AI

MDPI

Article

Understanding Modality-Specific Vulnerabilities in Vision–Language Models Under Adversarial Attacks

Maisha Binte Rashid [1,*] and Pablo Rivas [2,*]

[1] Department of Computer Science, Baylor University, Waco, TX 76798, USA
[2] Department of Computing Technology, Marist University, Poughkeepsie, NY 12601, USA
* Correspondence: maisha_rashid1@baylor.edu (M.B.R.); pablo.rivas@marist.edu (P.R.); Tel.: +1-(845)-575-3101 (P.R.)

Abstract

Vision–language models (VLMs), such as Contrastive Language–Image Pretraining (CLIP), are increasingly deployed in real-world applications, including content moderation, misinformation detection, and fraud analysis, making their robustness to adversarial attacks a critical concern. While adversarial robustness has been widely studied in unimodal models, modality-specific vulnerabilities in multimodal models remain underexplored. In this work, we analyze CLIP by applying gradient-based adversarial attacks to its vision and language modalities, both independently and jointly, and evaluating performance on two multimodal classification benchmarks: the Facebook Hateful Memes dataset and a large-scale Suspicious Car Parts dataset. Using Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks along with multiple adversarial retraining strategies, we show that adversarial perturbations on the image modality consistently cause the most severe and unstable performance degradation. These results demonstrate that the vision modality is the primary vulnerability in CLIP, highlighting the need for modality-specific defense strategies that focus more on the weaker modality in multimodal systems.

Keywords: vision-language models; adversarial attack; adversarial robustness

1. Introduction

Vision–language models (VLMs), such as Contrastive Language–Image Pretraining (CLIP) [1], integrate visual and textual modalities to perform complex tasks across safety-critical domains, including healthcare diagnostics, autonomous driving, and content moderation. As these models are increasingly used in high-risk areas such as detecting misinformation and identifying fraud, it is crucial to ensure they can withstand adversarial attacks and are not easily fooled. As these models are used more often in sensitive and important situations, people are becoming increasingly worried about how easily they can be tricked. Adversarial attacks use carefully crafted inputs to fool the model into making incorrect predictions, which can reduce trust in AI systems and make them unsafe for use in the real world. Protecting multimodal models is more difficult than protecting models that use only one type of data. This is because small changes to the image, the text, or both at the same time can affect how the model connects and understands information across modalities, which can affect its overall performance. Even so, there has not been much research to fully understand how these different parts of the model interact under adversarial attacks.

Most existing adversarial defense methods focus on models that use only one type of data, such as only images or only text. While these methods work well in those separate

Academic Editor: Peican Zhu

Received: 11 February 2026
Revised: 23 March 2026
Accepted: 3 April 2026
Published: 9 April 2026

domains, they do not consider that, in a multimodal model, the image and text parts may not be equally vulnerable [2–8]. Prior work on CLIP's adversarial robustness [9,10] has discussed its overall vulnerability but has not systematically compared the relative susceptibility of its vision and language encoders, nor determined which modality forms the weakest impact.

In this work, we address this gap by conducting a systematic empirical study of CLIP's adversarial robustness under controlled modality-specific attack scenarios. We apply gradient-based attacks to the vision modality, the language modality, and both modalities jointly, and evaluate performance across two multimodal classification benchmarks: the Facebook Hateful Memes dataset and a large-scale Suspicious Car Parts dataset constructed using a graph-based labeling methodology. To further isolate modality contributions, we introduce a selective freezing adversarial training protocol that alternates between freezing the vision and language encoders during retraining. This setup enables us to examine how each modality independently contributes to robustness and recovery under attack. Our focus is on the research question: *Which modality in CLIP is more adversarially vulnerable under structured perturbation systems?* To find an answer to this research question, this paper makes the following contributions:

- We present a detailed study that separately examines how the vision and language parts of CLIP respond to adversarial attacks. Specifically, we test each encoder independently using two commonly used adversarial attacks: Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) on two real-world multimodal datasets: the Facebook Hateful Memes dataset and a large-scale Suspicious Car Parts dataset.
- We demonstrate that the vision modality is the primary vulnerability in CLIP, consistently causing more performance degradation than text-only attacks, a finding that holds across both datasets and attack methods.
- We evaluate multiple adversarial retraining strategies involving selective encoder freezing, showing that modality-specific defenses focused on the vision encoder can achieve greater competitive robustness than full-model adversarial training.

Our results offer practical guidance on how to design smarter, modality-aware defenses that strengthen vision–language models without the high computational cost of retraining the entire model with adversarial data. This is especially important for real-world applications where computational resources are limited. By showing that the vision part of CLIP is more vulnerable and that focusing defenses on this encoder is effective, our work provides a clear path toward building more robust and trustworthy vision–language models that can better resist adversarial attacks in real-world use.

The remainder of this paper is organized as follows. Section 2 presents the problem statement along with the scope and assumptions of this study. Section 3 reviews related work on adversarial robustness in unimodal and multimodal models. Section 4 describes the datasets used, including the Hateful Memes and Suspicious Car Parts datasets. Section 5 details the proposed methodology, adversarial attack settings, and experimental configuration. Section 6 presents the experiments conducted. Section 7 discusses the experimental results and analysis. Section 8 discusses the findings, implications, and open research questions. Finally, Section 9 concludes the paper.

2. Problem Statement and Assumptions

2.1. Problem Definition

Let $f : \mathcal{X}_{\text{img}} \times \mathcal{X}_{\text{txt}} \to Y$ denote a multimodal classifier mapping image–text input pairs to a discrete label set Y. In this work, f is instantiated as CLIP (ViT-B/32) extended with a linear classification head, and $Y = \{0, 1\}$ for both binary classification tasks considered.

An adversarial attack on the image modality produces a perturbed image $x'_{\text{img}} = x_{\text{img}} + \delta_{\text{img}}$ such that $\|\delta_{\text{img}}\|_\infty \leq \epsilon$ and $f(x'_{\text{img}}, x_{\text{txt}}) \neq f(x_{\text{img}}, x_{\text{txt}})$. This definition applies to text-modality attacks operating in the continuous embedding space. The main question of this work is: *Which modality in CLIP is more adversarially vulnerable under structured perturbation systems?*

2.2. Scope, Limitations, and Assumptions

We analyze the research question under the following assumptions:

- The analysis assumes a white-box threat model, in which the attacker has access to model gradients and architecture during adversarial example generation. Under this assumption, only gradient-based attacks (FGSM and PGD) are evaluated. Combinatorial text-space attacks (e.g., word substitution) are excluded because they are incompatible with the continuous embedding-space perturbation framework applied uniformly across both modalities.
- The experiments are limited to a single CLIP architecture (ViT-B/32) and may not generalize directly to all vision–language models.
- The classification tasks used in this study represent specific multimodal applications, but they may not reflect all types of multimodal interactions that appear in other tasks, such as visual question answering or image caption generation.

Within these constraints, the study aims to provide insights into how adversarial perturbations affect different modalities in CLIP and whether robustness improvements can be achieved by focusing defensive strategies on the most vulnerable component.

3. Related Work

AI-driven secure architectures, and the robustness of intelligent systems in adversarial environments, are a growing interest [11]. In recent years, the field of adversarial machine learning, in particular, has gained significant attention, with researchers focusing on understanding and mitigating adversarial attacks that target machine learning models, particularly in areas such as image and text classification. Adversarial attacks involve the manipulation of input data in a way that can mislead models into producing incorrect outputs, posing serious risks to applications in security-sensitive domains [12,13].

