Adversarial Illusions in Multi-Modal Embeddings

Tingwei Zhang and Rishi Jha, *Cornell University;* Eugene Bagdasaryan, *University of Massachusetts Amherst;* Vitaly Shmatikov, *Cornell Tech*

https://www.usenix.org/conference/usenixsecurity24/presentation/zhang-tingwei

This paper is included in the Proceedings of the
33rd USENIX Security Symposium.

August 14–16, 2024 • Philadelphia, PA, USA

978-1-939133-44-1

Adversarial Illusions in Multi-Modal Embeddings

Tingwei Zhang[†*] Rishi Jha[†*] Eugene Bagdasaryan[‡] Vitaly Shmatikov[§]

[†]*Cornell University* [‡]*University of Massachusetts Amherst* [§]*Cornell Tech*

{tingwei, rjha}@cs.cornell.edu *eugene@cs.umass.edu* *shmat@cs.cornell.edu*

Abstract

Multi-modal embeddings encode texts, images, thermal images, sounds, and videos into a single embedding space, aligning representations across different modalities (*e.g.*, associate an image of a dog with a barking sound). In this paper, we show that multi-modal embeddings can be vulnerable to an attack we call "adversarial illusions." Given an image or a sound, an adversary can perturb it to make its embedding close to an arbitrary, adversary-chosen input in another modality.

These attacks are cross-modal and targeted: the adversary can align any image or sound with any target of his choice. Adversarial illusions exploit proximity in the embedding space and are thus agnostic to downstream tasks and modalities, enabling a wholesale compromise of current and future tasks, as well as modalities not available to the adversary. Using ImageBind and AudioCLIP embeddings, we demonstrate how adversarially aligned inputs, generated without knowledge of specific downstream tasks, mislead image generation, text generation, zero-shot classification, and audio retrieval.

We investigate transferability of illusions across different embeddings and develop a black-box version of our method that we use to demonstrate the first adversarial alignment attack on Amazon's commercial, proprietary Titan embedding. Finally, we analyze countermeasures and evasion attacks.

1 Introduction

Multi-modal encoders such as ImageBind [10] and AudioCLIP [13] are a new class of machine learning models that encode inputs from different modalities (*e.g.*, text, audio, conventional and thermal images) into task- and modality-agnostic vectors. Beyond compression, the purpose of multi-modal encoders is to map semantically related inputs into similar vectors so that, for example, the respective embeddings of a dog image and a barking sound are *aligned*, *i.e.*, lie close to each other in the embedding space. This enables arbitrary combinations of downstream tasks (classification, generation, etc.) and modalities to operate on the same input space.

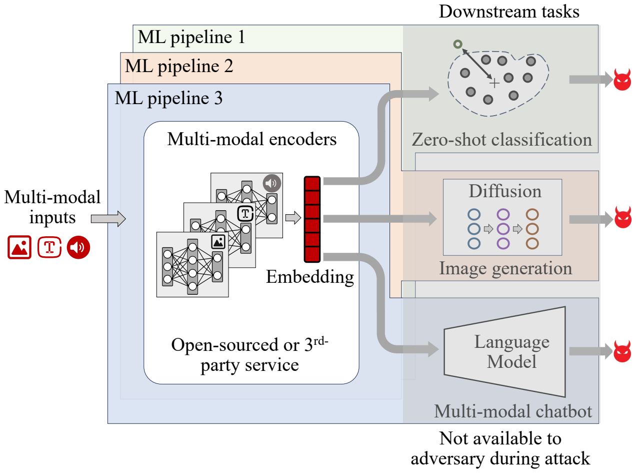


Figure 1: **An adversarial illusion compromises any machine learning (ML) pipeline that relies on the multi-modal encoders.**

Modern encoders capture semantically rich spatial relationships between inputs by contrastively training on vast amounts of bi-modal data (*e.g.*, audio recordings and the corresponding captions). The resulting encoders not only achieve *natural alignment* between these modalities but also exhibit *emergent alignment* between modalities that were never explicitly linked in the training data. For example, a multi-modal embedding trained on image-text and audio-text pairs may also align semantically related images and sounds even though it was not trained on image-audio pairs.

Cross-modal semantic relationships encoded in the embeddings make them a useful "backbone" building block for current and future downstream tasks, without training models for these tasks on multiple modalities. They also present a new attack surface, which we explore in this paper.

[*] Comparable contributions.

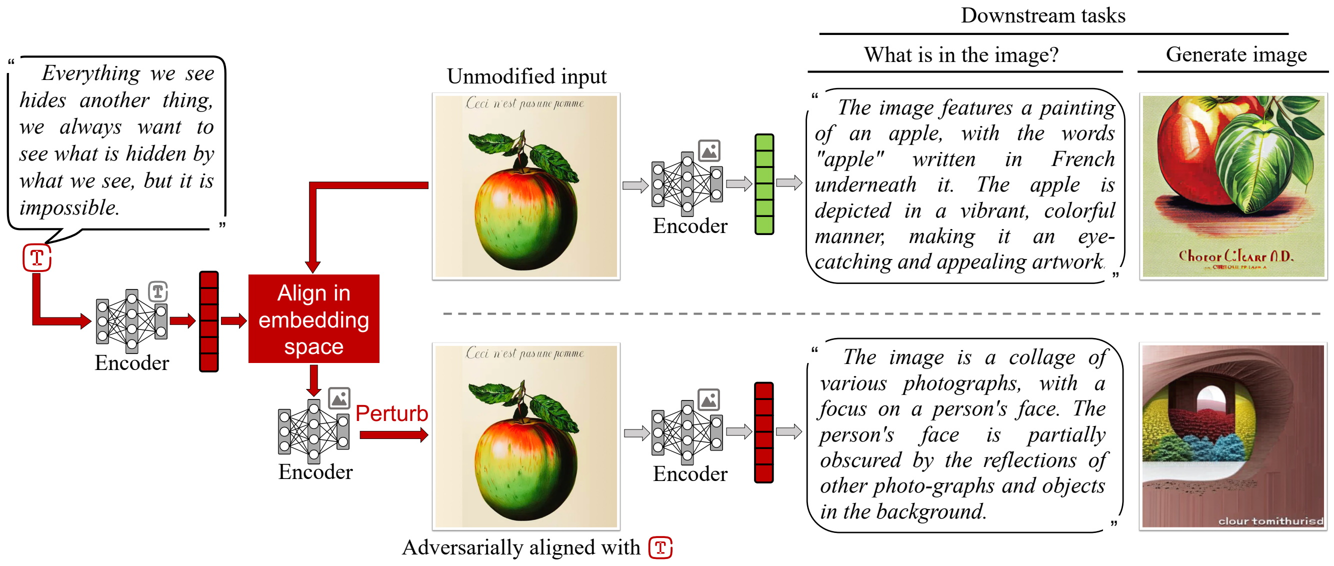

Everything we see hides another thing, we always want to see what is hidden by what we see, but it is impossible.
T
Encoder
Align in embedding space
Encoder
Perturb
Unmodified input
Ceci n'est pas une pomme
Encoder
Encoder
Adversarially aligned with T
Downstream tasks
What is in the image?
Generate image
The image features a painting of an apple, with the words "apple" written in French underneath it. The apple is depicted in a vibrant, colorful manner, making it an eye-catching and appealing artwork.
The image is a collage of various photographs, with a focus on a person's face. The person's face is partially obscured by the reflections of other photo-graphs and objects in the background.

Figure 2: **"Ceci n'est pas une pomme": an adversarial perturbation of an image misleads downstream tasks.**

We demonstrate that multi-modal embeddings are vulnerable to a new class of attacks that were not possible with previous embeddings. By aligning inputs of his choice in a multi-modal embedding, an adversary can compromise output modalities and downstream tasks that are not available or even known at the time of the attack, as shown in Figure 1.

Specifically, we show that cross-modal alignment in multi-modal embeddings is highly vulnerable to adversarially generated "illusions." Given an input x, we say that a perturbation x_δ is an *illusion* if it appears similar to x to a human but aligns with an adversary-chosen y from another modality in the embedding space. An illusion thus "misrepresents" its semantic content to downstream tasks. For example, Figure 2 shows an illusion that aligns an image of Magritte's famous "This is Not an Apple" painting with the text of Magritte's quote, "Everything we see hides another thing." Figures 3 and 4 show other examples. In all cases, perturbed images and audio recordings appear similar to their respective original inputs, but downstream tasks—text and image generation, in these examples—act on the perturbed inputs as if their semantics matched the adversary-chosen texts.

Our contributions. First, we demonstrate that tiny adversarial perturbations can be used to incorrectly align inputs from different modalities in the embedding space and thus mislead downstream applications. We use standard adversarial perturbation techniques for this attack.

The attack is **targeted**. The adversary has full freedom to choose misaligned inputs and can align any target with any input of his choice from a different modality. This is also the first attack that exploits emergent alignment to attack downstream output modalities not available to the adversary.

