

Q-MLLM: Vector Quantization for Robust Multimodal Large Language Model Security

Wei Zhao*, Zhe Li*, Yige Li✉, Jun Sun✉,

Singapore Management University

{wzhao, zheli, yigeli, junsun}@smu.edu.sg

Abstract—**Multimodal Large Language Models (MLLMs)** have demonstrated impressive capabilities in cross-modal understanding, but remain vulnerable to adversarial attacks through visual inputs despite robust textual safety mechanisms. These vulnerabilities arise from two core weaknesses: the continuous nature of visual representations, which allows for gradient-based attacks, and the inadequate transfer of text-based safety mechanisms to visual content. We introduce Q-MLLM, a novel architecture that integrates two-level vector quantization to create a discrete bottleneck against adversarial attacks while preserving multimodal reasoning capabilities. By discretizing visual representations at both pixel-patch and semantic levels, Q-MLLM blocks attack pathways and bridges the cross-modal safety alignment gap. Our two-stage training methodology ensures robust learning while maintaining model utility. Experiments demonstrate that Q-MLLM achieves significantly better defense success rate against both jailbreak attacks and toxic image attacks than existing approaches. Notably, Q-MLLM achieves perfect defense success rate (100%) against jailbreak attacks except in one arguable case, while maintaining competitive performance on multiple utility benchmarks with minimal inference overhead. This work establishes vector quantization as an effective defense mechanism for secure multimodal AI systems without requiring expensive safety-specific fine-tuning or detection overhead. Code is available at <https://github.com/QMLLM/Q-MLLM>.

I. INTRODUCTION

The rapid advancements in multimodal large language models (MLLMs) have equipped artificial intelligence systems with impressive capabilities to comprehend, reason, and generate based on both textual and visual modalities [47]. State-of-the-art MLLMs, such as LLaVA [31], Qwen-VL [3], and Flamingo [2], have demonstrated exceptional proficiency in tasks including image understanding, visual reasoning, and multimodal generation. These successes stem from the integration of powerful language models with visual encoders, enabling the fusion of images and textual inputs into unified representations for further reasoning and decision-making processes [14], [31], [55].

Despite these impressive advancements, recent findings reveal that existing MLLMs remain vulnerable to carefully

crafted adversarial inputs and harmful visual content, posing significant safety threats [16], [32], [33], [42]. Specifically, recent studies have identified two types of attacks against MLLMs. First, adversarially perturbed images can bypass the backbone LLM’s safety alignment, compelling them to generate responses that violate ethical guidelines or safety policies. Unlike textual embeddings, which involve discrete tokenization and embedding steps, visual representations within MLLMs are continuous, enabling attackers to introduce imperceptible perturbations optimized via gradient-based techniques [39], [42]. Second, inherent harmful visual content coupled with seemingly benign textual prompts can exploit gaps in cross-modal alignment, rendering the otherwise robust textual safety mechanisms useless in MLLMs. Evaluations across harmful image datasets have consistently demonstrated the vulnerability of current state-of-the-art models (e.g., LLaVA-1.5 and Qwen-VL) to such attacks, with near-zero defense success rate in defense against harmful visual inputs [33], [48].

Existing approaches to mitigate these vulnerabilities largely fall into three categories—safety fine-tuning methods, pre-image detection methods, and post-generation detection methods. Safety fine-tuning methods adapt the internal safety mechanisms via adversarial training or supervised training with additional toxic images [36], [49]. However, these methods are computationally intensive and typically require substantial task-specific datasets [56]. Alternatively, pre-image detection mechanisms—such as LlavaGuard [21] and Safe-CLIP [53]—filter harmful visuals before processing, but often lack sufficient capabilities to defend adversarial perturbation-based jailbreak attacks. Post-generation detection methods like ECSO [18], MLLM-Protector [41], and ETA [12] attempt to identify unsafe outputs after generation, yet pose significant overhead in computational resources and latency, thereby limiting their practical utility [21], [53]. Given the above limitations, there remains a critical need for more effective and computationally efficient approaches to safeguard MLLMs against these two kinds of threats.

In this work, we propose **Q-MLLM**, a novel MLLM architecture that employs two-level vector quantization at the embedded vision extractor of MLLM to introduce discrete bottlenecks in visual feature representations, substantially mitigating adversarial attacks while preserving multimodal reasoning capabilities. Inspired by recent adversarial defense approaches that exploit discretization barriers [17], [27], our method leverages vector quantization to block the gradient paths required for successful adversarial optimization. Specifically, we introduce hierarchical patch-level and semantic-level discretization of visual features, effectively transforming

* These authors contributed to the work equally and should be regarded as co-first authors. ✉ Corresponding authors: Yige Li and Jun Sun

vulnerable continuous embeddings into robust discrete tokens. Furthermore, recognizing that state-of-the-art MLLMs possess strong zero-shot classification capabilities (inherited from pre-trained visual encoders such as CLIP-ViT [43]), our method exploits these inherent competencies to efficiently detect harmful visual inputs by augmenting semantic-level embeddings for enhanced toxicity detection, enabling immediate rejection of harmful requests prior to further processing.

Through comprehensive experiments on established datasets and attack settings—including gradient-based jailbreak attacks such as ImgJP [39] and VAA [42], generation-based jailbreak attacks such as FigStep [16] and MM-SafetyBench [33], and toxic image datasets including HOD [19] and ToViLaG [48]—we demonstrate that Q-MLLM consistently achieves substantial improvements in safety: 98.4% average Defense Success Rate (DSR) against jailbreak attacks and 75.9% against toxic image attacks. These results significantly surpass existing defenses such as CAT [49] and SafeCLIP [53], highlighting the comprehensive protective capability of our dual quantization and enhanced semantic detection mechanisms. Importantly, evaluations on standard vision-language tasks reveal minimal trade-offs in task utility, with only minor degradations compared to baseline models, and a notably low false positive rate that preserves practical applicability.

In summary, the key contributions of our work are:

- **Novel Quantization-based Defense:** We introduce Q-MLLM, the first unified architecture to robustly and simultaneously defend against visual modality vulnerabilities—addressing both adversarial perturbations and inherent toxic visual content.
- **Computationally Efficient Safety Detection:** By employing enhanced semantic alignment for detection of toxic visual inputs, Q-MLLM achieves high accuracy and minimal inference overhead compared to state-of-the-art pre-image and post-generation detection frameworks.
- **Empirical Validation:** Extensive evaluations against multiple attack scenarios demonstrate our approach outperforms existing baselines, achieving significant improvements in defense success rates for jailbreak (up to 98.4%) and image (up to 75.9%) attack scenarios, while maintaining competitive task utility measured on established vision-language benchmarks.