3.1. Literature Review Method

The literature review was conducted through a structured search of major digital libraries, including IEEE Xplore, ACM Digital Library, arXiv, and Google Scholar. Keywords used in the search included **adversarial machine learning**, **adversarial attacks**, **vision–language models**, **adversarial robustness**, and **CLIP**. Studies were selected based on relevance to adversarial robustness in image models, text models, and multimodal architectures. Priority was given to recent publications proposing attack methods, defense strategies, and robustness analyses for vision–language models. The selected literature was then categorized into three groups: adversarial attacks on images, adversarial attacks on text, and adversarial attacks on multimodal models.

3.2. Adversarial Attack on Image

White-box attacks assume complete knowledge of the model architecture, parameters, and training data, allowing attackers to directly compute gradients and craft highly effective perturbations [14,15]. In contrast, black-box attacks operate with limited or no knowledge of the model internals, relying instead on query access to generate adversarial examples through techniques such as transfer attacks or gradient estimation [16,17]. Between these extremes lie gray-box attacks, where partial information about the model is available [18].

For image data, notable techniques include the Fast Gradient Sign Method (FGSM). FGSM, introduced by Goodfellow et al. [19], generates adversarial examples through a single-step gradient update in the direction that maximizes the loss function, offering computational efficiency at the cost of perturbation strength. This method creates adversarial perturbations by exploiting the gradient information of models [14,20,21]. PGD, developed by Madry et al. [15], extends this approach through iterative optimization with projection steps to maintain perturbations within specified bounds, producing more robust adversarial examples that have become a standard benchmark for evaluating model robustness. Additional sophisticated image attack methods include the Carlini & Wagner (C&W) attack [14], which formulates adversarial example generation as an optimization problem to find minimal perturbations, and DeepFool [22], which computes the minimum perturbation required to cross decision boundaries.

3.3. Adversarial Attacks on Text

Adversarial attacks on textual data present unique challenges due to the discrete nature of language and the constraints imposed by semantic coherence and grammatical correctness [23]. Unlike continuous image perturbations, text adversarial examples must maintain readability while altering model predictions. Character-level attacks introduce minor modifications such as typos, insertions, or deletions that exploit model vulnerabilities to noisy inputs [24,25]. Word-level attacks substitute words with synonyms or contextually similar alternatives, often using word embeddings to identify replacements that preserve semantic meaning while misleading classifiers [7,26]. Sentence-level attacks involve more substantial modifications, including paraphrasing or reordering, to evade detection while maintaining the original intent [27,28]. Recent work has explored the use of large language models to generate fluent adversarial text that is virtually indistinguishable from natural language [29].

3.4. Multimodal Adversarial Attacks

While adversarial robustness has been extensively studied for unimodal settings [2–4], with substantial progress in understanding attacks on images [5,6] and text [7,8] independently, research on adversarial attacks targeting multimodal models remains comparatively limited. Early work on multimodal adversarial attacks focused primarily on vision–language models. Xu et al. [30] demonstrated that perturbations in image inputs could significantly affect text-conditioned generation tasks, while Shayegani et al. [31] showed that carefully crafted visual inputs could manipulate language model outputs in vision–language models.

For CLIP specifically, several studies have examined adversarial vulnerabilities. Mao et al. [9] investigated the robustness of CLIP's zero-shot classification capabilities under adversarial perturbations, finding that the model exhibits vulnerabilities similar to those of traditional supervised models. Schlarmann and Hein [10] demonstrated that CLIP can be attacked through both image and text modalities, with attacks on the text encoder proving particularly effective due to the discrete optimization challenges in crafting adversarial text.

4. Dataset

4.1. Hateful Meme Dataset

For this experiment, we will use the Facebook Hateful Memes dataset, a multimodal dataset designed for studying hate speech detection in memes by combining image and text modalities. The dataset consists of 10,000 memes, each containing both an image and an overlaid text caption, making it well-suited for evaluating the adversarial robustness of

CLIP in a sentiment analysis context (See Figure 1). The dataset also includes a JSON file containing the image path, a text field with the meme's text, and a label indicating whether the meme is hateful. The dataset can be found in Kaggle [32]. The overview of the dataset is given below:

- Total samples: 10,000 memes
- Train set: 8500 memes
- Validation set: 500 memes
- Test set: 1000 memes
- Labels: Binary classification (0 = non-hateful, 1 = hateful)
- Number of samples per class for Training set: Class 0 = 5450, Class 1 = 3050

Figure 1. Random sample of images from Hateful Meme Dataset. The samples show the variety of colors, wording, and tone.

DataPreprocessing

All images are preprocessed using the official pipeline associated with the **openai/clip-vit-base-patch32** model to ensure consistency with CLIP's pretraining configuration. Each image is resized such that the shorter side is scaled to 224 pixels while preserving the original aspect ratio using bicubic interpolation, followed by a center crop to obtain a fixed 224 × 224 resolution. The cropped image is then converted to a floating-point tensor with pixel values scaled to the [0, 1] range and normalized using CLIP's predefined mean (0.48145466, 0.4578275, 0.40821073) and standard deviation (0.26862954, 0.26130258, 0.27577711). This normalization aligns the input distribution with the statistics used during CLIP pretraining. All adversarial perturbations are applied after preprocessing and normalization, ensuring that gradients are computed in the same input space used during training and evaluation.

To mitigate class imbalance in the Hateful Memes dataset, we performed data synthesis by augmenting the minority class, the 'Hateful' class. The augmentation is performed by duplicating underrepresented samples with slight text variations (e.g., paraphrasing or synonyms) and applying mild image augmentations (e.g., color jittering, slight rotation, and random cropping).

4.2. Suspicious Car-Parts Dataset

To determine whether an advertisement on a marketplace platform is likely to involve car parts that may appear suspicious of being stolen, we first collected data from two popular websites: Craigslist and OfferUp. Using a custom web scraping process, we gathered approximately 1,048,575 posts along with their associated images. An example of a post related to car parts from our dataset is shown in Figure 2. This marketplace corpus has also

supported related studies on visual representation learning and downstream analysis of car-parts listings, including single-modal visual embedding analysis and clustering-based evaluation [33,34].

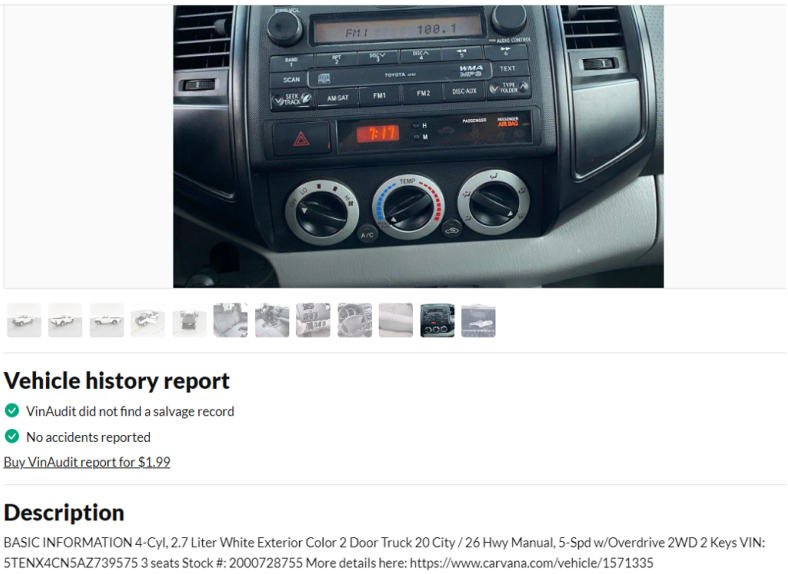


Figure 2. An example of a post from OfferUp, which includes several images and textual descriptions.