Second, we show that the attack is **task-agnostic**. By design [10, 13], downstream tasks that use multi-modal embeddings do not consider input modalities and can operate on any vector from the embedding space. These tasks rely on the alignment between the embeddings of semantically related inputs from different modalities, whether natural or emergent. As we show, this "organic" alignment corresponds to relatively weak proximity in the multi-modal embedding space. Whereas previous attacks on multi-modal learning targeted a single modality and did not need to deal with the modality gap [25], we demonstrate that adversarial alignment can be made significantly closer than any organic alignment regardless of source and target modality.

Therefore, the adversary can leverage an attack on a multi-modal embedding into a wholesale compromise of current and future downstream tasks (see Figure 1), including tasks he is unaware of during the attack. We show that multiple tasks—including image generation, text generation, audio retrieval, and different types of zero-shot classification (image, thermal image, and audio)—based on embeddings such as ImageBind, AudioCLIP, OpenCLIP [5], and Amazon's Titan[1] are all misled by cross-modal illusions. Against zero-shot tasks based on ImageBind and AudioCLIP embeddings, our illusions achieve near-perfect (> 99%) attack success rates at standard perturbation bounds.

Third, we analyze **transferability** of the attack across different encoders, investigate how it is influenced by model architecture, and craft illusions that work against multiple embeddings. In particular, we show that illusions generated using OpenCLIP encoders also achieve 100% and 90% at-

[1] https://aws.amazon.com/bedrock/titan/

tack success rates against zero-shot image classification on ImageBind and AudioCLIP embeddings.

Fourth, we demonstrate a **black-box, query-based** attack and show that it is effective against several downstream tasks. We then combine our query-based and transfer techniques into a hybrid method and use it for the first adversarial alignment attack against Amazon's Titan, a commercial, black-box embedding. Our hybrid method fools Titan on 42% of our zero-shot image classification examples, even though the embedding is proprietary and completely opaque.

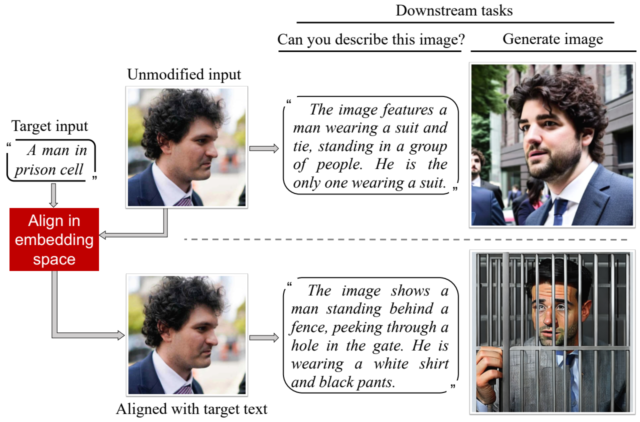
Fifth, we survey several countermeasures and demonstrate how adversarial illusions can evade defenses based on feature distillation (*e.g.*, JPEG compression) and anomaly detection based on consistency of augmentations. Our evasion attacks achieve 88% and 94% success rate against the JPEG defense. Even if multi-modal embeddings were certifiably robust to input perturbations (they are not), we show that certification may not prevent these attacks due to the tradeoff between insufficient robustness and excessive invariance.

The main technical innovations of this work are: (1) establishing that adversarial alignment in multi-modal embeddings can be made closer than any organic alignment regardless of input and target modality; (2) demonstrating that the scope of this attack is much broader than conventional adversarial examples because it is task-agnostic and affects all models based on the embedding, even those that do not accept the adversary's target modality as input; and (3) a new hybrid technique for query-based attacks on black-box embedding.

Significance of the attacks. Many modern ML-based systems operate on third-party content that is not created or controlled by their users. For example, generative ML systems analyze and summarize webpages and social media, ML-based personal assistants process emails and instant messages, and retrieval-augmented generation and recommender systems operate on databases of content from many users. Multi-modal embeddings extend input domains of these systems to new modalities without re-training or fine-tuning.

It is a core security principle that any third-party content is potentially adversarial and, therefore, cannot be trusted. Vulnerabilities described in this paper enable adversaries to craft multi-modal content that is interpreted differently by humans and ML systems. The resulting exploits can cause ML systems to retrieve or recommend adversary-manipulated content in response to certain queries, interpret it in a way favorable to the adversary, conceal dangerous or harmful content from ML-based detection tools, or else create an illusion of danger or harm where there is none (*e.g.*, see Fig. 10).

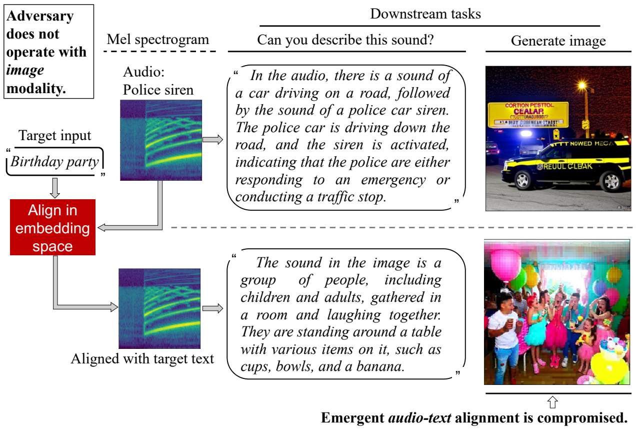
To facilitate research on the security of multi-modal embeddings, we released our code and models.[2]



" The image features a man wearing a suit and tie, standing in a group of people. He is the only one wearing a suit. "

" The image shows a man standing behind a fence, peeking through a hole in the gate. He is wearing a white shirt and black pants. "

Figure 3: **"Schadenfreude": a visual illusion against image and text generation.**



" In the audio, there is a sound of a car driving on a road, followed by the sound of a police car siren. The police car is driving down the road, and the siren is activated, indicating that the police are either responding to an emergency or conducting a traffic stop. "

" The sound in the image is a group of people, including children and adults, gathered in a room and laughing together. They are standing around a table with various items on it, such as cups, bowls, and a banana. "

Figure 4: **"Party time": an audio illusion against image and text generation.**

2 Background and Related Work

A machine learning (ML) model takes input $x \in \mathcal{X}$ and produces output $y \in \mathcal{Y}$ that can be, for example, a class label for x, a generated image, or generated text. Today, standard ML pipelines use an encoder θ (*e.g.*, CLIP [33]) to convert inputs into embedding vectors. Since training an encoder is expensive, many ML pipelines rely on pre-trained encoders (open-sourced or commercial). With a pre-trained encoder, it is sufficient to operate on embeddings, rather than raw inputs, in either a zero-shot fashion, or with a trained model. Many ML pipelines thus have two stages:

1. $\theta(x) = e$ — embedding of input
2. $\phi(e) = y$ — downstream task

Multi-modal embeddings. Multi-modal encoders map inputs $x^{(m)}$ from modalities $m \in \mathcal{M}$ into a single embedding

[2] https://github.com/ebagdasa/adversarial_illusions

space. Each modality m has its own encoder $\theta^{(m)}$. Encoders are trained on multi-modal tuples $(x^{(m)}, y^{(\bar{m})})$ that are semantically *aligned* (*e.g.*, a picture of wolves and the text "Wolves," as shown in Figure 5). Encoders are trained using contrastive learning [31] that pushes the representations of aligned inputs closer to each other in the embedding space and away from the representations of semantically different inputs.

Given a bi-modal dataset $\mathcal{D} = (X, Y) \in m \times \bar{m}$, we define an encoder's *alignment* on $\mathcal{D}$ as the mean cosine similarity between the embeddings of the elements in each $(x, y) \in \mathcal{D}$:

$$\text{align}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} \cos\left(\theta^{(m)}(x), \theta^{(\bar{m})}(y)\right) \qquad (1)$$

In this paper, we focus on images, sounds, and texts. As of this writing, ImageBind [10] and AudioCLIP [13] are the only open-source multi-modal embeddings that support these three modalities. Multi-modal embeddings exhibit "emergent" alignment between modalities. Semantically similar images and sounds (*e.g.*, a picture of a dog and an audio recording of a dog barking) have similar embeddings, even though the training data does not include image-audio tuples.

The distinction between natural (present in the training data) and emergent alignment is important for multi-modal encoders, but attacks in this paper are agnostic to the target modality and the source of its alignment. We generically define an encoder's *organic* alignment on datasets $\mathcal{D}$ where for all $(x, y) \in \mathcal{D}$, x and y are semantically related, regardless of whether it is learned from explicitly aligned training pairs or "emerged" from other modalities.

Downstream models. Multi-modal embeddings align semantically similar inputs across modalities. Thus, downstream models do not need to be trained on multiple modalities and can operate directly on the embeddings.