We believe that our findings and methods offer valuable insights and direction toward building safer, more reliable multimodal systems, laying the groundwork for future research into ensuring comprehensive cross-modal safety alignment.

II. PRELIMINARIES

In this section, we begin by explaining how state-of-the-art Multimodal Large Language Models (MLLMs) operate, with a focus on the role of continuous visual representations in their functioning. Next, we formalize our threat model, detailing two distinct attack strategies that exploit vulnerabilities in these systems.

A. Multimodal Large Language Models

Modern MLLMs integrate both visual and textual modalities to perform tasks such as image understanding, visual reasoning, and response generation. A typical MLLM architecture consists of the following key components:

- 1) **Visual Feature Extraction.** Given an input image $X_{\text{img}} \in \mathbb{R}^{H \times W \times C}$, a visual encoder F_v (often based on transformer architectures, e.g., CLIP-ViT) computes both pixel-level and semantic-level representations:

$$F_v(X_{\text{img}}) = Z = \{z_{\text{cls}}^0, Z_v^{1:N}\},$$

where the encoder output $Z \in \mathbb{R}^{(N+1) \times d_v}$ contains a global semantic embedding $z_{\text{cls}}^0 \in \mathbb{R}^{d_v}$ and N patch-level features $Z_v^{1:N} \in \mathbb{R}^{N \times d_v}$.

- 2) **Cross-modal Projection.** To bridge visual and language modalities, a projection module F_p maps the visual features into the language embedding space:

$$H_v = F_p(Z_v^{1:N}),$$

resulting in aligned visual features $H_v \in \mathbb{R}^{N \times d_h}$ that can be fused with textual embeddings. And normally, z_{cls}^0 is discarded during this process.

- 3) **Multimodal Fusion and Generation.** The input text X_t is tokenized and embedded to obtain $H_t \in \mathbb{R}^{L \times d_h}$. The fusion of both modalities is achieved by concatenating the visual and textual embeddings:

$$H_{\text{fusion}} = \text{Concat}(H_v, H_t).$$

This combined representation is then processed by the large language generation module F_{LLM} to produce the output:

$$y = F_{\text{LLM}}(H_{\text{fusion}}).$$

It is important to note that the continuous nature of visual feature embeddings (H_v) can be a source of vulnerability. Unlike the discretized token embeddings H_t used for textual inputs, the continuous representation is more amenable to gradient-based adversarial optimization. In particular, the absence of a discretization bottleneck (such as one achieved via vector quantization) opens the door to gradient-based manipulations, which attackers can exploit to trigger unsafe behaviors.

B. Threat Model

Target Model We consider state-of-the-art MLLMs that integrate visual and textual modalities through the architecture described above. These models typically employ safety mechanisms designed primarily for text inputs but may have insufficient safeguards for the visual modality. The primary vulnerability lies in the continuous nature of visual representations within the model’s intermediate layers. These representations are particularly vulnerable for two key reasons: (1) they allow adversaries to introduce adversarial perturbations through gradient-based optimization, and (2) they exhibit a significant safety alignment gap whereby text-based safety alignment fails to adequately transfer to visual content, leaving the model susceptible to generating unsafe responses when confronted with inherently harmful images.

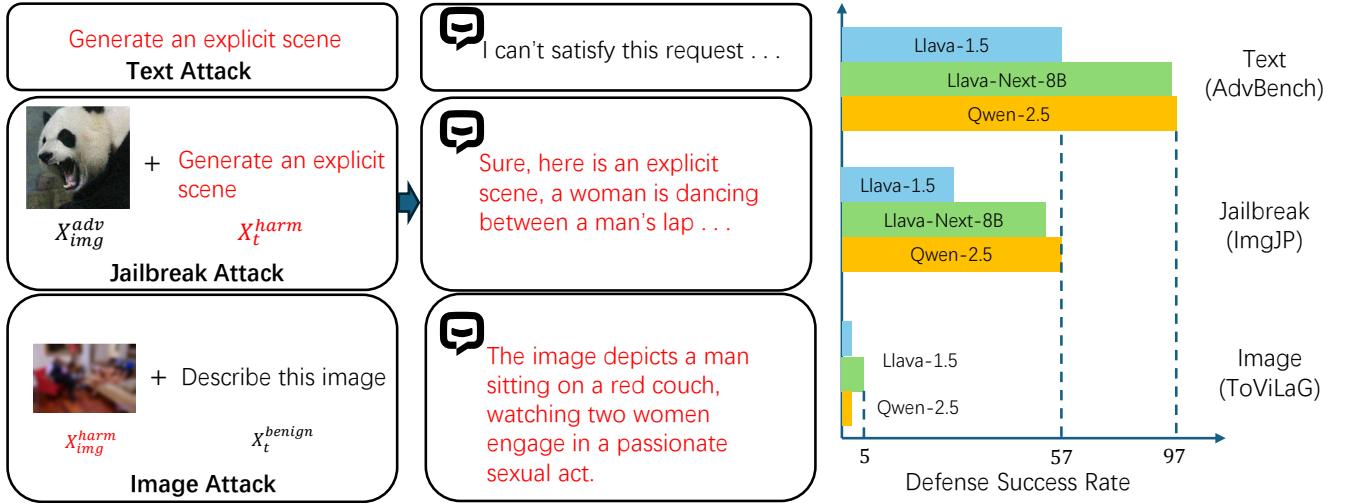


Fig. 1: Threat model for Multimodal Large Language Models (MLLMs), demonstrating two types of attacks: (1) jailbreak attacks combining adversarially perturbed images $X_{\text{img}}^{\text{adv}}$ with harmful text X_t^{harm} , and (2) image-based attacks using harmful images $X_{\text{img}}^{\text{harm}}$ with benign prompts X_t^{benign} . Defense success rates across different MLLMs reveal significant vulnerabilities in handling visual and multimodal threats.