For the Suspicious Car Parts dataset, image preprocessing follows the same standardized CLIP pipeline used for the Hateful Memes dataset to ensure consistency across experiments. Raw marketplace images are first loaded in RGB format without manual cropping or additional filtering. Each image is resized so that the shorter side is scaled to 224 pixels, preserving the aspect ratio using bicubic interpolation, followed by a center crop to obtain a 224 × 224 resolution. Dataset-specific augmentation or custom preprocessing has not been applied prior to adversarial training, to ensure that performance differences arise from adversarial perturbations rather than preprocessing discrepancies.

Since manual labeling of a dataset of this size was not practical, we developed an automated proxy-labeling method based on a relatedness graph. The goal of this graph was not to establish ground-truth theft, but to identify posts with elevated risk indicators based on repeated or coordinated posting behavior. To build the graph, we created a named entity recognition (NER) dataset that extracted key information from each post. The NER system identified several types of entities, including username, phone number, email, location, and longevity, following an entity-centric view of marketplace listings that has also been used in related SCP studies [35]. For graph construction, we primarily used extracted phone numbers.

Along with phone number, we also needed to know if multiple posts have the same image. For this purpose, we used a hash map generated during the data collection phase. In our dataset, image metadata is stored in a JSON format where the filename contains a unique hash identifier. By parsing the string following the first underscore in the file path (extracting the hash from . . ._09G07g. . .), we obtained a unique hash index for each image. We then compared these indices across the dataset, and a match indicated that two different posts contained the exact same image file. These signals reflect relatedness between posts, not proof that the listed items are stolen.

Using both phone number matches and shared image hashes, we built a relatedness graph where each node represented an individual post. An edge was added between two nodes if the posts shared a phone number or an image as shown in Figure 3. Once the graph was constructed, we computed all connected components (CCs). Each CC group's posts

are mutually accessible through these relationships, indicating a high degree of connection among the posts. The methodology is shown in Figure 4.

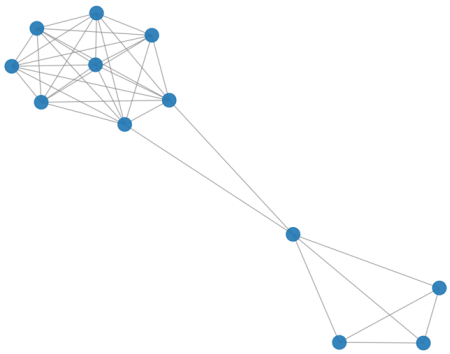


Figure 3. A connected component of size 12. The two subgraphs are connected by a particular node (colored in blue), e.g., a node that shares a phone number with five other nodes.

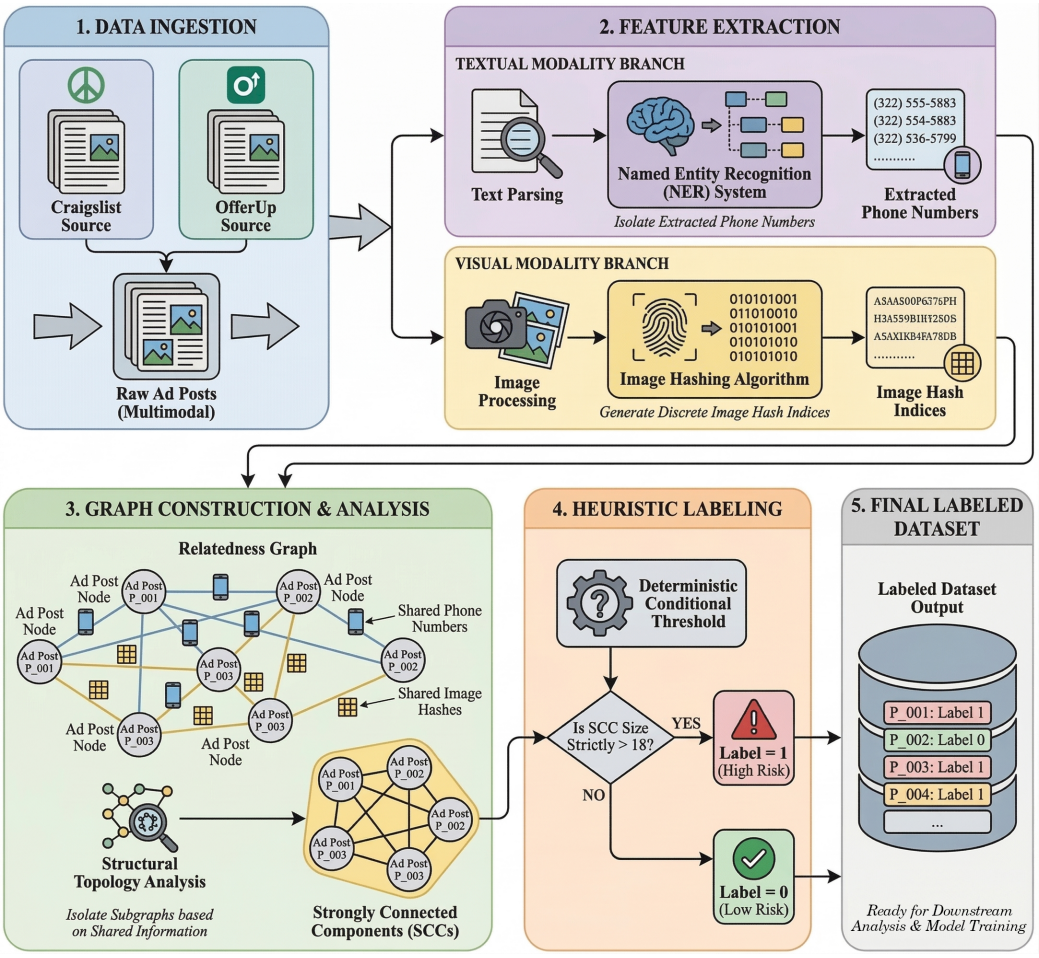


Figure 4. Labeling methodology of suspicious car parts posts.

From the graph, we identified a total of 51,447 connected components (CCs) of different sizes. To define a high-risk proxy class, we used a heuristic threshold of eighteen posts per connected component. Under our graph construction, larger components reflect stronger repeated or coordinated posting patterns induced by shared phone numbers and/or exact image reuse. We emphasize that this cutoff is an operational choice for proxy labeling, not a validated universal boundary for determining whether a listing contains stolen parts. Thus, posts belonging to CCs larger than eighteen were assigned the high-risk proxy label

(label 1), while posts in smaller components were assigned the low-risk proxy label (label 0). This procedure yielded approximately 15% of posts in the high-risk class [36]. We now present that prevalence only as a broad contextual outcome of the labeling procedure, rather than as direct validation from the prior literature. Future work should examine the sensitivity of results to alternative thresholds and, where feasible, compare proxy labels against manually audited subsets. After constructing the relatedness graph and labeling posts using CC membership, the dataset was split into training and validation sets while maintaining class distribution consistency across splits. Because only approximately 15% of posts were labeled as high-risk, preserving class balance on splits was necessary to ensure reliable evaluation using balanced accuracy. Marketplace datasets often contain repeated seller identifiers or contact information that can correlate with labels. In our dataset, such information was used only for labeling through graph construction and not as explicit input features. The model receives only the post text and associated images, reducing the likelihood that predictions are supported by phone numbers, URLs, or metadata rather than semantic content. Although labels are derived from graph components, the model input does not include phone numbers or image hashes, reducing direct leakage of labeling features. This dataset design also provides a useful test bed for related studies of marketplace-image representations and few-shot extraction from SCP listings [33–35].

The overview of the dataset is given below:

- Total samples: 1,048,575 posts
- Train set: 838,860 posts
- Validation set: 209,715
- Labels: Binary classification (0 = no risk, 1 = high risk)

To illustrate the pipeline in Figure 4, we discuss a simplified but representative example of how five hypothetical marketplace posts are processed from raw collection to their final labels. The example is constructed to reflect the types of patterns observed in the real dataset.