Zero-shot classification matches the embedding of an input to the closest class embedding (mean embedding of known class representatives). With multi-modal embeddings, zero-shot classification can classify inputs to classes in a different modality, *e.g.*, label an audio with an image class. *Retrieval*, similar to zero-classification, matches a query to the closest item in the embedding space.

Image generation takes an embedding and performs conditional generation using a diffusion model. ImageBind's image encoder is initialized from the CLIP visual encoder [33], thus diffusion models that operate on CLIP embeddings (*e.g.*, unCLIP [35]) can also operate on ImageBind embeddings. *Text generation* can use multi-modal embeddings as inputs to instruction-following language models, *e.g.*, PandaGPT [43].

Adversarial perturbations. An adversarial perturbation δ added to input x causes a model ϕ to produce an incorrect output, as in untargeted adversarial examples [11], or even a specific output y_t chosen by the adversary [4].

Adversarial examples for text target toxicity classification [15] and reading comprehension [19]. Recently, jailbreaking attacks on dialog systems, *i.e.*, chatbots, were demonstrated to bypass internal guardrails on hate speech [52]. All of these attacks target a specific downstream task.

Several recent papers [23, 49, 51] deploy adversarial perturbations against encoders that use contrastive learning. These attacks work by increasing contrastive loss and are necessarily *untargeted*: they prevent perturbed inputs from being embedded near the original input, but the adversary has no control over the placement or alignment of perturbed inputs.

In other recent work [2, 3, 32], adversarial perturbations are used for jailbreaking and prompt injection in multi-modal chatbots (a specific downstream task). In [39], adversarial perturbations against CLIP image embeddings are shown to affect downstream tasks. The attack is not cross-modal, the adversary has no arbitrary choice of inputs, and perturbations are very large and visible.

Dong et al. [6] present an untargeted attack on image embeddings and a targeted attack on text descriptions in a specific language model (rather than multi-modal embeddings). The attack causes the model to predict the main object in an image incorrectly. There is no analysis of how an embedding attack transfers to multiple downstream tasks, modalities beyond images and text, or robustness of defenses.

Zhao et al. [50] present several attacks against visual chatbots, evaluated via the embedding alignment (CLIP Score) between the adversarial image x_δ and target y_t. This is insufficient to show transferability to downstream tasks. For example, to attack a generation model ϕ based on embedding θ, it is not enough to show that $\theta(x_\delta)$ and $\theta(y_t)$ are aligned. The attack succeeds only if $\phi(\theta(x_\delta))$—*i.e.*, the image *generated* from the embedding, rather than the image used to produce the embedding—is classified to y_t. Modalities are limited to images and text, transferability across encoders is limited to CLIP as the target, and there is no analysis of defenses, even though adversarial images are foiled by simple countermeasures such as applying JPEG (see Section 6.1).

Our paper demonstrates that adversarial perturbations enable cross-modal, downstream task-agnostic adversarial alignment of arbitrary inputs in multi-modal embeddings.

Other attacks. Previous work considered collisions in NLP models [42], but—unlike this paper—it targets specific downstream tasks in a single modality (text), and the adversary does not have an arbitrary choice of colliding inputs.

Poisoning and backdoor attacks [17, 18] compromise embeddings at *training time*. By contrast, this paper focuses on *evaluation-time* attacks against "clean," unmodified models.

3 Threat Model

In this section, we briefly explain the adversary's goals and capabilities for different attack scenarios.

3.1 Adversary's Goals

The adversary seeks to influence a multi-modal embedding—and thus any downstream task or, in general, machine

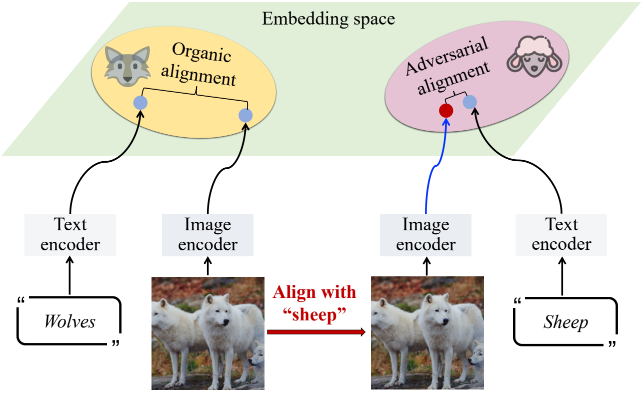


Figure 5: **Cross-modal, adversarial alignment in the embedding space.**

learning pipeline that uses the embedding, even if not known to the adversary at the time of the attack—by crafting a semantically meaningful input we call an *adversarial illusion* so that its embedding is aligned with another, attacker-chosen input in a different modality. Technically, the adversary's goal is to maximize *adversarial alignment* between the adversary's inputs x_δ in modality m and (arbitrarily chosen) targets y_t in modality $\bar{m}$—see Figure 5.

Cross-modal adversarial alignment extends the scope of the attack beyond the modalities supported by a particular ML pipeline. For example, there may not even exist an input in the right modality, *e.g.*, there may not be an available image corresponding to the adversary's target text (see an example in Figure 2). Furthermore, the attack may target downstream applications that do not accept the adversary's modality as input. For example, consider an accessibility application that takes images and generates audio describing their contents. An adversarial alignment attack enables an adversary to use any text of his choice as the target and attack this application with an image aligned with the chosen text—even though the application does not accept text inputs.

3.2 Adversary's Capabilities

We consider four distinct attack settings. In each setting, the adversary has different levels of access to the target encoders and auxiliary information.

White-box. Several multi-modal encoders are available as open-sourced code. In this setting, the adversary has access to all of their internal details, including architecture, parameters, and gradients (but not downstream tasks). This level of access allows the adversary to backpropagate gradients from the loss functions defined in Section 4 directly to adversarial illusions, similar to conventional adversarial examples [11, 29].

Transfer. If the adversary does not have access to the target encoder (because it is proprietary or closed-source or charges too much for queries), he may use white-box access to other, *surrogate* encoders to create adversarial illusions. The success of the transfer attack depends on the similarity between the surrogate and target models (which may not be known to the adversary in advance) and how well the surrogate attack generalizes [11, 24].

In Section 4.2 and Section 5.3, we show that backpropagating gradients from multiple surrogates helps generate illusions that succeed against target encoders to which the adversary does not have access.

Query-based. Many commercial embeddings are only available as APIs. In this setting, the adversary has query access to the target encoder but no surrogates. The adversary has no visibility into the model being queried.

Most query-based attacks rely on estimating the gradients of the target model, require additional auxiliary information, do not work well for targeted attacks, or fail to scale to a large number of queries [45]. Instead, our attack optimizes over a gradient-free objective (Section 4.3) to create effective adversarial illusions (Section 5.4).

Hybrid. Finally, we investigate a realistic, combined threat model in which the adversary has access to white-box surrogates (*e.g.*, open-source encoders) and query-only access to the target (*e.g.*, commercial embeddings). The adversary still has no prior knowledge about the target but can "warm-start" the query-based search with adversarial illusions crafted to transfer across surrogate encoders.

In Section 4.4 and Section 5.4, we show that this strategy, starting with white-box access to surrogate encoders, yields a successful query-only attack on Amazon's Titan embeddings.

4 Crafting Cross-Modal Illusions

Let $m, \bar{m} \in \mathcal{M}$ be the modalities supported by a multi-modal encoder $\theta = \{\theta^{(i)}\}_{i \in \mathcal{M}}$. The adversary's goal is to generate a perturbation δ such that the embedding $\theta^{(m)}(x_\delta)$ of $x_\delta = x^{(m)} + \delta$ is close to a target $\theta^{(\bar{m})}(y_t^{(\bar{m})})$, by optimizing the following objective:

$$\arg\min_{\delta} \left[\mathcal{L}(x_\delta, y_t) = \mathcal{L}\left(x^{(m)} + \delta, y_t^{(\bar{m})}; \theta\right) \right] \quad (2)$$

In the rest of this section, we define the objective functions $\mathcal{L}$ for each threat model from Section 3.

4.1 White-Box Attack

Although ImageBind [10] and AudioCLIP [13] use dot product during training, modalities that are not naturally aligned (*e.g.*, audio and text in ImageBind) have different normalizations. Hence, for our white-box attack, we omit the norms, use cosine similarity, and minimize the following objective:

$$\mathcal{L}_{\mathrm{WB}}(x_\delta, y_t) = 1 - \cos\left(\theta^{(m)}(x_\delta), \theta^{(\bar{m})}(y_t)\right) \quad (3)$$

To optimize Equation (3), we iteratively update perturbation δ with Projected Gradient Descent (PGD) [29].

4.2 Transfer Attack

For the transfer setting, we adopt a variation on the ensemble-based approach of [27], extending our white-box attack to an ensemble of K surrogate models $\Theta = \{\theta_i\}_{i=1}^{K}$. Each model's weight is parameterized by $\lambda = [\lambda_1 \, ... \, \lambda_K]$ where $\lambda_i = \frac{1}{K}$. Concretely, we minimize the following objectives:

$$\mathcal{L}_{\mathrm{T}}(x_\delta, y_t) = \sum_{\theta\in\Theta} \lambda_i \mathcal{L}_{\mathrm{WB}}(x_\delta, y_t; \theta) \qquad (4)$$

As in the white-box setting, we use a variant of PGD, varying the target model in each iteration of training.