Adversary Capabilities and Objectives The adversary operates under a white-box setting with complete access to the model architecture, parameters, and gradient information. This access enables the adversary to craft adversarial perturbations using gradient-based optimization techniques. The adversary’s primary objective is to circumvent the model’s safety alignment to generate harmful, offensive, or prohibited content. We assume the adversary can observe model outputs, cannot directly modify the model parameters or training data, and possesses the technical capability to manipulate input images either through calculated perturbations or by selecting specific harmful content.

Problem Definition Our threat model focuses on two primary vulnerabilities in existing MLLM architectures demonstrated in Figure 1. The first vulnerability concerns the inherent susceptibility of continuous image representations to adversarial perturbations. Unlike text embeddings, which undergo tokenization and discretization, visual features remain continuous throughout the processing pipeline, making them fundamentally more susceptible to adversarial manipulations. In a typical jailbreak attack, an attacker takes a benign image $X_{\text{img}}^{\text{benign}}$ and applies an imperceptible perturbation δ :

$$X_{\text{img}}^{\text{adv}} = X_{\text{img}}^{\text{benign}} + \delta, \quad \text{with } \|\delta\|_p \leq \epsilon,$$

where ϵ is a small perturbation budget and $\|\cdot\|_p$ denotes an L_p -norm. When this perturbed image is processed, it generates a compromised visual representation H_v^{adv} . Meanwhile, when a harmful text prompt X_t^{harm} is processed alone, it produces a text representation H_t^{harm} that would normally trigger the model’s safety mechanisms in F_{LLM} , resulting in a rejection of the harmful request. However, when H_v^{adv} is fused with H_t^{harm} :

$$y^{\text{adv}} = F_{\text{MLLM}}(X_{\text{img}}^{\text{adv}}, X_t^{\text{harm}}) = F_{\text{LLM}}(H_v^{\text{adv}} \oplus H_t^{\text{harm}}),$$

the adversarially perturbed visual representation effectively bypasses the safety mechanisms, causing the model to generate

the corresponding harmful content. This occurs because the perturbation in H_v^{adv} is specifically optimized to neutralize or mislead the safety mechanisms when combined with H_t^{harm} .

The second vulnerability arises from a fundamental misalignment between visual and textual modalities. This issue highlights the inability to adequately transfer text-based safety mechanisms to the visual domain due to inherent differences in representation across modalities. In this scenario, an attacker feeds the model with an inherently harmful image $X_{\text{img}}^{\text{harm}}$ paired with a benign text prompt:

$$y^{\text{harm}} = F_{\text{MLLM}}(X_{\text{img}}^{\text{harm}}, X_t^{\text{benign}}).$$

While the language model has robust safety alignment for text-based harmful prompts, this safety alignment cannot be transferred to the vision modality, resulting in the model inadvertently producing unsafe outputs even with neutral prompts. In Section IV, we demonstrated that state-of-the-art MLLMs such as Llava-1.5 [31], Llava-Next-8B [30] and Qwen-2.5 [4] have almost zero defense against different types of harmful images.

The fundamental security challenge is therefore twofold: (1) protecting against attacks that exploit the continuous nature of visual representations, and (2) addressing the cross-modal alignment gap that prevents complete transfer of text-based safety mechanisms to visual inputs. Our work proposes to address the first vulnerability by introducing vector quantization during the visual feature extraction process, creating a discrete bottleneck that significantly impairs gradient-based attack methods. For the second vulnerability, we develop enhanced cross-modal safety alignment techniques that better bridge the gap between visual and textual safety mechanisms.

III. METHOD

In this section, we introduce **Q-MLLM**, a MLLM architecture integrating two-level vector quantization to enhance

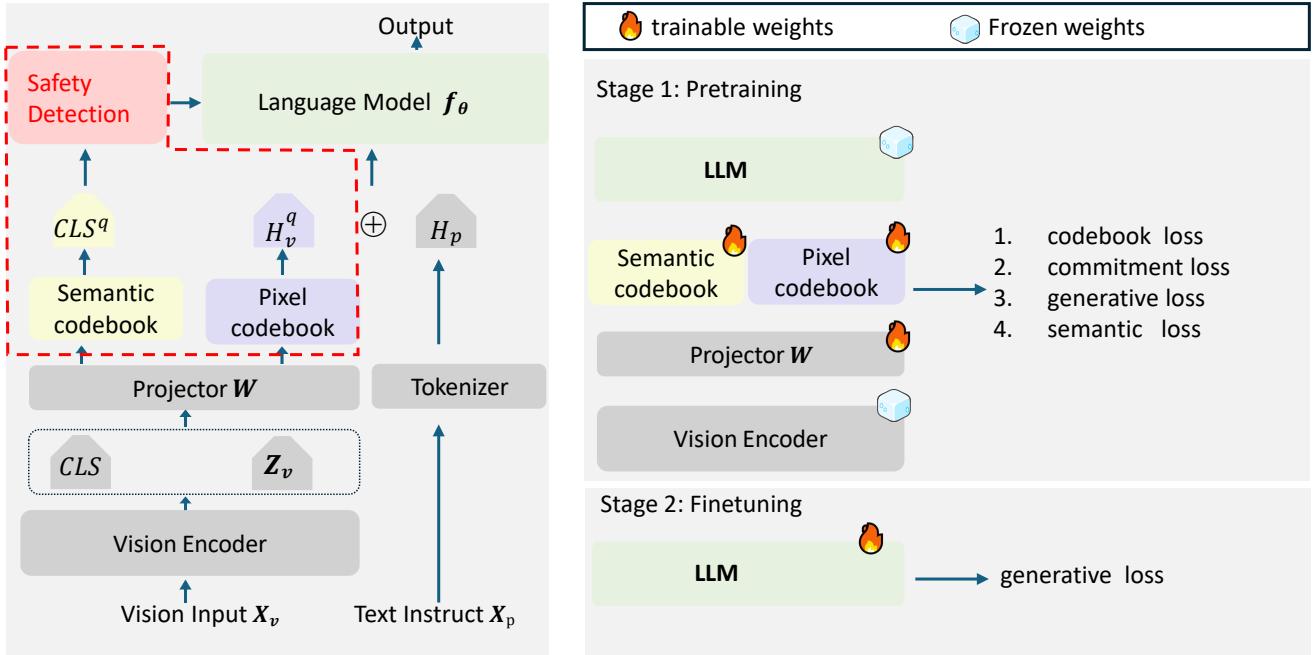


Fig. 2: Overview of Q-MLLM architecture and training methodology. Left: Q-MLLM employs hierarchical vector quantization on vision encoder representations through semantic and patch-level codebooks, generating discrete tokens for enhanced multimodal integration robustness. Right: The training pipeline comprises two distinct phases—Stage 1 involves codebook and projector pretraining with multi-objective loss functions while maintaining frozen vision encoder and LLM parameters; Stage 2 performs LLM fine-tuning through generative loss optimization.

resilience against visual-based adversarial manipulation. We first describe the proposed modifications to standard MLLM architectures, particularly highlighting our hierarchical vector quantization. Next, we detail the training strategy comprising pretraining and fine-tuning stages.