Image hashes are computed from the image filename column. In Table 1, posts P1, P2, and P3 all share phone number 2452242825. Additionally, P1 and P2 share image hash 09G07g, indicating they contain the exact same image. An undirected graph is constructed, with each node representing a post. An edge is added between two nodes if they share a phone number or an image hash. Applying this rule, we have P1–P2 (shared phone 2542242825 AND shared image hash 09G07g) and P1–P3 (shared phone 2542242825). Thus, we have a resulting graph containing a connected component of size 3 for posts P1, P2, and P3. As the size of the connected component is less than 18, we can label posts P1, P2, and P3 as 0 for low risk.

Table 1. Example of Suspicious Car Parts dataset.

Post	Title	Description	NER Phone Number	Image Filename
P1	BMW 3-series front bumper	Good condition, minor scratches	2542242825	eSWztVe_09G07g_600x450.jpg: 1186545.jpg
P2	BMW bumper	All power options work. Cash is preferred	2542242825	rWsftZt_09G07g_600x450.jpg: 3445561.jpg
P3	Car sale BMW	All power options work. Cash is preferred	254-224-2825	qRsftGr_0BR15g_600x450.jpg: 6712364.jpg
P4	Honda Civic headlights	Good condition. Cash only	796-444-1295	gRebTe_0TORr2g_600x450.jpg: 5535513.jpg

5. Methodology

We want to evaluate the robustness of visual language models such as CLIP in a classification task using the Hateful Memes and Suspicious Car Parts datasets. For this purpose, we designed a series of experiments to understand which modality impacts adversarial robustness. Our main goal is to assess how adversarial perturbations affect CLIP's performance when applied to its image and text modalities, individually and together, and also to understand the efficiency of adversarial training as a defense mechanism. The specific methods used in this study are described below, with a focus on adversarial example generation, selective modality freezing, and robustness evaluation under different conditions.

In Figure 5, we show that from the hateful meme dataset, each input meme consists of an image and its corresponding text caption. To process this multimodal input, we first extract features from both modalities separately. The image is passed through CLIP's vision encoder, which converts it into a high-dimensional embedding that captures its visual semantics. Simultaneously, the text associated with the meme is tokenized using CLIP's tokenizer and then processed by the text encoder to generate a corresponding text embedding. These two embeddings are then jointly fed into the CLIP model, which is designed to align and understand cross-modal relationships between visual and textual inputs.

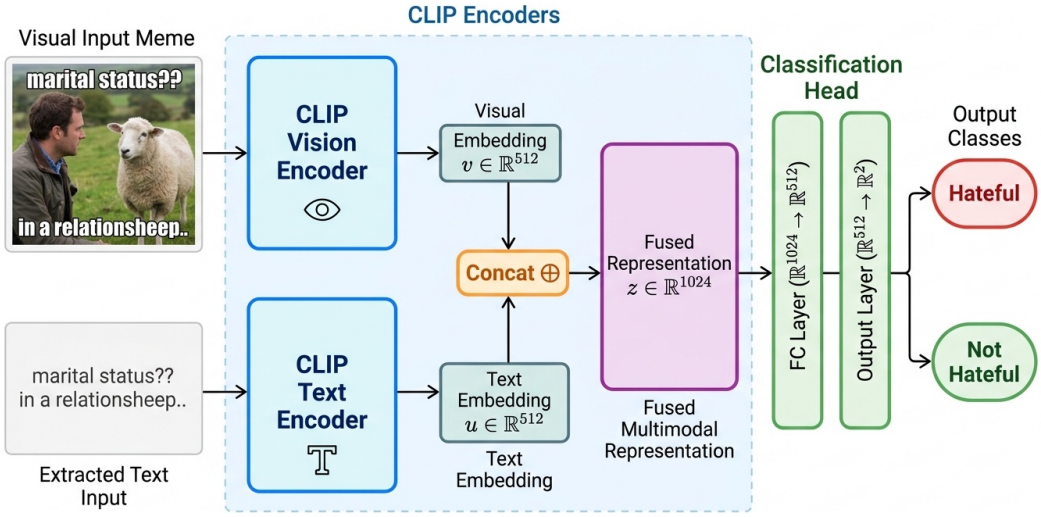


Figure 5. Baseline architecture of methodology of binary classification with CLIP using Hateful Memes dataset.

Given an input pair (x_{img}, x_{txt}), the pretrained CLIP encoders independently produce two L2-normalized embeddings. The image encoder (*ViT-B/32*) maps x_{img} to a vision embedding $v \in \mathbb{R}^{512}$, and the text encoder maps x_{txt} to a text embedding $u \in \mathbb{R}^{512}$. Text inputs are tokenized using the CLIP processor with padding and truncation before being passed to the encoder. Both embeddings are L2-normalized, following the standard CLIP implementation:

$$\hat{v} = \frac{v}{\|v\|_2}, \quad \hat{u} = \frac{u}{\|u\|_2}, \quad \text{where } \hat{v}, \hat{u} \in \mathbb{R}^{512}.$$

The two normalized embeddings are fused by concatenation along the feature dimension to form a single multimodal representation $z \in \mathbb{R}^{1024}$. Concatenation was chosen over element-wise operations (addition, multiplication) because it preserves the full, independent information content of both modalities in a single vector, allowing the subsequent classification head to learn modality-specific discriminative features rather than forcing them into a shared activation pattern. The fused representation $z \in \mathbb{R}^{1024}$ is passed through a two-layer classification head:

1. Fully connected layer: Input dimension is 1024, Output dimension is 512.
2. Output layer: Input dimension is 512, Output dimension is 2 (binary classification)

Since our task is supervised binary classification, the model is trained end-to-end using cross-entropy loss with L2 weight decay ($\lambda = 1 \times 10^{-4}$) applied to the classification head parameters. For a batch of N samples, the training objective is:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i \mid z_i) + \lambda \|W\|_2^2, \quad (1)$$

where $p(y_i \mid z_i)$ is the softmax probability assigned to true label $y_i \in \{0, 1\}$, and W denotes the classification head weight matrix. The Adam optimizer is used with a learning rate of 5×10^{-4}.

- **Baseline:** At the beginning, we established a performance baseline for CLIP on the Hateful Memes dataset. The dataset contains multimodal image–text pairs labeled as hateful or non-hateful. We use CLIP's pretrained vision encoder to process the images and its text encoder to process the corresponding text extracted from the dataset's JSON file. The image and text embeddings are then combined and passed through a classification head to predict whether a meme is hateful or not. We evaluate the model using accuracy and F1-score for this classification task. This baseline process is illustrated in Figure 5.
- **Adversarial Example with Perturbed Image:** We applied perturbations to the images using the Fast Gradient Sign Method (FGSM), an adversarial attack that modifies pixel values in the direction of the gradient of the loss function with respect to the input image [19]. We began our approach with the PGD, making it a suitable choice for an initial exploration of adversarial vulnerabilities. The magnitude of the perturbation $\epsilon = 8/255$ is to ensure subtle but effective alterations. Text input remains unchanged and we measure the resulting drop in CLIP classification performance.
- **Adversarial Example with Perturbed Text:** To generate adversarial examples for the text modality, we used a similar approach to the image-based attack by applying the FGSM directly to the text embeddings. Unlike images, textual inputs are discrete and cannot be directly perturbed using gradient-based methods without breaking syntactic validity or tokenizer constraints. To maintain semantic consistency while enabling gradient-based adversarial perturbations, we apply FGSM perturbations in the continuous text-embedding space produced by CLIP's text encoder. This allowed us to evaluate CLIP's performance under text-specific adversarial perturbations and assess the susceptibility of the language encoder to such attacks.
- **Image–Text Adversarial Example:** To evaluate the combined effect of adversarial attacks on both modalities, we simultaneously applied perturbations to the image and text inputs using FGSM. This joint attack scenario created a more realistic adversarial setting, where an attacker targeted visual and textual components to compromise the model's performance. It showed a comprehensive assessment of the robustness of CLIP when faced with multimodal adversarial threats.
- **Adversarial Training with Frozen Language Encoder:** To strengthen the robustness of CLIP and understand vulnerabilities specific to each modality, we conducted adversarial training while selectively freezing the image or text encoder. In this phase, we generate adversarial examples from the image modality and retrain the model while keeping the language encoder's parameters fixed. This allowed us to assess the ability of the visual encoder to adapt to adversarial perturbations. During training, only the vision encoder's parameter was updated.