4.3 Query-Based Attack

In the black-box setting, the adversary does not have access to the target model's parameters. We adopt the Square attack [1], a score-based method that does not rely on local gradient information. When the adversary's goal is embedding alignment, *i.e.*, cosine distance between the respective embeddings of some input and a target, the attack can be made stronger if the adversary knows non-target labels $\mathcal{Y}nt$ (*e.g.*, in the case of image-text alignment, $\mathcal{Y}nt$ are labels that neither describe the image, nor are targets of the attack). For $f(x_\delta, y_t) = 1 - \mathcal{L}_{WB}(x_\delta, y_t)$ (the embedding cosine distance), we update δ with a randomized search scheme, minimizing

$$\mathcal{L}_{BB}(x_\delta, y_t) = -f(x_\delta, y_t) - \log\left(\sum_{y\in\mathcal{Y}nt} e^{f(x_\delta, y)}\right) \qquad (5)$$

This is a contrastive approach that maximizes proximity to a specific target while minimizing proximity to non-targets.

4.4 Hybrid Attack

In the hybrid attack setting, we "warm-start" (*i.e.*, initialize) a query-based attack with the illusions generated locally by the transfer attack. Even if they do not fully transfer to the target model, we observe that they are still closer in the embedding space to the targeted region than the original input [44].

5 Experimental Evaluation

In this section, we evaluate adversarial illusions under the four threat models from Section 3 on two standard multi-modal datasets, four embedding families, and four downstream tasks (the fifth task is used for illustration). For randomly paired, multi-modal tuples, in each of the threat models, our attack produces adversarial illusions that (1) are close to their "source" in the input space, (2) are close to their randomly chosen target in the embedding space, and (3) mislead multiple downstream tasks with a high success rate.

5.1 Setup

Models. We use four encoder families. For our white-box attacks, we use ImageBind [10] and AudioCLIP [13], popular open-source text, image, and audio encoders. For our transfer attacks, we use the OpenCLIP family (ViT and ResNet50) [5, 7, 14] as surrogate models. For our query-based attack, we use Amazon's proprietary, black-box Titan encoder.

Datasets. We evaluate our (`image`, `text`) and (`audio`, `text`) adversarial alignment on random, 100-datapoint subsets of the ImageNet [36], AudioSet [8], AudioCaps [22], and LLVIP [21] datasets. Before subsetting, we permute the datasets by randomly matching source inputs in one modality to target inputs in another modality. The adversary is thus asked to produce a perturbation for each source that aligns its embedding with a random target, which is strictly harder than an untargeted attack.

Following is a brief summary of each dataset:

- *ImageNet* is an image-classification dataset that naturally aligns images and WordNet text labels. Before evaluating on the downstream tasks, each label y is injected into the templates introduced in [33] (*i.e.*, $y \leftarrow$ "A photo of a $\{y\}$.")
- *AudioSet* is an audio-classification dataset that associates audio samples and WordNet text labels. Labels are embedded as-is. AudioCLIP is the only encoder trained on (`audio`, `text`) data, thus audio-text alignment is "emergent" for the other encoders.
- *AudioCaps* is an audio-retrieval dataset based on AudioSet that associates audio samples with text captions.
- *LLVIP* is designed for low-light vision tasks, comprising paired RGB and thermal images of street scenes. We use the annotations in the dataset to crop pedestrian and random bounding boxes and label the cropped images as "Person" or "Street."

Downstream tasks. We evaluate our attack on the following downstream tasks:

- *Zero-shot Classification: Images, Thermal images, Audio.* Given a set of labels $\mathcal{Y}$, input x is assigned $y \in \mathcal{Y}$ whose embedding is closest to the embedding of x. For LLVIP thermal images, Y is {"Person", "Street"}.
- *Retrieval: Audio.* Similar to classification, audio input x "retrieves" the closest caption y in the embedding space.
- *Generation: Images.* Images are generated from embeddings using BindDiffusion [26] and evaluated using an image classifier. Because available generative models sometimes fail to produce correct outputs even on the embeddings of clean images (the generated image does not match the semantics of the input image), we only evaluate our attack on sources for which generation produces correctly classified images. We further explain this limitation below.
- *Generation: Text.* Text is generated from embeddings using PandaGPT [43] and evaluated using OpenAI's ChatGPT as a text classifier, as described below.

Evaluation metrics. To evaluate the effect of our attacks, we measure raw adversarial alignment and, for each downstream task, the attack success rate with Top-1 (T-1) and Top-5 (T-5) accuracies. For most of our experiments, the error bars

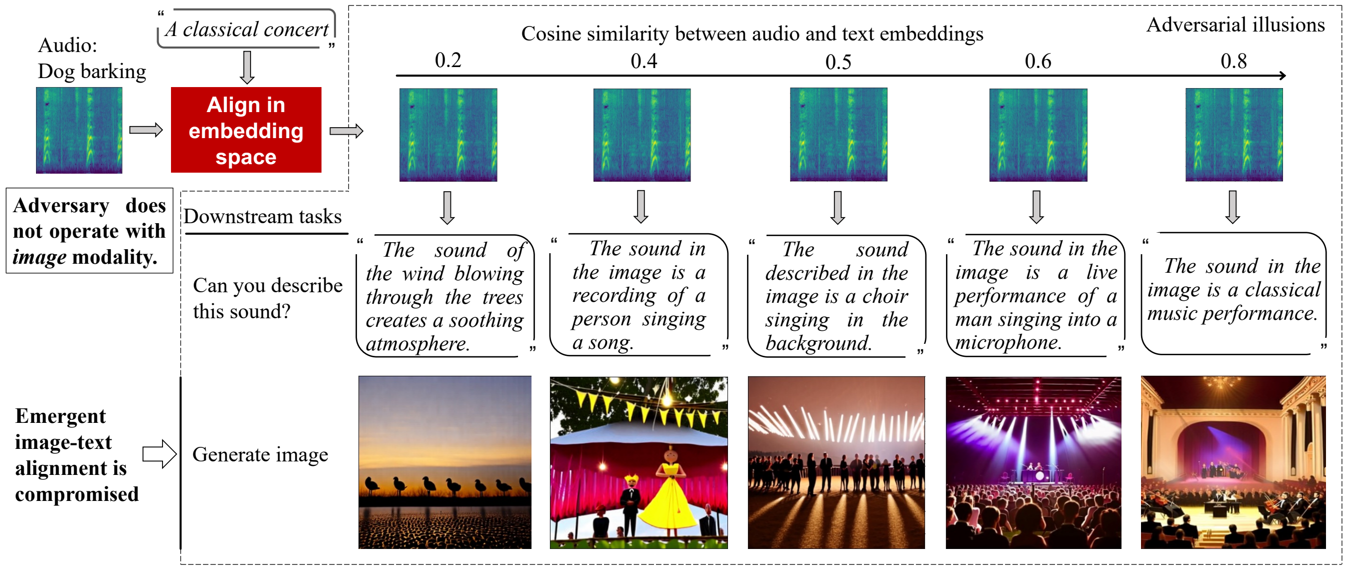

Audio:
Dog barking
" A classical concert "
Align in embedding space
Adversary does not operate with *image* modality.
Cosine similarity between audio and text embeddings
Adversarial illusions
0.2
0.4
0.5
0.6
0.8
Downstream tasks
Can you describe this sound?
" The sound of the wind blowing through the trees creates a soothing atmosphere. "
" The sound in the image is a recording of a person singing a song. "
" The sound described in the image is a choir singing in the background. "
" The sound in the image is a live performance of a man singing into a microphone. "
" The sound in the image is a classical music performance. "
Emergent image-text alignment is compromised
Generate image

Figure 6: **"Symphony of Woofs": similarity between an adversary-chosen input and the resulting illusion.**

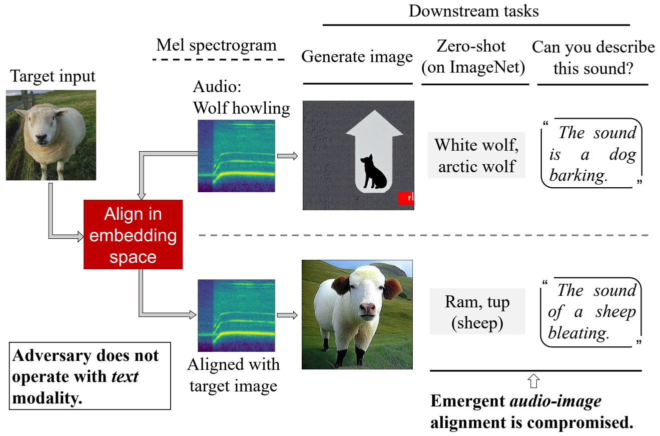

Downstream tasks
Mel spectrogram
Generate image
Zero-shot (on ImageNet)
Can you describe this sound?
Target input
Audio: Wolf howling
White wolf, arctic wolf
" The sound is a dog barking. "
Align in embedding space
Ram, tup (sheep)
" The sound of a sheep bleating. "
Adversary does not operate with *text* modality.
Aligned with target image
Emergent *audio-image* alignment is compromised.