A. Q-MLLM: Vector-Quantized Multimodal Architecture

Continuous visual representations in existing MLLM remain susceptible to gradient-based attacks. To address this vulnerability, Q-MLLM discretizes visual representations at two levels: pixel-patch (spatial) level and global-semantic level. This two-level approach creates a hierarchical discretization bottleneck that significantly enhances robustness against adversarial manipulations while preserving multimodal reasoning capabilities.

1) Two-Level Vector Quantization

The two-level vector quantization mechanism transforms continuous visual embeddings into discrete tokens, analogous to how text inputs are tokenized in language models.

First, hierarchical features are extracted from the input image. Given an image \mathbf{X}_{img} , the vision encoder F_v computes both global semantic and patch-level embeddings:

$$\{\mathbf{z}_{\text{cls}}^0, \mathbf{Z}_v^{1:N}\} = F_v(\mathbf{X}_{\text{img}}), \quad (1)$$

where $\mathbf{Z}_v^{1:N} \in \mathbb{R}^{N \times d_v}$ represents N patch embeddings, and $\mathbf{z}_{\text{cls}}^0 \in \mathbb{R}^{d_v}$ captures the global semantic representation of the image. For a CLIP-patch14-336 encoder, $d_v = 1024$ and $N =$

576. This step extracts hierarchical vision features from the image.

Next, these features are projected into a shared latent space aligned with the textual embedding space. A linear projection layer F_h transforms both patch and semantic embeddings:

$$\{\mathbf{h}_{\text{cls}}, \mathbf{H}_v\} = F_h(\{\mathbf{z}_{\text{cls}}^0, \mathbf{Z}_v^{1:N}\}), \quad (2)$$

where $\mathbf{H}_v \in \mathbb{R}^{N \times d_h}$ represents the pixel-level patch embeddings, and $\mathbf{h}_{\text{cls}} \in \mathbb{R}^{d_h}$ is the global semantic embedding. Here, the visual features are mapped into the latent space for alignment with textual embeddings.

To discretize these embeddings, Q-MLLM applies vector quantization (VQ). For the global semantic embedding \mathbf{h}_{cls} , the nearest vector from a semantic codebook $\mathcal{C}_{\text{cls}} \in \mathbb{R}^{K \times d_h}$ is selected:

$$\tilde{\mathbf{h}}_{\text{cls}} = \mathbf{e}_k, \quad k = \arg \min_i \|\mathbf{h}_{\text{cls}} - \mathbf{e}_i\|_2^2, \quad (3)$$

where \mathbf{e}_i represents the i -th vector in the codebook. This quantizes the global semantic embedding by mapping it to the nearest vector in the semantic codebook.

Similarly, each patch embedding \mathbf{H}_v^j (for $j = 1, \dots, N$) is quantized using a separate patch-level codebook $\mathcal{C}_{\text{patch}} \in \mathbb{R}^{P \times d_h}$:

$$\tilde{\mathbf{H}}_v^j = \mathbf{e}_{k_j}, \quad k_j = \arg \min_i \|\mathbf{H}_v^j - \mathbf{e}_i\|_2^2. \quad (4)$$

This step quantizes the patch-level embeddings by mapping each to the nearest vector in the patch-level codebook.

This dual quantization solution generates discrete latent representations that inherently resist gradient manipulation while maintaining spatial and semantic coherence necessary for multimodal reasoning.

2) Safety Signal Detection

The safety detection mechanism in Q-MLLM leverages the quantized global semantic embedding $\tilde{\mathbf{h}}_{\text{cls}}$ for identifying harmful content. This process consists of two phases: constructing a safety mapping and detecting violations during inference as demonstrated in Algorithm 1.

In the mapping phase, Q-MLLM generates a compact dataset \mathcal{D}_{map} containing representative examples across multiple toxic categories (e.g., 50 images per category) and neutral images (e.g., 500 images). Note that \mathcal{D}_{map} only serves as a lightweight calibration step to identify which indices represent toxic content and does not participate in the training process (details of \mathcal{D}_{map} can be found in Appendix F). For each image $I_i \in \mathcal{D}_{\text{map}}$, the semantic embedding $\mathbf{h}_{\text{cls}}^i$ is extracted, and the nearest codebook vector is identified:

$$k_i = \arg \min_j |\mathbf{h}_{\text{cls}}^i - \mathbf{e}_j|_2^2. \quad (5)$$

Each image is assigned a codebook index by mapping its semantic embedding to the closest codeword.

The distribution of toxic categories across codebook indices is tracked, and a mapping function $M(k)$ is defined. For each codebook index k , if the dominant toxic category exceeds a threshold τ , the index is classified as belonging to that category; otherwise, it is labeled neutral:

$$M(k) = \begin{cases} \arg \max_c P(c|k), & \text{if } \max_c P(c|k) > \tau, \\ \text{neutral}, & \text{otherwise.} \end{cases} \quad (6)$$

This mapping associates each codebook index with a category or marks it as neutral based on the proportions of toxic categories.

During inference, the quantized global semantic embedding $\tilde{\mathbf{h}}_{\text{cls}}$ is processed through this mapping function:

$$\hat{y} = M(\arg \min_j |\mathbf{h}_{\text{cls}} - \mathbf{e}_j|_2^2). \quad (7)$$

The input is classified as toxic or neutral based on the safety mapping.

If \hat{y} corresponds to a toxic category, the model rejects the input before proceeding, ensuring robust safety with minimal computational overhead.

3) Multimodal Fusion and Generation

For non-toxic inputs, the visual and textual embeddings are fused for further processing. The quantized patch embeddings $\tilde{\mathbf{H}}_v$ are concatenated with the textual embeddings \mathbf{H}_{text} to form a multimodal input sequence:

$$\mathbf{H}_{\text{fusion}} = \text{Concat}(\tilde{\mathbf{H}}_v, \mathbf{H}_{\text{text}}). \quad (8)$$

The fused multimodal representation combines quantized visual embeddings with textual embeddings.