- **Multimodal Perturbation under Frozen Text Encoder:** We repeated the previous phase, but in this phase, we created both text and image adversarial examples to perform adversarial training along with clean images and texts. This setup evaluated whether enhancing the robustness of the vision encoder alone can effectively defend against multimodal attacks.
- **Text-Modality Attack with Frozen Vision Encoder:** We generate adversarial examples from the text modality and retrain the CLIP model while keeping the vision encoder frozen. This approach isolates the ability of the language encoder to handle adversarial perturbations and allows us to evaluate its independent contribution to the overall robustness of the model.
- **Multimodal Adversarial Training with Frozen Image Encoder:** We repeated the previous setup by training on adversarial examples from both the image and text modalities while keeping the vision encoder frozen. This allowed us to assess the language encoder's capacity to manage combined adversarial attacks and its impact on maintaining model robustness under dual perturbations.
- **Full Adversarial Training:** We conducted adversarial training using perturbed inputs from both the image and text modalities without freezing any component of the model. This training approach allowed both the vision and language encoders to jointly adapt to adversarial perturbations, serving as a comprehensive defense strategy to maximize overall model robustness.

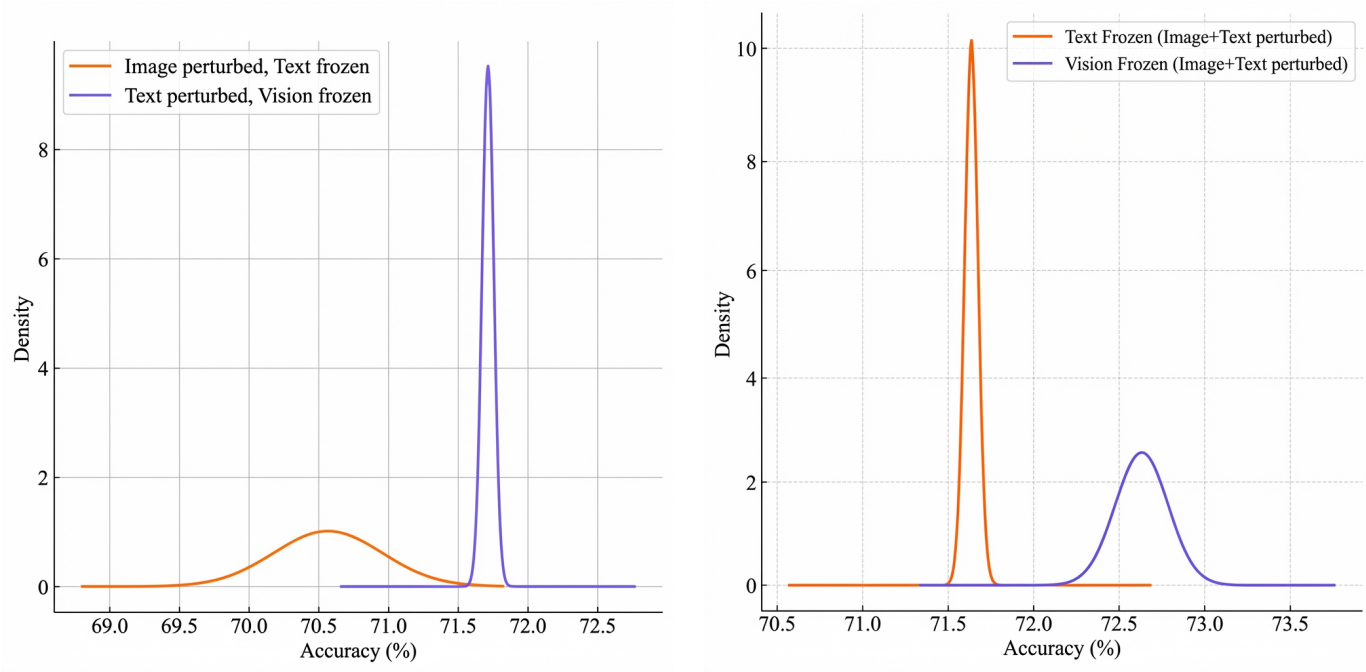
By repeating this methodology on the Suspicious Car Parts dataset, we wanted to identify which modality, image or text, is more vulnerable to adversarial attack in the CLIP model. As we compare frozen and unfrozen adversarial training, our goal is to determine whether modality-specific defenses can achieve robustness comparable to that of full-model training, to make adversarial training more efficient.

6. Experiments

6.1. Evaluation Method

For quantitative evaluation, we use F1-score and balanced accuracy to assess CLIP's robustness on the two datasets, particularly under adversarial attacks. These metrics will help measure performance degradation caused by adversarial perturbations and improvements gained through adversarial training. We compare results against baseline CLIP performance on clean data, expecting a significant drop under attack and partial recovery with adversarial training.

To account for randomness in training and adversarial example generation, each experiment was repeated five times using different random seeds. For each phase, we report the average performance across these runs using the evaluation metrics described earlier (F1-score for the Hateful Memes dataset and balanced accuracy for the Suspicious Car Parts dataset). Along with the average results, we also examine how the results vary between runs by analyzing the distribution of performance under each experimental setting. These distributions are visualized in Figure 6 to illustrate the stability of the model under different adversarial training setups. This evaluation helps ensure that the reported robustness trends are consistent and not driven by a single stochastic training instance.



(**a**) (**b**)

Figure 6. Normal distribution of model performance on five independent runs under different adversarial training configurations. (**a**) Performance distributions when only one modality is adversarially perturbed. (**b**) Performance distributions when both the image and text modalities are adversarially perturbed simultaneously.

6.2. Experimental Details

We have designed the experiments to evaluate the robustness of CLIP against various adversarial scenarios on the Hateful Memes dataset. For this purpose we used the following setup:

- *Model Configuration:* We used CLIP as the base model for our experiments. We used the pretrained model, which is **openai/clip-vit-base-patch32** from the HuggingFace Transformers library. This model uses a Vision Transformer (ViT) [37] model architecture for image encoding; the 'ViT-B' base variant with a patch size of 32. Both the image and text encoder extract 512-dimensional image and text features. The classification head is a fully connected layer that maps the 1024-d concatenated multimodal embedding to 2 output logits. The 1024-d input derives from concatenating the two independently L2-normalized 512-d embeddings produced by CLIP's image and text encoders, respectively.
- *Hyperparameter Configuration:* In this experiment, we used **Adam** optimizer with a learning rate of 0.0005. We used L2 regularization to prevent overfitting. All experiments were conducted on an L40s GPU with 35 GB of dedicated memory. A batch size of 32 is used for both datasets across all training conditions. This value balances GPU memory utilization on the L40s (35 GB) with stable gradient estimates, and is consistent with common practice for fine-tuning CLIP models. The training is run for a maximum of 10 epochs for the Hateful Memes dataset and 5 epochs for the Suspicious Car Parts dataset. The reduced epoch count for Suspicious Car Parts reflects its substantially larger training set. We applied early stopping with a patience of 3 epochs based on validation balanced accuracy (Suspicious Car Parts) or validation F1-score (Hateful Memes); if the monitored metric does not improve for 3 consecutive epochs, training is halted and the checkpoint with the best validation metric is restored

for evaluation. This prevents overfitting during adversarial retraining, where the model is exposed to perturbed inputs that can cause unstable loss.