Figure 7: **"Wolf in sheep's clothing": an audio illusion against image generation, zero-shot classification, and text generation.**

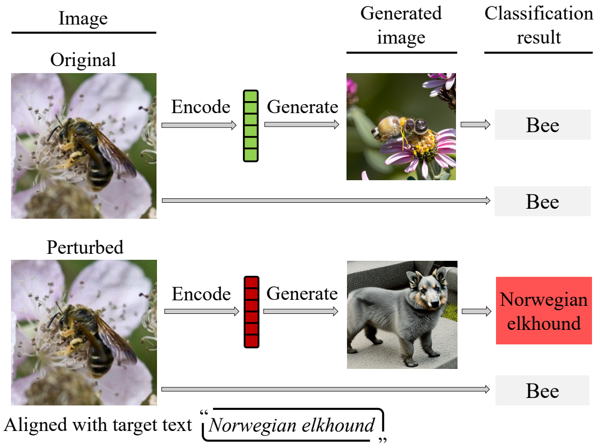


Figure 8: **"Beewulf": the original and perturbed images, images generated from their respective embeddings, and their classification.**

are standard deviations computed over a 100-sample subset, while for Amazon's Titan, we measure over 50 samples. As explained above, the classification accuracy of images, thermal images, and audio, as well as audio retrieval, is measured in a zero-shot fashion, while generated images are classified using a pre-trained vision Transformer (ViT-B/16) [7].

Limitations of downstream tasks. Generative models based on multi-modal embeddings are a new, rapidly evolving field, and publicly available implementations are limited. We only consider BindDiffusion [26] and PandaGPT [43]. BindDiffusion uses the unCLIP model [35] trained on CLIP embeddings [33]. Even without adversarial perturbations, images generated by BindDiffusion appears from the embeddings of many images, sounds, and texts are rather poor. PandaGPT, although trained on ImageBind embeddings, was fine-tuned only on image-text pairs. Therefore, in some cases, it interprets embeddings of sounds as if they were images (see Figure 4). Our attack is agnostic with respect to downstream models, and we expect that improvements in their quality will make the attack more effective.

Our evaluation of text generation from thermal images is, too, limited by the capabilities of PandaGPT, which does not

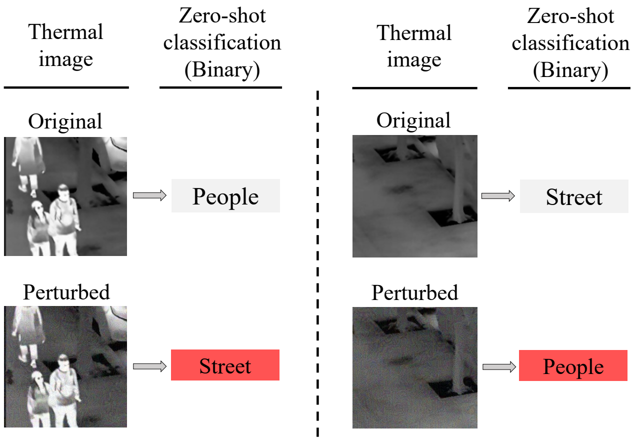
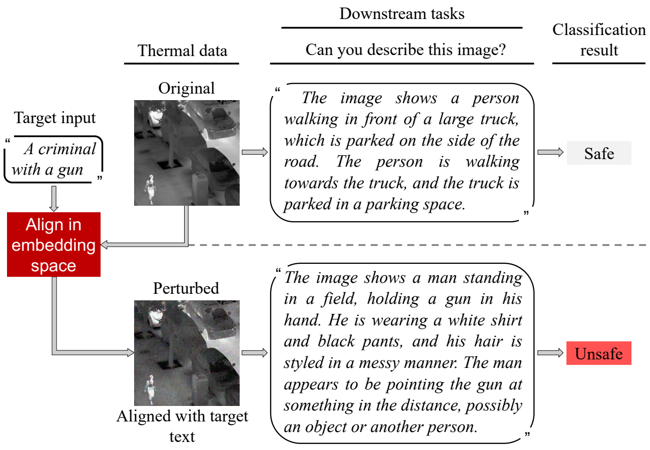


Figure 9: **"The invisible man": the original and perturbed thermal images and their zero-shot classification.**



" A criminal with a gun "

" The image shows a person walking in front of a large truck, which is parked on the side of the road. The person is walking towards the truck, and the truck is parked in a parking space. "

" The image shows a man standing in a field, holding a gun in his hand. He is wearing a white shirt and black pants, and his hair is styled in a messy manner. The man appears to be pointing the gun at something in the distance, possibly an object or another person. "

Figure 10: **"Surveillance": a thermal-image illusion against text generation.**

work well on ImageBind thermal embeddings. Therefore, while we use the thermal encoder for zero-shot classification, we use the RGB image encoder for generation. We align thermal images with the target text "A criminal with a gun" and generate text from the resulting embedding by querying PandaGPT "What's in the image?" To classify the generated text, we use the following prompt to ask GPT-4o whether the text describes an unsafe activity: "Answer with yes or no, does this response describe anything unsafe?"

Hyperparameters. For the white-box experiments, we present our results at various maximum perturbations: $\epsilon = \frac{\epsilon_V}{255}$ where $\epsilon_V \in \{1, 4, 8, 16, 32\}$ for images, $\epsilon = \epsilon_A \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$ for audio. Illusions are trained for $T = 7{,}500$ iterations in both cases. For the transfer experiments, we set the standard $\epsilon = \frac{16}{255}$ and evaluate for $T = 300$ iterations. For the query-based experiments (including the hybrid attack), we use the same hyperparameters as the Square attack [1] with the maximum perturbation $\epsilon = \frac{16}{255}$, query limit $N = 100{,}000$, and other hyperparameters unchanged.

5.2 White-Box Attack Results

As we demonstrate in the rest of this section, access to the gradients of the target encoder (ImageBind and AudioCLIP, in this case) yields a powerful attack based on a direct optimization of the adversarial illusion with respect to the desired alignment. In each of our experiments, we were able to generate illusions that are far closer in the embedding space to their target than even the target's "naturally aligned" counterpart. Consider an illusion x_δ and a target pair (x_t, y_t), say, an image of a dog and its caption. We achieve $\text{align}(\{x_\delta, y_t\}) > \text{align}(\{x_t, y_t\})$.

Figures 2 to 4, 6 and 7 illustrate the effect of our illusions on text and image generation. Figures 4 and 7 show audio illusions against generation and zero-shot classification, respectively. Figures 9 and 10 show thermal-image illusions against zero-shot classification and text generation, respectively. Figure 6 shows how increasing alignment between an audio perturbation and the adversary's text affects downstream tasks. As cosine similarity between the embeddings increases, the interpretation of the perturbed barking-dog audio changes until it is interpreted as a classical concert by downstream tasks. Figure 8 shows how an image generated from the embedding of an adversarial illusion is classified to the adversary-chosen text by an image classifier.

Zero-shot image classification. Table 1 shows the results for ImageBind and AudioCLIP. Whereas organic alignment is around 0.29 and 0.14, respectively, our attack with pixel perturbations as small as $\frac{1}{255}$ produces adversarial alignment that's twice as strong at 0.57 and 0.37. The corresponding Top-1 accuracy is a near-perfect 93% and 90%, respectively. Meanwhile, at the standard perturbation bound from the literature, $\frac{16}{255}$, our perturbed images induce adversarial alignment of 0.96 and 0.78 and are classified as the target label with 100% accuracy for both models.

Image generation. Table 2 shows that, with a maximum pixel perturbation of $\frac{16}{255}$, images generated from the embeddings of our illusions reach 64% and 92% attack success rate (*i.e.*, classification to the target label), *better* than the classification accuracy of images generated from the embeddings of the original images (42% and 64%, respectively).

Classification accuracy of the original images is 85% and 99%, respectively. By contrast, almost none of the adversarial images are classified to the label of the original image from which it was produced.

Text generation. Table 3 shows that with a maximum pixel perturbation of $\frac{16}{255}$, 64% of texts generated from the embeddings of our illusions are classified as unsafe by GPT-4o, vs. 0% for texts generated from the embeddings of original images.

Table 1: **Zero-shot image classification.** Our illusions successfully fool zero-shot image classification. We measure attack success rates (%) and adversarial alignment of cross-modal illusions against ImageBind and AudioClip. ϵ_V is the perturbation bound $\epsilon = \frac{\epsilon_V}{255}$. The first two rows are baselines (no perturbations applied to inputs). Standard deviations are reported.