Algorithm 1 Safety Signal Detection in Q-MLLM

Require: Safety mapping dataset \mathcal{D}_{map} , Toxic mapping threshold τ , Codebook $\mathbf{C}_{\text{cls}} = \{\mathbf{e}_j\}_{j=1}^K$, Vision Encoder F_v , Modal Projector F_h

1: **Phase 1: Safety Mapping Construction**

- 2: Initialize dictionary D where each key k maps to an empty category counter
- 3: For each image $I_i \in \mathcal{D}_{\text{map}}$ with category c_i :
- 4: Extract semantic embedding $\mathbf{h}_{\text{cls}}^i = F_h(F_v(I_i))$
- 5: Find nearest codeword: $k_i = \arg \min_j |\mathbf{h}_{\text{cls}}^i - \mathbf{e}_j|_2^2$
- 6: Update dictionary: $D[k_i][c_i] += 1$
- 7: Initialize mapping function M as empty dictionary
- 8: For each codebook index k in D :
- 9: Calculate total count: $\text{total} = \sum_c D[k][c]$
- 10: For each category c in $D[k]$: $P(c|k) = D[k][c]/\text{total}$
- 11: Find dominant category: $c_{\text{dom}} = \arg \max_c P(c|k)$
- 12: If $P(c_{\text{dom}}|k) > \tau$: $M[k] = c_{\text{dom}}$
- 13: Else: $M[k] = \text{neutral}$

14: **Phase 2: Inference-time Safety Detection**

- 15: **function** DetectSafety(I)
- 16: Extract semantic embedding: $\mathbf{h}_{\text{cls}} = F_h(F_v(I))$
- 17: Find nearest codeword index: $k = \arg \min_j |\mathbf{h}_{\text{cls}} - \mathbf{e}_j|_2^2$
- 18: **Return** safety prediction: $\hat{y} = M[k]$

This fused representation is processed by the language model F_{LLM} to generate the output:

$$\mathbf{y} = F_{\text{LLM}}(\mathbf{H}_{\text{fusion}}). \quad (9)$$

The discrete visual tokens ensure robust defense against adversarial attacks, while the fusion mechanism maintains high multimodal reasoning performance.

As a result, the quantized visual representations function explicitly as discrete tokens, reducing susceptibility to adversarial attacks while maintaining the integrity of multimodal reasoning. The complete inference pipeline of Q-MLLM is outlined in Algorithm 2.

B. Training Q-MLLM

We adopt a carefully structured two-stage training approach for the Q-MLLM architecture to ensure robust multimodal representation learning while maintaining resilience to adversarial manipulation. Our strategy comprises (1) a pretraining stage targeting the visual projection and dual-level vector quantization modules, and (2) a fine-tuning stage focused on enhancing multimodal reasoning and generation capabilities under discretized input constraints.

Training Dataset. We utilize the publicly available LLaVA training dataset containing approximately 558K image-text pairs for pretraining and 665K multimodal conversation examples for fine-tuning.

- **Pretraining data:** Each sample is a single image-text pair $(\mathbf{X}_{\text{image}}, \mathbf{X}_{\text{inst}} \cdot \mathbf{X}_{\text{caption}})$, where the input prompt instructs the model to describe the image, and the target is the corresponding caption.
- **Instruction-tuning data:** Each sample follows a multi-turn format: $(\mathbf{X}_{\text{image}}, \mathbf{X}_{\text{inst}}^1 \cdot \mathbf{X}_{\text{r}}^1, \dots, \mathbf{X}_{\text{inst}}^q \cdot \mathbf{X}_{\text{r}}^q)$, where $\mathbf{X}_{\text{image}}$ is the input image, and each $(\mathbf{X}_{\text{inst}}^i, \mathbf{X}_{\text{r}}^i)$ is an instruction-response pair in a dialogue format.

Algorithm 2 Q-MLLM: Inference Process

Require: Image input \mathbf{X}_{img} , Text prompt \mathbf{X}_{text} , Vision encoder F_v , Projection layer F_h , Semantic codebook $\mathcal{C}_{\text{cls}} \in \mathbb{R}^{K \times d_h}$, Patch codebook $\mathcal{C}_{\text{patch}} \in \mathbb{R}^{P \times d_h}$, Safety mapping function M , Language model F_{LLM}

- 1: **Phase 1: Visual Feature Extraction and Projection**
- 2: Extract hierarchical features: $\mathbf{z}_{\text{cls}}^0, \mathbf{Z}_v^{1:N} = F_v(\mathbf{X}_{\text{img}})$
- 3: Project to language model dimension: $\mathbf{h}_{\text{cls}}, \mathbf{H}_v = F_h(\mathbf{z}_{\text{cls}}^0, \mathbf{Z}_v^{1:N})$
- 4: **Phase 2: Dual-Level Vector Quantization**
- 5: **Global-semantic quantization**
- 6: Find nearest semantic codeword: $k_{\text{cls}} = \arg \min_i |\mathbf{h}_{\text{cls}} - \mathbf{e}_i|^2$ where $\mathbf{e}_i \in \mathcal{C}_{\text{cls}}$
- 7: Quantize semantic vector: $\tilde{\mathbf{h}}_{\text{cls}} = \mathbf{e}_{k_{\text{cls}}}$
- 8: **Pixel-patch level quantization**
- 9: **for** $j = 1$ to N **do**
- 10: Find nearest patch codeword: $k_j = \arg \min_i |\mathbf{H}_v^j - \mathbf{e}_i|^2$ where $\mathbf{e}_i \in \mathcal{C}_{\text{patch}}$
- 11: Quantize patch vector: $\tilde{\mathbf{H}}_v^j = \mathbf{e}_{k_j}$
- 12: **end for**
- 13: **Phase 3: Safety Assessment**
- 14: Apply safety mapping: $\hat{y} = M(k_{\text{cls}})$
- 15: **if** $\hat{y} \neq \text{neutral}$ **then**
- 16: **return** Safety warning response
- 17: **end if**
- 18: **Phase 4: Multimodal Fusion and Text Generation**
- 19: Embed text input: $\mathbf{H}_{\text{text}} = \text{TextEmbedding}(\mathbf{X}_{\text{text}})$
- 20: Concatenate modalities: $\mathbf{H}_{\text{fusion}} = \text{Concat}(\tilde{\mathbf{H}}_v, \mathbf{H}_{\text{text}})$
- 21: Generate output text: $\mathbf{y} = F_{\text{LLM}}(\mathbf{H}_{\text{fusion}})$
- 22: **return** \mathbf{y}

1) Pretraining Phase. During pretraining, we freeze both the vision encoder and the language model, training only the visual projection F_h and the associated vector quantization codebooks. This selective training strategy serves two essential purposes. First, it maintains the pretrained knowledge in both the vision and language components while adapting only the components necessary for our defense mechanism. Second, the frozen foundation ensures stable performance on vision tasks, while the trained projection and codebooks develop effective quantization that disrupts adversarial attacks and improves toxic content detection.