- *Attack Configuration and Threat Model:* To evaluate robustness under realistic adversaries, we apply both FGSM and PGD attacks under a white-box, untargeted l_∞-bounded threat model. FGSM generates adversarial examples in a single gradient step, making it a computationally efficient baseline that establishes a lower bound on model vulnerability. PGD extends FGSM through iterative optimization with projection, producing stronger adversarial examples that represent a more reliable upper bound on vulnerability. The perturbation budget for image-based attacks is set to $\epsilon = 8/255$ under the l_∞ norm, which is a commonly adopted standard in the adversarial robustness literature for evaluating perceptually constrained attacks on image classifiers. This magnitude ensures that perturbations remain visually imperceptible while being sufficiently strong to meaningfully affect model robustness.
  The PGD step size is set to $\alpha = 2/255$, which balances sufficient gradient signal per step against the risk of overshooting the ϵ-ball boundary. Five iterations are used, which is sufficient to produce near-convergent adversarial examples within the small $\epsilon = 8/255$ budget without high computational cost during adversarial retraining. The same perturbation budget and iteration settings are consistently applied across all modality-specific and multimodal attack scenarios to ensure fair comparison between vision-only, text-only, and joint perturbations.
  For text embedding attacks, the perturbation magnitude is chosen to match the scale of the image perturbation budget relative to embedding norms, ensuring that attacks remain bounded within a controlled l_∞ constraint in the continuous embedding space.

7. Result and Analysis

In Tables 2 and 3, we see that adversarial perturbations applied to the vision modality consistently degrade performance across both datasets. On the Hateful Memes dataset, attacking only the vision modality reduces the F1-score from 75 to 62.34, which is slightly better than attacking the language modality (62.29). In the Suspicious Car Parts dataset, attacking the vision modality reduces balanced accuracy from 79 to 74, matching the degradation caused by language-only attacks (74.61). When both modalities are attacked simultaneously, the model shows the largest performance drop, indicating that CLIP is not robust across modalities and is vulnerable to multimodal perturbations.

Table 2. F1-scores of CLIP under various attack and adversarial training conditions on Hateful Meme dataset. C = Clean, P = Perturbed.

Method	Vision Modality	Text Modality	F1-Score
Baseline	C	C	75.00
Attacking Vision Model	P	C	62.34
Attacking Language Model	C	P	62.29
Attacking Vision and Language Models	P	P	61.12
Retraining Vision Model and Freezing Language Model	P	C/P	71.65/71.76
Retraining Language Model and Freezing Vision Model	C/P	P	72.76/70.8
Retraining Vision and Language Models	P	P	71.39

Table 3. Balanced accuracy of CLIP under various attack and adversarial training conditions on Suspicious Car Parts dataset. C = Clean, P = Perturbed.

Method	Vision Modality	Text Modality	Accuracy
Baseline	C	C	79.00
Attacking Vision Model	P	C	74.00
Attacking Language Model	C	P	74.61
Attacking Vision and Language Models	P	P	74.00
Retraining Vision Model and Freezing Language Model	P	C/P	77.87/75.11
Retraining Language Model and Freezing Vision Model	C/P	P	75.12/74.80
Retraining Vision and Language Models	P	P	77.92

We evaluate adversarial retraining strategies by selectively training one modality while freezing the other. On the Hateful Memes dataset, we observe that the model experiences a significant drop in performance when adversarial perturbations are applied to the image modality. When we use perturbed images or keep the vision encoder frozen during adversarial retraining, performance degrades substantially (71.65 and 70.80, respectively). This indicates that the model struggles to maintain robustness when visual representations are compromised or when it cannot learn from the data. Similarly, in the Suspicious Car Parts dataset, during the retraining phase, the worst performance is obtained either when the image encoder is frozen or when adversarial perturbations are applied to the images (75.11 and 74.80, respectively). In both cases, the model fails to recover its performance, suggesting that the visual component plays a critical role in determining the overall robustness of the model.

Tables 2 and 3 reveal that CLIP shows the greatest vulnerability when adversarial perturbations are applied to the image modality, regardless of whether the accompanying text is clean or perturbed. We ran each phase five times, plotted the results, and generated a normal distribution as shown in Figure 6. In Figure 6a, where only one modality is perturbed, the accuracy drops more noticeably when images are attacked and the text encoder is frozen, while the model remains more stable when text embeddings are perturbed and the vision encoder is frozen. This indicates that the vision modality is more vulnerable to adversarial perturbations, as image attacks cause greater performance degradation and variability. In Figure 6b, where both the image and text modalities are adversarially perturbed, the model still achieves slightly better performance when the vision encoder is frozen compared to when the text encoder is frozen. However, we also observe more variability in performance when the image encoder is frozen. This increased variability suggests that the vision modality is less stable and more sensitive to adversarial perturbations. In other words, small changes in the visual input lead to larger and less predictable fluctuations in performance, indicating that the vision encoder is more prone to attack and therefore less robust than the language encoder.

The confusion matrices in Figure 7 show that the retraining language and freeze vision model with both perturbed image and text inputs has the highest misclassification rates, meaning it performs the worst as shown in Figure 7a. This indicates that when the vision encoder is frozen, the model becomes highly vulnerable to adversarial attacks, especially from image perturbations. However, allowing the vision encoder to adapt improves robustness. Overall, these results confirm that the vision modality is the weaker part of CLIP under adversarial conditions.