Alignment		ImageBind		AudioCLIP	
	ϵ_V	Top-1	align(·)	Top-1	align(·)
Organic	–	67%	0.2878 ± 0.05	27%	0.1445 ± 0.03
Adversarial	–	0%	0.0892 ± 0.05	1%	0.0670 ± 0.03
	1	93%	0.5741 ± 0.15	90%	0.3560 ± 0.08
	4	100%	0.8684 ± 0.06	100%	0.6510 ± 0.09
	8	100%	0.9241 ± 0.04	100%	0.7323 ± 0.07
	16	100%	0.9554 ± 0.02	100%	0.7841 ± 0.01
	32	100%	0.9692 ± 0.01	100%	0.8074 ± 0.06

Table 2: **Image generation.** Our illusions successfully fool the downstream classification of generated images. We use BindDiffusion to generate images from ImageBind embeddings and measure Top-k attack success rate (%)—marked in bold—for the original images, adversarial illusions (white-box), and images generated from their respective embeddings.

Input type	Top-1		Top-5	
	Original label	Target label	Original label	Target label
Original image x	85%	0%	99%	0%
Adversarial illusion x_δ	77%	0%	95%	0%
BindDiffusion(ImageBind(x))	42%	0%	64%	2%
BindDiffusion(ImageBind(x_δ))	0%	**64**%	1%	**92**%

Table 3: **Text generation.** Our illusions successfully fool downstream classification of generated text. We use PandaGPT to generate texts from the ImageBind embeddings and measure attack success rate (%)—marked in bold—on texts generated from the original thermal images and the corresponding adversarial illusions (white-box).

	Safe	Unsafe
PandaGPT(ImageBind(x))	1.00	0.00
PandaGPT(ImageBind(x_δ))	0.32	**0.68**

Zero-shot audio classification and audio retrieval. Tables 4 and 5 demonstrate that our illusions are effective at attacking audio retrieval and classification. For ImageBind, audio data is represented as MEL spectrograms, where decibel levels serve as audio analogs of pixels. Meanwhile, for AudioCLIP, we use time series data of the same structure as in [13]. The optimization procedure is the same as in Section 4.1.

On AudioCaps, adversarial alignment significantly outperforms the organic baseline with 0.01 input perturbation. With 0.05 or higher perturbation, audio illusions have high adversarial alignment and near-perfect attack success rate on top-1 accuracy of both retrieval and classification. In ImageBind, this alignment is "emergent" because no (`audio`, `text`) data was used for training the ImageBind encoders.

Zero-shot thermal image classification. Table 6 shows that our illusions achieve perfect Top-1 accuracy with perturbations of at least $\frac{8}{255}$: all perturbed thermal images of people are classified as "street" and all perturbed thermal images of street are classified as "people."

For all perturbation bounds, adversarial alignment for thermal images is lower than for RGB images. We have two conjectures for this discrepancy. First, initial adversarial alignment (without any perturbation) is lower for thermal images than for RGB images, suggesting that thermal images are less susceptible to adversarial perturbations. Second, thermal images have only one channel vs. three channels in RGB images. Consequently, there are $2 \times 255 \times 255$ fewer pixels available for manipulation in thermal images.

Table 4: **Zero-shot audio classification.** Our illusions successfully fool zero-shot audio classification. We measure Top-k attack success rates (%) and adversarial alignment of cross-modal illusions against ImageBind and AudioClip. ϵ_A is the perturbation bound. The first two rows are baselines (no perturbations applied to inputs). Standard deviations are reported.

Alignment		ImageBind			AudioCLIP		
	ϵ_A	Top-1	Top-5	align(·)	Top-1	Top-5	align(·)
Organic	-	23%	54%	0.1087 ± 0.1	-	-	-
Adversarial	-	1%	6%	0.0112 ± 0.05	0%	1%	0.0510 ± 0.03
	0.005	13%	34%	0.1090 ± 0.07	93%	99%	0.3441 ± 0.08
	0.010	67%	95%	0.2772 ± 0.10	98%	100%	0.3622 ± 0.07
	0.050	100%	100%	0.7755 ± 0.06	99%	100%	0.3983 ± 0.06
	0.100	100%	100%	0.8656 ± 0.05	99%	100%	0.4060 ± 0.06
	0.500	100%	100%	0.9130 ± 0.04	99%	100%	0.4139 ± 0.06

Table 5: **Zero-shot audio retrieval.** Our illusions successfully fool zero-shot audio retrieval. We measure Top-k attack success rates (%) and adversarial alignment of cross-modal illusions against ImageBind and AudioClip audio retrieval. ϵ_A is the perturbation bound. The first two rows are baselines (no perturbations applied to inputs). Standard deviations are reported.

Alignment		ImageBind			AudioCLIP		
	ϵ_A	Top-1	Top-5	align(·)	Top-1	Top-5	align(·)
Organic	-	13%	32%	0.2141 ± 0.11	-	-	-
Adversarial	-	0%	0%	0.0248 ± 0.07	0%	1%	0.0827 ± 0.03
	0.005	1%	10%	0.1263 ± 0.09	90%	100%	0.4141 ± 0.05
	0.010	48%	68%	0.3319 ± 0.12	95%	100%	0.4419 ± 0.05
	0.050	99%	100%	0.8641 ± 0.06	99%	100%	0.4733 ± 0.05
	0.100	99%	100%	0.9295 ± 0.05	99%	100%	0.4796 ± 0.05
	0.500	99%	100%	0.9578 ± 0.04	99%	100%	0.4870 ± 0.05

5.3 Transfer Attack Results

We now show that access to the target encoder's gradients is not necessary to create adversarial illusions. It may be sufficient to perform a white-box attack on a surrogate encoder and "transfer" the resulting illusions to the target. Table 7 shows that the transfer attack can achieve high zero-shot image classification accuracy, although results depend on the similarity between the surrogate and the target.

Both ImageBind and AudioCLIP add modalities to, respectively, pre-trained ViT and ResNet-50 versions of the popular CLIP embedding, albeit in different ways. In ImageBind, all modalities are mapped into the ViT-based CLIP's embedding space. AudioCLIP fine-tunes ResNet-50-based CLIP on audiovisual data, thereby "merging" the audio, image, and text embedding spaces. This difference seems to matter. In our experiments, we found that illusions did not transfer to a checkpoint at the end of AudioCLIP's fine-tuning process but did transfer to a partially trained version provided for compatibility with CLIP (and ostensibly close to CLIP's embedding space). Table 7 includes this partially trained checkpoint.

With a single surrogate model, we observed that illusions transfer well within but not across model architectures. Even so, as we show in the bottom row of Table 7, this generalizability gap can be addressed by jointly optimizing a single illusion across multiple surrogates (as in Section 4.2). We measured the transfer of illusions generated using both OpenCLIP models as surrogates and found that the same illusions achieve high adversarial alignment in both ImageBind and AudioCLIP's embedding spaces (see Figure 11). The existence of illusions that attack multiple embedding spaces simultaneously indicates the possibility of *universal* illusions. We leave this line of research to future work.

Table 6: **Zero-shot thermal image classification.** Our illusions successfully fool zero-shot thermal image classification. We measure attack success rates (%) and adversarial alignment of cross-modal illusions against ImageBind. Perturbation bounds are $\frac{\epsilon_V}{255}$. The first two rows are baselines (no perturbation). Standard deviations are reported.

Alignment		ImageBind	
	ϵ_V	Top-1	align(·)
Organic	–	68%	0.0757 ± 0.04
Adversarial	–	32%	0.0534 ± 0.02
	1	39%	0.0679 ± 0.03
	4	87%	0.1651 ± 0.03
	8	100%	0.2841 ± 0.03
	16	100%	0.4210 ± 0.03
	32	100%	0.5576 ± 0.03

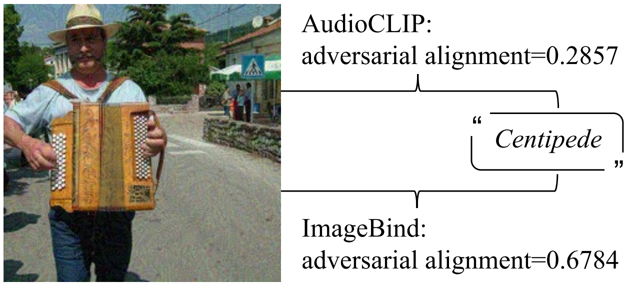


Figure 11: **An example of a transferable illusion that misleads both ImageBind and AudioCLIP embeddings.**

5.4 Query-Based and Hybrid Attack Results

In this section, we show that the adversary may not need access to gradients or surrogates at all as long as he has query access to the target encoder. With a limited number of queries, the adversary can craft adversarial illusions that succeed against downstream zero-shot classification and generation tasks with high accuracy. Because commercial black-box embeddings charge for each use of their APIs, we aim to minimize the number of queries and discontinue the experiments (other than those with ImageBind-full) when adversarial alignment is sufficient to fool zero-shot classification.