Vector Quantization Loss. We implement a gradient approximation technique to enable backpropagation through the otherwise non-differentiable discrete codebook selection process. This approach allows gradient-based optimization of both spatial-patch and semantic-level vector quantization components. Specifically, the vector quantization loss consists of two standard terms:

The *Codebook Loss*, which optimizes the codebook vector towards visual encoder outputs, defined as:

$$\mathcal{L}_{\text{codebook}} = |\text{VQ}(\mathbf{x}) - \text{sg}[\mathbf{x}]|^2_2, \quad (10)$$

where $\text{VQ}(\mathbf{x})$ represents the vector quantization process described in previous subsection.

The *Commitment Loss*, ensuring that the visual projections commit to selected codebook vector:

$$\mathcal{L}_{\text{commit}} = |\mathbf{x} - \text{sg}[\text{VQ}(\mathbf{x})]|^2_2, \quad (11)$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation. Thus, the

total quantization objective integrates these two elements as:

$$\mathcal{L}_{\text{vq}} = \mathcal{L}_{\text{codebook}} + \lambda_{\text{commit}} \mathcal{L}_{\text{commit}}. \quad (12)$$

Semantic Alignment Loss. To ensure Q-MLLM can effectively detect and defend against toxic visual content, we introduce a semantic alignment loss designed explicitly to optimize the quantized semantic embedding $\tilde{\mathbf{h}}_{\text{cls}}$ for enhanced global-semantic representations.

Specifically, this loss minimizes the distance between the quantized semantic embedding and the caption’s comprehensive representation derived from the language model’s final layer, capturing the image’s global semantic information through $\mathbf{H}_{\text{caption}}$ obtained during pretraining from image-caption pairs:

$$\mathcal{L}_{\text{semantic}} = |\tilde{\mathbf{h}}_{\text{cls}} - \mathbf{H}_{\text{caption}}|^2_2. \quad (13)$$

This objective aligns multimodal semantic representations without directly coupling them to generation processes, thus providing a reliable latent vector for downstream safety detection tasks.

Generative Loss. The generative loss is a standard autoregressive language modeling objective, defined as the negative log-likelihood of generating target textual tokens y_t conditioned on the discretized multimodal embeddings $\mathbf{H}_{\text{fusion}}$ and previous tokens:

$$\mathcal{L}_{\text{generative}} = - \sum_{t=1}^T \log p(y_t | \mathbf{H}_{\text{fusion}}, y_{<t}). \quad (14)$$

Combined Pretraining Objective. Overall, the composite pretraining loss integrates all the losses described above:

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{\text{generative}} + \lambda_1 (\mathcal{L}_{\text{vq-patch}} + \mathcal{L}_{\text{vq-cls}}) + \lambda_2 \mathcal{L}_{\text{semantic}}, \quad (15)$$

where $\mathcal{L}_{\text{vq-patch}}$ and $\mathcal{L}_{\text{vq-cls}}$ denote patch-level and semantic-level VQ losses, respectively. This integrated loss formulation guides learning across both patch-level and semantic-level representations, while maintaining the security advantages of discrete vector quantization. The resulting architecture provides inherent defense against adversarial manipulation by creating a non-differentiable bottleneck that fundamentally disrupts gradient-based attacks rather than relying on pattern-specific detection mechanisms.

2) Fine-Tuning Phase. In the fine-tuning stage, we freeze the visual projection and vector quantization parameters, focusing optimization on the pretrained language model using multimodal conversation data.

The fine-tuning objective is solely based upon standard conversational generative loss:

$$\mathcal{L}_{\text{fine-tune}} = \mathcal{L}_{\text{lm}}, \quad (16)$$

calculated on conversational response tokens.

The rationale behind freezing visual quantization components is to preserve the security guarantees conferred by discrete visual encoding. By preventing further updates to visual projections post-pretraining, we ensure stability in the discrete encoding mechanism. Consequently, the language model must

implicitly adapt its reasoning exclusively through discrete multimodal embeddings, thereby inherently reinforcing security robustness, while enhancing multimodal dialogue generation performance.

IV. EXPERIMENTAL EVALUATION

In this section, we comprehensively evaluate Q-MLLM with various experimental settings. Particularly, we would like to answer the following research questions:

- **RQ1:** What is the safety and utility performance of Q-MLLM?
- **RQ2:** How does Q-MLLM defend against toxic image attacks?
- **RQ3:** How does Q-MLLM defend against jailbreak attacks?

While RQ1 aims to measure the safety and utility achieved by Q-MLLM, RQ2 and RQ3 aim to conduct in-depth analysis to pinpoint the underlying reasons for Q-MLLM’s enhanced safety against two different kinds of attacks.

A. Experiment Setup

Image Attack Following the settings in [50], we adopt the toxic datasets comprising HOD [19], which contains 10631 toxic images depicting alcohol, cigarette, gun, insulting gesture, bloody scenes and knife, and ToViLaG [48], which includes 5600 toxic images representing pornographic content. We adopt the standard safety testing method described in [48], presenting toxic images to models with text instructions to describe their content.

