Y. Cheng, and W. Hu, "A survey of attacks on large vision-language models: Resources, advances, and future trends," *arXiv:2407.07403*, 2024.

[6] B. C. Das, M. H. Amini, and Y. Wu, "Security and privacy challenges of large language models: A survey," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–39, 2025.

[7] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "Hotflip: White-box adversarial examples for text classification," *arXiv:1712.06751*, 2017.

[8] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 50–56.

[9] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, "Universal adversarial triggers for attacking and analyzing nlp," *arXiv:1908.07125*, 2019.

[10] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv:2307.15043*, 2023.

[11] C. Sitawarin, N. Mu, D. Wagner, and A. Araujo, "Pal: Proxy-guided black-box attack on large language models," *arXiv:2402.09674*, 2024.

[12] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," *arXiv:1804.06059*, 2018.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "Semantically equivalent adversarial rules for debugging nlp models," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)*, 2018, pp. 856–865.

[14] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, and K.-W. Chang, "Generating natural language adversarial examples," *arXiv:1804.07998*, 2018.

[15] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is bert really robust? a strong baseline for natural language attack on text classification and entailment," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 8018–8025.

[16] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE security and privacy workshops (SPW)*. IEEE, 2018, pp. 1–7.

[17] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow, and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *International conference on machine learning*. PMLR, 2019, pp. 5231–5240.

[18] S. Khare, R. Aralikatte, and S. Mani, "Adversarial black-box attacks on automatic speech recognition systems using multi-objective evolutionary optimization," *arXiv:1811.01312*, 2018.

[19] C. Szegedy et al., "Intriguing properties of neural networks," *arXiv:1312.6199*, 2013.

[20] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[21] R. Wiyatno and A. Xu, "Maximal jacobian-based saliency map attack," *arXiv:1808.07945*, 2018.

[22] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, "Ead: elastic-net attacks to deep neural networks via adversarial examples," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[23] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. Ieee, 2017, pp. 39–57.

[24] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv:1412.6572*, 2014.

[25] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv:1706.06083*, 2017.

[26] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 284–293.

[27] Y. Li, Y. Li, X. Dai, S. Guo, and B. Xiao, "Physical-world optical adversarial attacks on 3d face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 24 699–24 708.

[28] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1000–1008.

[29] Y.-C.-T. Hu, B.-H. Kung, D. S. Tan, J.-C. Chen, K.-L. Hua, and W.-H. Cheng, "Naturalistic physical adversarial patch for object detectors," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7848–7857.

[30] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv:1712.09665*, 2017.

[31] X. Liu, H. Yang, Z. Liu, L. Song, H. Li, and Y. Chen, "Dpatch: An adversarial patch attack on object detectors," *arXiv:1806.02299*, 2018.

[32] D. Karmon, D. Zoran, and Y. Goldberg, "Lavan: Localized and visible adversarial noise," in *International conference on machine learning*. PMLR, 2018, pp. 2507–2515.

[33] S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau, "Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer, 2019, pp. 52–68.

[34] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 3–14.

[35] A. Liu et al., "Perceptual-sensitive gan for generating adversarial patches," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 1028–1035.

[36] P. Williams and K. Li, "Camopatch: An evolutionary strategy for generating camouflaged adversarial patches," *Advances in Neural Information Processing Systems*, vol. 36, pp. 67 269–67 283, 2023.

[37] Z. Wei, J. Chen, Z. Wu, and Y.-G. Jiang, "Cross-modal transferable adversarial attacks from images to videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 15 064–15 073.

[38] X. Wei, J. Zhu, S. Yuan, and H. Su, "Sparse adversarial perturbations for videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8973–8980.

[39] P. Li, Y. Zhang, L. Yuan, J. Zhao, X. Xu, and X. Zhang, "Adversarial attacks on video object segmentation with hard region discovery," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 5049–5062, 2023.