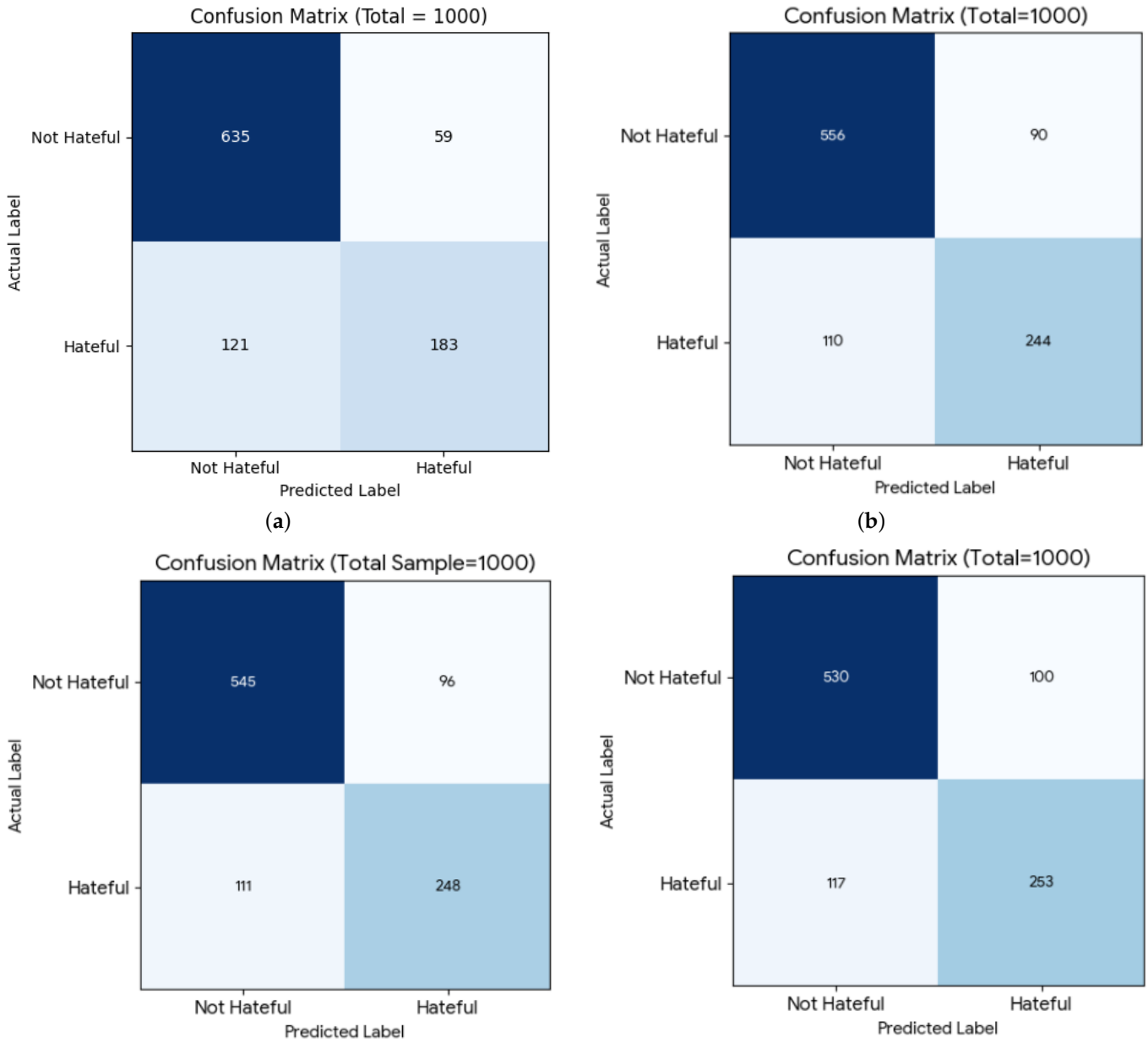


Figure 7. Confusion matrix for hateful meme classification under four conditions. (**a**) Retraining the language model and freezing the vision model with perturbed images and text. (**b**) Retraining the language model and freezing the vision model with perturbed text and clean images. (**c**) Retraining the vision model and freezing the language model with perturbed images and text. (**d**) Retraining the vision model and freezing the language model with perturbed images and clean text.

8. Discussion

8.1. Reasons for Vision Modality Vulnerability

This study consistently shows that the vision modality in CLIP is more sensitive to adversarial perturbations than the text modality. In both datasets used in this work (the Hateful Memes and Suspicious Car Parts datasets), attacks on the image modality resulted in larger performance drops and greater variation across experimental runs. These results indicate that the visual encoder is a major point of vulnerability in CLIP's multimodal architecture.

This observation can be explained by several architectural characteristics of CLIP. Images are represented as continuous pixel values, which makes them naturally vulnerable to small gradient-based perturbations. Even slight changes in pixel intensity can pass through the vision encoder and significantly alter the resulting embedding. However, text inputs are discrete and must first be converted into token embeddings, which provides some inherent robustness because attacks cannot directly modify the original tokens.

CLIP's image encoder uses a Vision Transformer (ViT-B/32) architecture, which divides each 224 × 224 image into a 7 × 7 grid of 49 non-overlapping 32 × 32-pixel patches. Each patch is linearly projected into a 768-dimensional token embedding and processed jointly by the Transformer's self-attention layers. Because adversarial perturbations are applied uniformly across all pixel dimensions before patch extraction, a single $\epsilon = 8/255$ perturbation simultaneously corrupts all 49 patch embeddings. The self-attention mechanism then propagates these corrupted patch representations across the entire sequence, allowing a perturbation originating in one spatial region to distort the global image representation through attention-weighted aggregation.

Text perturbations in this study are applied in the continuous embedding space rather than directly modifying the discrete tokens. While this allows gradient-based attacks to be applied, it still constrains the magnitude of meaningful semantic changes compared to pixel-level perturbations in images. As a result, adversarial modifications to text embeddings tend to produce smaller shifts in the joint representation than equivalent perturbations in the visual modality.

8.2. Statistical Variability and Stability of Adversarial Conditions

The normal distribution plots in Figure 6 show an important insight that mean performance alone does not reveal. When adversarial perturbations are applied to images, the results show wider distributions (higher variance) across the five runs compared to when only text is perturbed. In Figure 6a, the distribution for the image-perturbed/text-frozen condition is much wider than the text-perturbed/vision-frozen condition, indicating that image attacks not only reduce performance but also cause greater instability between runs.

In Figure 4b, where both modalities are perturbed, the vision-encoder-frozen distribution is wider than the text-encoder-frozen distribution, even though the frozen-vision condition achieves slightly higher mean accuracy. This widening of variance under frozen-vision conditions during multimodal attack suggests that when the vision encoder is prevented from adapting during adversarial retraining, the model's performance becomes more sensitive to random initialization and batch ordering, presumably because the text encoder alone cannot consistently compensate for the corrupted visual representations in different training runs.

From a deployment perspective, variability is just as important as mean performance. A system with slightly lower accuracy but stable results may be more reliable than one with higher mean accuracy but unpredictable failures, especially in high-risk applications like content moderation or fraud detection. Our results show that retraining strategies that allow the vision encoder to adapt not only improve average performance but also lead to more consistent and predictable behavior. This important practical benefit can be overlooked when results are reported only as mean values without considering variability.

8.3. Open Research Questions and Future Directions

While this work provides empirical evidence of modality-specific adversarial vulnerability in CLIP, several open research questions remain. It is not clear whether the vulnerability patterns observed in this study generalize to other multimodal architec-

tures beyond CLIP. Future work should evaluate modality-specific robustness in newer vision–language models to determine whether the models show similar patterns.

The interaction between modalities under adversarial perturbations warrants further investigation. In multimodal systems, perturbations applied to one modality may indirectly influence the representation of another modality through cross-modal alignment mechanisms. Understanding these interactions could lead to improved defense strategies.

This study focuses on gradient-based white-box attacks under controlled perturbation constraints. Future research should investigate additional threat models, including black-box attacks, generative adversarial methods, and real-world adversarial scenarios, to better understand the robustness of multimodal systems in practical deployment environments.

9. Conclusions

This paper presented a systematic study of modality-specific adversarial vulnerabilities in the CLIP vision–language model. By applying adversarial perturbations independently to image and text modalities and evaluating robustness across two multimodal datasets, we demonstrated that perturbations applied to the visual modality consistently produce greater performance degradation and variability than perturbations applied to textual inputs. These findings suggest that the visual encoder represents a primary vulnerability in CLIP's architecture. The results highlight the importance of analyzing robustness at the modality level and suggest that future defense strategies for multimodal systems may benefit from prioritizing the protection of the most vulnerable component.

Author Contributions: Conceptualization, M.B.R. and P.R.; methodology, M.B.R. and P.R.; validation, M.B.R.; investigation, M.B.R. and P.R.; writing—original draft preparation, M.B.R. and P.R.; writing—review and editing, M.B.R. and P.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was executed while P.R. was funded by the National Science Foundation under grant NSF CISE-CNS Award 2136961 (link: https://www.nsf.gov/awardsearch/showAward?AWD_ID=2136961&HistoricalAwards=false, access date: 11 February 2026).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: This study utilized the publicly available Hateful Meme dataset (link: https://www.kaggle.com/datasets/parthplc/facebook-hateful-meme-dataset, access date: 10 February 2025). The images originate from public sources and are employed here exclusively for non-commercial scientific research. As the data were anonymized and publicly released by the original creators, no further ethics approval was required under the applicable guidelines.

Data Availability Statement: The facial images used in this study are sourced from the Hateful Meme dataset (link: https://www.kaggle.com/datasets/parthplc/facebook-hateful-meme-dataset, access date: 10 February 2025), which is publicly available for non-commercial research purposes [32]. The authors do not have permission to redistribute the raw image data. All facial images presented in this paper are used solely for academic research in accordance with the dataset's license terms.