Jailbreak Attack For jailbreak attacks, we first evaluate the recently proposed ImgJP Attack method [39] which applies image perturbations to induce affirmative responses to toxic queries. Following their original setup, we assess the performance based on the 520 harmful prompts from AdvBench [57]. We also compare against the Visual Adversarial Attack (VAA) [42], which directly optimizes image noise to maximize the likelihood of generating toxic text. For VAA, we follow their original setup and optimize on 40 harmful prompts. Both ImgJP and VAA represent white-box attack methods that require gradient information for adversarial optimization. Finally, we evaluate FigStep [16] and MM-SafetyBench [33], which directly embed adversarial harmful content into images. FigStep incorporates prohibited content through typography (adding text directly into images), while MM-SafetyBench visualizes harmful content by using diffusion models to generate images based on harmful prompts. Both FigStep and MM-SafetyBench constitute black-box attack methods designed as model-agnostic approaches, attempting to bypass safety alignment by directly embedding harmful features in the image rather than relying on adversarial noise perturbations.

Safety Evaluation Metric We employ the defence success rates (DSR) as our primary metric, which indicates whether the model refuses to produce toxic responses when presented with toxic input. A higher DSR indicates that the model is more inclined to reject requests for generating descriptions of toxic content. For neutral images, we similarly follow the

standard safety testing procedure; and report the false positive rate (FPR), defined as the number of images misclassified as non-neutral divided by the total number of neutral images. GPT-4o [22] is used to determine whether the responses generated by the model are toxic, thereby facilitating the evaluation of both DSR and FPR. Detailed prompt templates are provided in Appendix A.

Utility Evaluation Metric We apply ScienceQA [35] to measure scientific reasoning capabilities, using its 21k multimodal multiple-choice questions across diverse science topics. Following LLaVA [31], we evaluate zero-shot accuracy on the image subset. For hallucination assessment, we employ POPE [24], which tests models across three COCO-derived splits (random, common, and adversarial), reporting F1 scores for each condition. Note that these benchmarks are widely adopted in the literature [8], [30], [50]. Additional details about these benchmarks and benchmark result on MMvet can be found in Appendix C.

Vanila MLLMs The open-source MLLMs and LLMs employed in our experiments include: LLaVA-1.5 [31] with its base LLM Vicuna-7B-v1.5 [10], Llava-next-8B [30] with its base LLM Llama-3-8B-Instruct [15], Qwen2.5-VL [4] with its base LLM Qwen2.5-7B-Chat [3].

Q-MLLM Setup We implemented our Q-MLLM following similar settings of Llava-1.5-7B. For Q-MLLM-7B, the baseline LLM is Llama2-7B [10] and for Q-MLLM-8B, the baseline LLM is Llama-3-8B [15]. Additionally, we have implemented Q-MLLM based on the InstructBlip-7B architecture, with detailed implementation procedures and comprehensive result analysis provided in the Appendix D. For both Q-MLLMs they share the same settings as below: CLIP-encoder CLIP-336ppx-14patch [43], semantic codebook size $K = 128$, pixel codebook size $P = 16000$, commitment weight $\lambda_{commit} = 0.25$, overall vq-loss weight $\lambda_1 = 0.5$, semantic loss weight $\lambda_2 = 0.1$ and toxic mapping threshold $\tau = 0.6$. We conducted the training of our Q-MLLM using one H100nvl GPU with float16 precision, employing a batch size of 8 for the pretraining phase and reducing to a batch size of 2 during the fine-tuning stage. For all inference and evaluation procedures, we utilized the same H100nvl hardware but switched to full float32 precision to ensure maximum accuracy in our experimental results.

Defense Baseline We compare our approach with a comprehensive set of baselines, as illustrated in Table I, which categorizes different methods based on their capabilities to defend against image attacks and jailbreak attacks. A checkmark (✓) indicates the method’s effectiveness against the corresponding attack type, while a cross (✗) indicates limited or no defense capability. We implemented all defense baselines using LLaVA-1.5. For R2D2 [36] and CAT [49], we first fine-tuned the LLM decoder with these methods before connecting it to the visual encoder and cross-modal adapter. These methods focus on defending against jailbreak attacks. TGA [50] applies a novel vision-language alignment training method with Llava-1.5 architecture and training data for defending against toxic image. LlavaGuard [21] and SafeCLIP [53] use pre-image detection to filter out toxic visual content before the model processes it, making them effective against image attacks. ECSO [18], MLLM-Procter [41], and ETA [12] use post-

TABLE I: Baseline Defense Methods.

Defense baseline	Image	Jailbreak	Description
R2D2 [36]	✗	✓	Robustness finetuning
CAT [49]	✗	✓	Robustness finetuning
TGA [50]	✓	✗	Robustness finetuning
LlavaGuard [21]	✓	✗	Pre-Image Detection
SafeCLIP [53]	✓	✗	Pre-Image Detection
ECSO [18]	✓	✓	Post-Generation Detection
MLLM-Procter [41]	✓	✓	Post-Generation Detection
ETA [12]	✓	✓	Post-Generation Detection

generation detection to identify harmful content after the model generates a response, providing protection against both types of attacks. Since our Q-MLLM-7B and all defense baselines share the similar settings of LLaVA-1.5, our comparison focuses specifically on comparing their safety performance.

B. RQ1: How effective is the safety and utility performance of Q-MLLM

Defense Against Jailbreak Attacks Defense results against jailbreak attacks are summarized in Table II. Our analysis reveals that vanilla MLLMs retain certain defensive capabilities against jailbreak attacks due to their underlying text-based safety alignment provided by the backbone LLM.

Among jailbreak defense baselines, CAT demonstrates strong performance against image perturbation-based attacks, achieving 83.1% DSR against ImgJP and 95.0% DSR against VAA. However, CAT exhibits decreased effectiveness against attacks like Figstep and MM-SafetyBench that embed harmful content directly into images rather than utilizing adversarial perturbations.

Post-detection methods exhibit comparable performance against jailbreak attacks, with MLLM-Protector (91.7% average DSR) and ETA (92.1% average DSR) significantly outperforming ECSO (79.7% average DSR). The superior performance stems from their utilization of specially trained harmful text detectors, whereas ECSO relies on the MLLM itself for detection. Nevertheless, these more effective detectors necessitate additional GPU memory to load specialized harmful classifier models during response generation.