Zero-shot image classification. Table 8 shows that our query-based attack achieves near-perfect success rates against the black-box versions of ImageBind and AudioCLIP (98% and 100%, respectively). Observe that adversarial alignment (0.26 and 0.16) is similar to the models' organic alignment (0.29 and 0.14). This is sufficient to fool zero-shot classification.

Image generation. While zero-shot classification is compromised even by relatively weak adversarial alignment (comparable to organic alignment), we observed that higher adversarial alignment is necessary for the downstream generation model to generate images that classify to a label corresponding to the adversary's target. Concretely, our early-stopped ImageBind illusions achieved negligible accuracy on the downstream generation task. When we increased the number of queries, adversarial alignment increased to 0.37, and adversarial accuracy of downstream generation improved to 38%.

Amazon's Titan. Finally, we apply our attack to Amazon's commercial, black-box Titan embedding. Table 8 shows that, without the warm start, the query-based attack achieves a 30% success rate against zero-shot classification. With the hybrid attack, 3% of the illusions transfer directly, and when coupled with the query-based attack, the success rate reaches 42%. This suggests that hybrid illusions start closer to the targets in the embedding space, making black-box optimization easier. We do not know the internal details of Titan embeddings but conjecture, based on these results, that either these embeddings are at least partly based on the CLIP, or our hybrid attack may generalize beyond related embedding spaces.

Amazon currently charges \$0.00006 per image query to the Titan API, so the average cost is \$125.51 per 100 images for the query-based attack and \$108.11 per 100 images for the hybrid attack. For comparison, Google's Gemini Pro embedding currently charges \$0.0025 per image query.

6 Countermeasures

In this section, we survey several types of defenses against adversarial perturbations and analyze whether they can protect multi-modal embeddings.

6.1 Feature Distillation

Adversarial input perturbations can be considered as features for the model [16]. One defense is to apply a transformation that preserves essential visual features while destroying adversarial features, *e.g.*, JPEG compression [28]. In our case, adding a JPEG compression layer before encoding input images lowers zero-shot classification accuracy of illusions from 100% to 5% and 27% for ImageBind [10] and AudioCLIP [13], respectively.

Unfortunately, this defense can be evaded. After we add a differentiable approximation [40] of JPEG compression to our perturbation method, the resulting illusions are effective against both original (72% and 98% zero-shot classification accuracy for ImageBind and AudioCLIP, respectively) and JPEG-protected embeddings (88% and 94%)—see Table 9.

Defense-GAN [37] is a similar defense that uses a separate GAN to encode each input and generate a supposedly equivalent replacement. This defense assumes that (1) the adversary cannot craft adversarial perturbations for both the original encoder and the GAN, and (2) the GAN accurately re-generates

Table 7: **Transfer attack.** On related embedding spaces, our illusions transfer. We measure attack success rates (%) of our transfer attack with different surrogate and target models. "Ours" represents an ensemble of surrogates consisting of the ViT and ResNet versions of OpenCLIP. $\epsilon_V = \frac{16}{255}$.

Surrogate Model	Target model			
	ImageBind	OpenCLIP-ViT	AudioCLIP	OpenCLIP-RN50
ImageBind	-	100%	0%	0%
OpenCLIP-ViT	100%	-	1%	0%
AudioCLIP	0%	0%	-	92%
OpenCLIP-RN50	0%	0%	88%	-
Ours	**100%**	**100%**	**90%**	**100%**

Table 8: **Query-based and hybrid attack results.** Our black-box illusions work on a number of encoders. We measure organic alignment, adversarial alignment, zero-shot classification (ZSC) attack success rate (%), average number of queries, and Top-k image generation (Gen) attack success rate (%) for different encoders. ImageBind-full is run for a full 100,000 queries (no early stopping). Titan-hybrid is our hybrid attack on Amazon's Titan encoder. $\epsilon_V = \frac{16}{255}$. Standard deviations are reported.

Encoder	Org. align(·)	Adv. align(·)	ZSC	Avg. queries	Gen (Top-1)	Gen (Top-5)
ImageBind	0.2905 ± 0.06	0.2609 ± 0.04	98%	18,942	0	3%
ImageBind-full	0.2905 ± 0.06	0.3659 ± 0.07	98%	100,000	**38%**	**58%**
AudioCLIP	0.1406 ± 0.04	0.1674 ± 0.03	100%	4,112	-	-
Titan	0.4652 ± 0.04	0.4024 ± 0.07	30%	20,919	-	-
Titan-hybrid	0.4652 ± 0.04	0.4215 ± 0.06	**42%**	**18,019**	-	-

all features of the original inputs that are important for downstream tasks. The efficacy of this defense is uncertain because it was only evaluated on the MNIST dataset. Furthermore, it lacks versatility across modalities and is not task-agnostic.

6.2 Anomaly Detection

Semantically similar inputs should be embedded into similar representations. Therefore, a plausible defense is to compare the embedding of an input with the embeddings of very similar inputs, *e.g.*, various augmentations. Table 10 shows that this defense is effective: average alignment between unmodified images and their augmentations is 0.8828, significantly higher than 0.3725 average alignment between the corresponding adversarial illusions and their augmentations.

The evasion attack from Section 6.1 defeats this defense, too: observe that the values in the x_{jpeg} column overlap with those in the x column. This illustrates the limitations of anomaly detection methods against defense-aware adversaries [41]. Anomaly detection using high-dimensional statistics or enforcing a specific geometry for each modality are interesting topics for future work.

6.3 Certified Robustness

Robustness to perturbations. Another plausible defense is to somehow train multi-modal encoders so that they are robust in the following sense: if two inputs are within δ of each other, their respective embeddings should be within γ (using some suitable metrics for δ and γ over the input space and embedding space, respectively).

Certified robustness [12, 34] ensures that any input perturbation within a certain bound has only a limited impact on the output of the model. It is defined for classification tasks and cannot be directly applied to multi-modal encoders in a downstream task-agnostic way. Given a specific downstream task, it may be possible to perform interval-bound propagation [12] over multi-modal encoders.

Unfortunately, it is unclear which metric to use for δ. For many plausible metrics, there is a fundamental tradeoff between robustness and invariance [46]. If δ is too small, the adversary can use perturbations bigger than δ to evade the defense (*i.e.*, the model is insufficiently robust). If δ is too big, semantically unrelated inputs will be mistakenly aligned (*i.e.*, the model is too invariant).

This observation was made for l_p bounds and image clas-

Table 9: **JPEG-resistant illusions.** Our JPEG-resistant illusions evade JPEG compression defenses. We measure Top-k attack success rates (%) and adversarial alignment against zero-shot classification for the original and JPEG-resistant adversarial illusions. Illusions are generated with 200 iterations to avoid over-fitting. $\epsilon_V = \frac{16}{255}$. Standard deviations are reported.

Alignment	ImageBind		AudioCLIP	
	Top-1	align(·)	Top-1	align(·)
Adversarial illusion x_δ	100%	0.7679 ± 0.07	100%	0.5048 ± 0.04
JPEG(x_δ)	5%	0.1722 ± 0.06	27%	0.0722 ± 0.03
JPEG-resistant adversarial illusion x_{jpeg}	72%	0.3623 ± 0.08	98%	0.2345 ± 0.04
JPEG(x_{jpeg})	**88**%	0.3968 ± 0.07	**94**%	0.2079 ± 0.04

Table 10: **Generating illusions that evade anomaly detection.** Alignment between augmentations and unperturbed inputs(x), adversarial illusions (x_δ), and JPEG-resistant illusions (x_{jpeg}). Numbers are computed over 100 random ImageNet images. $\epsilon_V = \frac{16}{255}$. Standard deviations are reported.

Augmentation method	ImageBind			AudioCLIP		
	x	x_δ	x_{jpeg}	x	x_δ	x_{jpeg}
JPEG()	0.8784 ± 0.04	0.2257 ± 0.07	0.7303 ± 0.08	0.8512 ± 0.06	0.2149 ± 0.08	0.6198 ± 0.09
GaussianBlur()	0.7493 ± 0.09	0.2108 ± 0.07	0.6447 ± 0.09	0.8068 ± 0.08	0.2070 ± 0.07	0.5637 ± 0.09
RandomAffine()	0.8698 ± 0.06	0.2557 ± 0.10	0.7428 ± 0.09	0.8756 ± 0.06	0.2124 ± 0.11	0.6175 ± 0.10
ColorJitter()	0.8903 ± 0.07	0.4987 ± 0.19	0.8245 ± 0.10	0.9197 ± 0.05	0.4248 ± 0.19	0.7486 ± 0.12
RandomHorizontalFlip()	0.9713 ± 0.05	0.4006 ± 0.09	0.8670 ± 0.07	0.9688 ± 0.02	0.2949 ± 0.07	0.7070 ± 0.07
RandomPerspective()	0.9377 ± 0.07	0.6437 ± 0.38	0.8739 ± 0.15	0.9398 ± 0.07	0.6161 ± 0.41	0.8105 ± 0.21
Average	0.8828 ± 0.07	0.3725 ± 0.14	0.7805 ± 0.09	0.8937 ± 0.06	0.3284 ± 0.15	0.6779 ± 0.11

sifiers in [46]. We extend it to l_∞ bounds, showing (by counterexample) that no δ is sufficiently robust *and* does not mistakenly align semantically different inputs.