Acknowledgments: The authors want to thank Laurie Giddens, Gisela Bichler, Stacie Petter, Tomas Cerny, Alejandro Rodriguez Perez, and Javier Turek for earlier discussions regarding the SCP dataset. The authors thank the reviewers for their useful feedback. The authors acknowledge the use of generative AI tools only for spell-checking (Grammarly, free) and to aid in rephrasing a few unclear sentences during peer review (ChatGPT-5.2). All writing, research design, experimental implementation, data analysis, and scientific conclusions were developed and validated entirely by the authors with no generative AI.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the writing of the manuscript; or in the decision to publish the results.

References

1. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. *arXiv* **2021**, arXiv:2103.00020. [CrossRef]
2. Zhang, J.; Li, C. Adversarial examples: Opportunities and challenges. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 2578–2593. [CrossRef] [PubMed]
3. Kong, Z.; Xue, J.; Wang, Y.; Huang, L.; Niu, Z.; Li, F. A survey on adversarial attack in the age of artificial intelligence. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 4907754. [CrossRef]
4. Han, X.; Zhang, Y.; Wang, W.; Wang, B. Text adversarial attacks and defenses: Issues, taxonomy, and perspectives. *Secur. Commun. Netw.* **2022**, *2022*, 6458488. [CrossRef]
5. Guo, C.; Rana, M.; Cissé, M.; van der Maaten, L. Countering adversarial images using input transformations. *arXiv* **2017**, arXiv:1711.00117. [CrossRef]
6. Dubey, A.; van der Maaten, L.; Yalniz, Z.; Li, Y.; Mahajan, D. Defense against adversarial images using web-scale nearest-neighbor search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
7. Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.; Srivastava, M.; Chang, K.W. Generating natural language adversarial examples. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, 31 October–4 November 2018. [CrossRef]
8. Wang, W.; Wang, R.; Wang, L.; Wang, Z.; Ye, A. Towards a robust deep neural network in texts: A survey. *arXiv* **2019**, arXiv:1902.07285. [CrossRef]
9. Mao, C.; Chiquier, M.; Wang, H.; Yang, J.; Vondrick, C. Understanding zero-shot adversarial robustness for large-scale models. *arXiv* **2022**, arXiv:2212.07016. [CrossRef]
10. Schlarmann, C.; Hein, M. On the adversarial robustness of multi-modal foundation models. *arXiv* **2023**, arXiv:2308.10741. [CrossRef]
11. Pawar, P.P.; Kumar, D.; Addula, S.R.; Naveed Cheema, Q.; Haq, A.U.; Kumar Meesala, M. A Blockchain-Based IoT Framework for Smart Homes: Enhancing Energy Prediction and Security with LSTM and Equilibrium Optimization. In Proceedings of the 2025 International Conference on Intelligent and Cloud Computing (ICoICC), Bhubaneswar, India, 2–3 May 2025; pp. 1–8. [CrossRef]
12. Ijiga, O.; Idoko, I.; Ebiega, G.; Olajide, F.; Olatunde, T.; Ukaegbu, C. Harnessing adversarial machine learning for advanced threat detection: AI-driven strategies in cybersecurity risk assessment and fraud prevention. *Open Access Res. J. Sci. Technol.* **2024**, *11*, 1–4. [CrossRef]
13. Pitropakis, N.; Panaousis, E.; Giannetsos, T.; Anastasiadis, E.; Loukas, G. A taxonomy and survey of attacks against machine learning. *Comput. Sci. Rev.* **2019**, *34*, 100199. [CrossRef]
14. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*; IEEE: New York, NY, USA, 2017; pp. 39–57. [CrossRef]
15. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018. [CrossRef]
16. Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; Swami, A. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, Abu Dhabi, United Arab Emirates, 2–6 April 2017; pp. 506–519. [CrossRef]
17. Ilyas, A.; Engstrom, L.; Athalye, A.; Lin, J. Black-box adversarial attacks with limited queries and information. In Proceedings of the International Conference on Machine Learning. PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2137–2146. [CrossRef]
18. Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. Ensemble adversarial training: Attacks and defenses. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018. [CrossRef]
19. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572. [CrossRef]
20. Ren, K.; Zheng, T.; Qin, Z.; Liu, X. Adversarial attacks and defenses in deep learning. *Engineering* **2020**, *6*, 346–360. [CrossRef]
21. Engstrom, L.; Ilyas, A.; Athalye, A. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv* **2018**, arXiv:1807.10272. [CrossRef]
22. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582. [CrossRef]
23. Zhang, W.E.; Sheng, Q.Z.; Alhazmi, A.; Li, C. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol. (TIST)* **2020**, *11*, 1–41. [CrossRef]
24. Ebrahimi, J.; Rao, A.; Lowd, D.; Dou, D. Hotflip: White-box adversarial examples for text classification. *arXiv* **2017**, arXiv:1712.06751. [CrossRef]

25. Gao, J.; Lanchantin, J.; Soffa, M.L.; Qi, Y. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW)*; IEEE: New York, NY, USA, 2018; pp. 50–56. [CrossRef]
26. Ren, S.; Deng, Y.; He, K.; Che, W. Generating natural language adversarial examples through probability weighted word saliency. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1085–1097. [CrossRef]
27. Ribeiro, M.T.; Singh, S.; Guestrin, C. 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), Melbourne, Australia, 15–20 July 2018; pp. 856–865. [CrossRef]
28. Iyyer, M.; Wieting, J.; Gimpel, K.; Zettlemoyer, L. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv* **2018**, arXiv:1804.06059. [CrossRef]
29. Wang, B.; Xu, C.; Wang, S.; Gan, Z.; Cheng, Y.; Gao, J.; Awadallah, A.H.; Li, B. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. *arXiv* **2021**, arXiv:2111.02840. [CrossRef]
30. Xu, Y.; Wu, B.; Shen, F.; Fan, Y.; Zhang, Y.; Shen, H.T.; Liu, W. Exact adversarial attack to image captioning via structured output learning with latent variables. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4135–4144. [CrossRef]
31. Shayegani, E.; Dong, Y.; Abu-Ghazaleh, N. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv* **2023**, arXiv:2307.14539. [CrossRef]
32. Chokhra, P. Facebook Hateful Meme Dataset. 2020. Available online: https://www.kaggle.com/datasets/parthplc/facebook-hateful-meme-dataset?resource=download (accessed on 10 February 2025).
33. Armijo, C.; Rivas, P. Exploring Visual Embedding Spaces Induced by Vision Transformers for Online Auto Parts Marketplaces. In Proceedings of the AAAI 2025 Workshop on AI for Social Impact: Bridging Innovations in Finance, Social Media, and Crime Prevention at the 39th Annual AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February 2025; pp. 1–8. [CrossRef]
34. Okonkwo, M.; Rivas, P. Assessing the Efficacy of DinoV2-Based Embeddings in Clustering Visual Data from C2C Marketplaces. In Proceedings of the AIR-RES 2025: The 2025 International Conference on the AI Revolution: Research, Ethics, and Society, Las Vegas, NV, USA, 14–16 April 2025; pp. 1–12.
35. Rivas, P.; Yero Salazar, J.; Turek, J.S.; Khanal, B. Few-Label SetFit for C2C Online Marketplace Listings: Multi-Label Classification and Entity Extraction for Identifying Potentially Illicit Ads. In Proceedings of the LatinX in AI Research Workshop at NeurIPS 2025, San Diego, CA, USA, 2 December 2025.
36. Gounev, P.; Bezlov, T. From the economy of deficit to the black-market: Car theft and trafficking in Bulgaria. *Trends Organ. Crime* **2008**, *11*, 410–429. [CrossRef]
37. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]