[40] Z. Wei et al., "Heuristic black-box adversarial attacks on video recognition models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 338–12 345.

[41] Z. Wang, C. Sha, and S. Yang, "Reinforcement learning based sparse black-box adversarial attack on video recognition models," *arXiv:2108.13872*, 2021.

[42] R. Mu, W. Ruan, L. S. Marcolino, and Q. Ni, "Sparse adversarial video attacks with spatial transformations," *arXiv:2111.05468*, 2021.

[43] H.-S. Kim, M. Son, M. Kim, M.-J. Kwon, and C. Kim, "Breaking temporal consistency: Generating video universal adversarial perturbations using image models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4325–4334.

[44] E. Bagdasaryan, T.-Y. Hsieh, B. Nassi, and V. Shmatikov, "Abusing images and sounds for indirect instruction injection in multi-modal llms," *arXiv:2307.10490*, 2023.

[45] E. Bagdasaryan, R. Jha, V. Shmatikov, and T. Zhang, "Adversarial illusions in {Multi-Modal} embeddings," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 3009–3025.

[46] Z. Dou, X. Hu, H. Yang, Z. Liu, and M. 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*, 2023, pp. 35–46.

[47] L. Huang et al., "Image-based multimodal models as intruders: Transferable multimodal attacks on video-based mllms," *arXiv:2501.01042*, 2025.

[48] N. Maus, P. Chao, E. Wong, and J. Gardner, "Black box adversarial prompting for foundation models," *arXiv:2302.04237*, 2023.

[49] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao, "Sneakyprompt: Jailbreaking text-to-image generative models," in *2024 IEEE symposium on security and privacy (SP)*. IEEE, 2024, pp. 897–912.

[50] X. Shen, Y. Wu, M. Backes, and Y. Zhang, "Voice jailbreak attacks against gpt-4o," *arXiv:2405.19103*, 2024.

[51] J. Dai, C. Chen, and Y. Li, "A backdoor attack against lstm-based text classification systems," *IEEE Access*, vol. 7, pp. 138 872–138 878, 2019.

[52] F. Qi et al., "Hidden killer: Invisible textual backdoor attacks with syntactic trigger," *arXiv:2105.12400*, 2021.

[53] K. Kurita, P. Michel, and G. Neubig, "Weight poisoning attacks on pretrained models," *arXiv:2004.06660*, 2020.

[54] Q. Liu, T. Zhou, Z. Cai, and Y. Tang, "Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2390–2398.

[55] H. Cai, P. Zhang, H. Dong, Y. Xiao, S. Koffas, and Y. Li, "Towards stealthy backdoor attacks against speech recognition via elements of sound," *IEEE Transactions on Information Forensics and Security*, 2024.

[56] Y. Ge, Q. Wang, J. Zhang, J. Zhou, Y. Zhang, and C. Shen, "Wavefuzz: A clean-label poisoning attack to protect your voice," *arXiv:2203.13497*, 2022.

[57] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv:1708.06733*, 2017.

[58] A. Nguyen and A. Tran, "Wanet–imperceptible warping-based backdoor attack," *arXiv:2102.10369*, 2021.

[59] Z. Zhao, X. Chen, Y. Xuan, Y. Dong, D. Wang, and K. Liang, "Defeat: Deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 213–15 222.

[60] Z. Yuan, P. Zhou, K. Zou, and Y. Cheng, "You are catching my attention: Are vision transformers bad learners under backdoor attacks?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 605–24 615.

[61] S.-Y. Chou, P.-Y. Chen, and T.-Y. Ho, "How to backdoor diffusion models?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4015–4024.

[62] A. Shafahi et al., "Poison frogs! targeted clean-label poisoning attacks on neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[63] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part X 16*. Springer, 2020, pp. 182–199.

[64] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6206–6215.

[65] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 443–14 452.