Figure 12 shows that δ must be greater than 0.3 because there exist perturbations bigger than 0.3 that produce a visually similar image, which should be aligned with the original. This figure also shows that δ must be less than 0.3 because there exist perturbations smaller than 0.3 that produce a visually different image, which should *not* be aligned with the original. Thus, there is no "right" l_∞ value for δ.

Robustness to semantics-preserving transformations. Ultimately, embeddings should be trained so that *semantically* close inputs, within and across modalities, are encoded into similar vectors in the embedding space. There is prior work on adversarially robust "perceptual similarity" metrics [9]. Perceptual similarity only captures semantic similarity insofar as images belong to the same class, *i.e.*, it only aligns images with the labels of a specific image classifier. Alignment in multi-modal embeddings is supposed to cover much broader semantic similarity.

For text, there exist semantics-preserving transformations (*e.g.*, substituting words with synonyms), and models can be trained to be robust with respect to these transformations [20, 30, 48]. Our attacks exploit perturbations in images and audio, where semantic similarity is not as well defined as in text. In the image domain, robustness under different distance metrics results in vastly different performance [47]. We are not aware of any metric on the input space that corresponds to task-agnostic semantic similarity, which is required to achieve certified robustness for embeddings.

Furthermore, to ensure that the embeddings of semantically similar inputs are close to each other, the range of encoders for different modalities should overlap with very fine-grained control over changes in the embedding given a change in the input. With such control, it is unclear how to prevent an adversary from crafting inputs closer to the target than organically aligned inputs, considering that organic alignment in the existing embeddings is relatively weak (see Section 5.2).

In practice, it does not appear that any existing multi-modal embedding uses certified robustness.

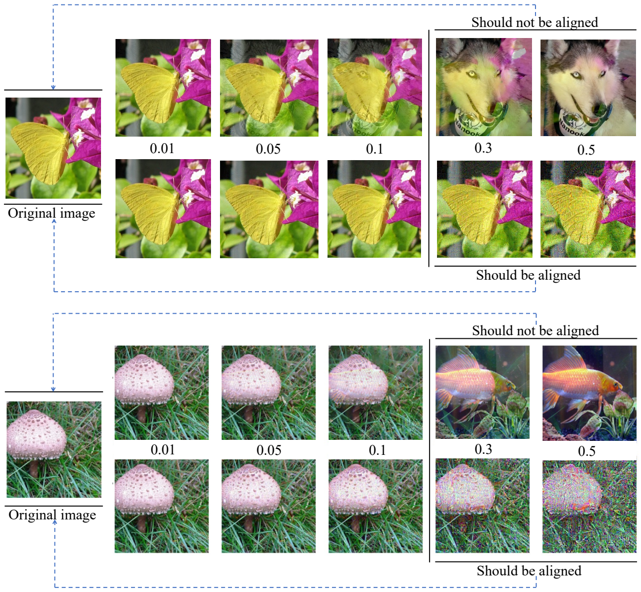


Figure 12: **Perturbations of the same magnitude can produce semantically different images or adversarial illusions.**

6.4 Adversarial Training

When training classifiers, one way to ensure that small perturbations do not affect the model's output is to add perturbed images with correct class labels to the training data [29, 38]. In contrastive learning, adversarial training can use perturbed images as positive samples [23, 49, 51]. The efficacy of this defense depends on the specific downstream task because some tasks (*e.g.*, fine-grained classification) may require distinguishing representations that differ less than γ (maximum embedding distance between an input and its perturbations). Adversarial contrastive training negatively impacts relatively coarse downstream tasks such as CIFAR-10 and -100 [49]. In multi-modal contrastive learning, the data are sparse, and some modalities are not directly aligned. Thus, we expect that adversarially trained embeddings significantly degrade the accuracy of downstream tasks.

7 Conclusions

We demonstrated that aligned multi-modal embeddings are highly vulnerable to adversarial perturbations that create *cross-modal illusions*, *i.e.*, inputs in one modality that are aligned with semantically unrelated inputs in another modality. The attack works across both natural and emergent alignment and is task-agnostic: adversarial alignment fools even downstream tasks of which the adversary is not aware when generating adversarial inputs. Furthermore, it enables the adversary to use alignment targets in modalities that are not accepted by downstream applications.

We evaluated the attack for several multi-modal embeddings, including open-source embeddings such as ImageBind and AudioCLIP and proprietary, black-box embeddings such as Amazon's Titan. We showed that the attack successfully misleads downstream tasks based on multi-modal embeddings, including audio retrieval, zero-shot classification of images, thermal images, and audio, and generation of text and images. Finally, we discussed potential countermeasures and evasion attacks.

Ethical Considerations

The only purpose of analyzing the vulnerability of multi-modal embeddings to adversarial inputs is to help develop more robust embeddings and motivate research on defenses.

Acknowledgments

This work was performed at Cornell Tech and partially supported by the NSF grant 1916717.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square Attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision (ECCV)*, 2020.

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

[3] Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? *arXiv:2306.15447*, 2023.

[4] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *IEEE Symposium on Security and Privacy Workshops*, 2018.

[5] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[6] 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:2309.11751*, 2023.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas

Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Machine Learning (ICML)*, 2021.

[8] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[9] Sara Ghazanfari, Siddharth Garg, Prashanth Krishnamurthy, Farshad Khorrami, and Alexandre Araujo. R-LPIPS: An adversarially robust perceptual similarity metric. *arXiv:2307.15157*, 2023.

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[12] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy A. Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[13] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending CLIP to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving Google's perspective API built for detecting toxic comments. *arXiv:1702.08138*, 2017.

[16] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[17] Rishi Dev Jha, Jonathan Hayase, and Sewoong Oh. Label poisoning is all you need. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[18] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. BadEncoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *IEEE Symposium on Security and Privacy (S*&P), 2022.

[19] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[20] Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[21] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. LLVIP: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[22] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[23] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[24] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Machine Learning (ICML) Workshops*, 2017.

[25] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[26] Zhijie Lin, Lijuan Liu, Yangzihao Wang, and Xiangyu Xu. BindDiffusion: One diffusion model to bind them all. https://github.com/sail-sg/BindDiffusion/tree/main, 2023.

[27] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations (ICLR)*, 2017.

[28] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: DNN-oriented JPEG compression against adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[30] Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations (ICLR)*, 2017.

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

[32] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *International Conference on Machine Learning (ICML) Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[34] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv:2204.06125*, 2022.

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[37] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations (ICLR)*, 2018.

[38] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[39] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Plug and Pray: Exploiting off-the-shelf components of multi-modal models. *arXiv:2307.14539*, 2023.

[40] Richard Shin and Dawn Song. JPEG-resistant adversarial images. In *NeurIPS Workshop on Machine Learning and Computer Security*, 2017.

[41] Chawin Sitawarin, Zachary Golan-Strieb, and David Wagner. Demystifying the adversarial robustness of random transformation defenses. In *AAAI Workshop on Adversarial Machine Learning and Beyond*, 2022.

[42] Congzheng Song, Alexander M Rush, and Vitaly Shmatikov. Adversarial semantic collisions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[43] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. PandaGPT: One model to instruction-follow them all. *arXiv:2305.16355*, 2023.

[44] Fnu Suya, Jianfeng Chi, David Evans, and Yuan Tian. Hybrid batch attacks: Finding black-box adversarial examples with limited queries. In *USENIX Security Symposium (USENIX Security)*, 2020.

[45] Fnu Suya, Anshuman Suri, Tingwei Zhang, Jingtao Hong, Yuan Tian, and David Evans. SoK: Pitfalls in evaluating black-box attacks. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024.

[46] Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *International Conference on Machine Learning (ICML)*, 2020.

[47] Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[48] Jin Yong Yoo and Yanjun Qi. Towards improving adversarial training of NLP models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[49] Qiying Yu, Jieming Lou, Xianyuan Zhan, Qizhang Li, Wangmeng Zuo, Yang Liu, and Jingjing Liu. Adversarial contrastive learning via asymmetric InfoNCE. In *European Conference on Computer Vision (ECCV)*, 2022.

[50] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[51] Ziqi Zhou, Shengshan Hu, Ruizhi Zhao, Qian Wang, Leo Yu Zhang, Junhui Hou, and Hai Jin. Downstream-agnostic adversarial examples. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[52] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv:2307.15043*, 2023.