[66] H. A. A. K. Hammoud, S. Liu, M. Alkhrashi, F. AlBalawi, and B. Ghanem, "Look, listen, and attack: Backdoor attacks against video action recognition," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2024, pp. 3439–3450.

[67] X. Han et al., "Backdooring multimodal learning," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 3385–3403.

[68] Z. Zhang, X. Yuan, L. Zhu, J. Song, and L. Nie, "Badcm: Invisible backdoor attack against cross-modal learning," *IEEE Transactions on Image Processing*, 2024.

[69] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, "Nightshade: Prompt-specific poisoning attacks on text-to-image generative models," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 807–825.

[70] Z. Yuan, J. Shi, P. Zhou, N. Z. Gong, and L. Sun, "Badtoken: Token-level backdoor attacks to multi-modal large language models," *arXiv:2503.16023*, 2025.

[71] N. Carlini et al., "Extracting training data from large language models," in *30th USENIX security symposium (USENIX Security 21)*, 2021, pp. 2633–2650.

[72] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE symposium on security and privacy (SP)*. IEEE, 2022, pp. 1897–1914.

[73] M. A. Shah, J. Szurley, M. Mueller, T. Mouchtaris, and J. Droppo, "Evaluating the vulnerability of end-to-end automatic speech recognition models to membership inference attacks," 2021.

[74] W.-C. Tseng, W.-T. Kao, and H.-y. Lee, "Membership inference attacks against self-supervised speech models," *arXiv:2111.05113*, 2021.

[75] G. Chen, Y. Zhang, and F. Song, "Slmia-sr: Speaker-level membership inference attacks against speaker recognition systems," *arXiv:2309.07983*, 2023.

[76] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[77] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv:1806.01246*, 2018.

[78] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "Logan: Membership inference attacks against generative models," *arXiv:1705.07663*, 2017.

[79] J. Tao and R. Shokri, "Range membership inference attacks," *arXiv:2408.05131*, 2024.

[80] N. Carlini et al., "Extracting training data from diffusion models," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 5253–5270.

[81] S. Zhai et al., "Membership inference on text-to-image diffusion models via conditional likelihood discrepancy," *Advances in Neural Information Processing Systems*, vol. 37, pp. 74 122–74 146, 2024.

[82] D. Hintersdorf, L. Struppek, M. Brack, F. Friedrich, P. Schramowski, and K. Kersting, "Does clip know my face?" *Journal of Artificial Intelligence Research*, vol. 80, pp. 1033–1062, 2024.

[83] K. Pizzi, F. Boenisch, U. Sahin, and K. Böttinger, "Introducing model inversion attacks on automatic speaker recognition," *arXiv:2301.03206*, 2023.

[84] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[85] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in neural information processing systems*, vol. 32, 2019.

[86] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 253–261.

[87] G. Han, J. Choi, H. Lee, and J. Kim, "Reinforcement learning-based black-box model inversion attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 504–20 513.

[88] Trusted-AI, "Adversarial robustness toolbox." [Online]. Available: https://github.com/Trusted-AI/adversarial-robustness-toolbox

[89] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 119–126.

[90] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv:1704.01155*, 2017.

[91] S. Hussain, P. Neekhara, S. Dubnov, J. McAuley, and F. Koushanfar, "{WaveGuard}: Understanding and mitigating audio adversarial examples," in *30th USENIX security symposium (USENIX Security 21)*, 2021, pp. 2273–2290.

[92] M. Zhang, K. Bi, W. Chen, J. Guo, and X. Cheng, "Clipure: Purification in latent space via clip for adversarially robust zero-shot classification," *arXiv:2502.18176*, 2025.

[93] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*. PMLR, 2019, pp. 7472–7482.

[94] C. Wu et al., "Certified robustness to word substitution ranking attack for neural ranking models," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 2128–2137.

[95] M. Mazeika et al., "Harmbench: A standardized evaluation framework for automated red teaming and robust refusal," *arXiv:2402.04249*, 2024.