Our Q-MLLM method demonstrates significantly improved robustness against all jailbreak attacks, achieving an exceptional average DSR of 98.4%. Against ImgJP attacks, Q-MLLM-7B attains perfect 100% DSR across all perturbation levels. This effectiveness arises from our image feature quantization process, which disrupts the gradient-based adversarial perturbation by introducing a stop-gradient operation during backpropagation. By enforcing a discretization bottleneck, the quantization maps perturbations into a finite codebook space, effectively neutralizing such attacks. For more sophisticated attacks like VAA (97.5% DSR), Q-MLLM’s quantization approach significantly constrains the adversarial optimization process by preventing attackers from establishing effective perturbation paths through the non-differentiable barrier created by our vector quantization mechanism. When evaluating on FigStep (96.6% DSR) and MM-SafetyBench (96.5% DSR), which embed harmful content directly into images, the quantization process degrades the fidelity of harmful content through discrete mapping, thus attenuating its capacity to

trigger unsafe responses. The quantized features inherently resist the transmission of adversarial semantics, resulting in superior defense performance compared to ETA (improvement of 6.3%), MLLM-Protector (improvement of 6.7%), and CAT (improvement of 9.7%). Q-MLLM-8B exhibited comparable results, confirming our method’s effectiveness across different backbone LLMs.

Defense Against Toxic Image Attacks Defense results against image attacks are presented in Table III. Current state-of-the-art MLLMs without vision-safety alignment demonstrate negligible defense capability against harmful images (approximately 1.0-1.3% average DSR). While aligned for text safety, these models lack specific visual safety alignment, consistently generating inappropriate content when presented with toxic visual input.

Among image attack defenses, TGA achieves limited performance (21.2% average DSR), relying solely on vision-language alignment training without harmful data fine-tuning. Pre-image detection methods including LlavaGuard (49.1% average DSR) and SafeCLIP (66.8% average DSR) demonstrate stronger results, with SafeCLIP exhibiting superior performance through the application of category-specific harmful image descriptions on CLS token. Meanwhile, LlavaGuard exhibits superior defense performance on Insulting Gesture toxic category owing to its fine-tuning for specific category detection. Post-generation detection methods such as ECSO, MLLM-Protector, and ETA show comparable overall performance (52-55% average DSR), though effectiveness varies across toxic categories.

Our Q-MLLM-7B achieves superior performance against toxic image attacks with an average DSR of 75.9%, outperforming the next best method (SafeCLIP) by 9.1%. This improvement derives from our enhanced CLS token detection mechanism. While SafeCLIP utilizes the CLS token from CLIP’s original pretraining process, Q-MLLM further enhances CLS token by aligning them with captions during MLLM training. This additional alignment stage significantly improves the classification efficacy of these tokens for toxic content detection. When detected potentially harmful content with the enhanced CLS token, the system immediately issues a refusal response, bypassing further processing of the visual content. This direct rejection mechanism enables effective filtering across diverse toxic categories, resulting in 92.3% DSR for pornographic category, 76.2% for alcohol imagery, and over 80% for weapons (guns and knives). The superior detection performance validates the effectiveness of our enhanced CLS token alignment and direct rejection strategy. Q-MLLM-8B exhibited comparable safety performance on both types of attacks, confirming our method’s effectiveness across different backbone LLMs.

Performance on Vision-Language Benchmarks To evaluate whether our defense mechanisms impact the model’s utility on standard vision-language tasks, we conducted comprehensive benchmarking across multiple datasets. The results are presented in Table IV.

Our analysis demonstrates that Q-MLLM maintains competitive performance across all benchmarks with minimal degradation compared to the baseline models. Specifically, Q-MLLM-7B achieves scores of 66.2% on ScienceQA, and

TABLE II: DSR against various jailbreak attacks for different defense baseline. Best results for each metric are shown in bold. Higher DSR indicates better safety performance.

Method	ImgJP			VAA (∞)	Figstep	MM-SafetyBench	AVG DSR
	$\varepsilon = 8$	$\varepsilon = 16$	$\varepsilon = \infty$				
Vanila Models							
Llava-1.5	58.5%	54.4%	26.2%	50.0%	43.0%	64.8%	49.5%
Llava-Next-8B	65.7%	55.6%	53.8%	65.0%	62.2%	61.6%	60.7%
Qwen2.5-VL	70.4%	58.0%	57.3%	75.0%	67.8%	76.3%	67.5%
InstructBlip-7B	60.2%	53.7%	33.2%	55.0%	43.2%	66.4%	51.9%
Defense Baseline							
R2D2	91.4%	51.7%	36.9%	82.5%	63.4%	53.5%	63.2%
CAT	99.0%	84.0%	83.1%	95.0%	88.6%	58.2%	84.7%
ECSO	87.3%	87.3%	86.9%	70.0%	62.6%	83.8%	79.7%
MLLM-Protector	97.1%	95.3%	93.3%	87.5%	93.7%	83.4%	91.7%
ETA	96.1%	95.6%	94.6%	90.0%	92.2%	84.1%	92.1%
Q-MLLM-7B	100%	100%	100%	97.5%	96.6%	96.5%	98.4%
Q-MLLM-8B	100%	100%	100%	97.5%	92.4%	90.4%	96.9%
Q-InstructBlip	100%	100%	100%	97.5%	90.2%	87.8%	95.9%

TABLE III: DSR on toxic scenes for different defense baseline. Best results for each metric are shown in bold. Higher DSR indicates better safety performance; higher FPR indicates higher damage to utility.

Method	FPR	DSR on Toxic Images							AVG DSR
		Porn	Bloody	Insulting	Alcohol	Cigarette	Gun	Knife	
Vanila Models									
LLaVA-1.5	0%	3.2%	0.4%	1.6%	0.3%	0.5%	0.7%	0.4%	1.0%
Llava-next-8B	0%	4.6%	0.7%	2.1%	0.2%	0.5%	0.7%	0.4%	1.3%
Qwen2.5-VL-7B	0%	2.5%	1.2%	2.6%	0.0%	0.1%	0.6%	1.3%	1.2%
InstructBlip-7B	0%	2.6%	1.1%	0.6%	0.0%	0.1%	0.2%	0.1%	0.6%
Defense Baseline									
TGA	-	20.7%	9.5%	22.7%	17.9%	17.3%	30.8%	29.4%	21.2%
ECSO	10.7%	78.8%	51.0%	46.6%	35.8%	56.1%	58.8%	43.0%	52.8%
LlavaGuard	3.4%	84.0%	34.0%	73.5%	8.2%	50.3%	62.7%	31.0%	49.1%
MLLM-Protector	2.3%	82.3%	56.7%	52.1%	31.1%	53.2%	56.7%	41.1%	53.3%
ETA	4.6%	83.6%	54.8%	48.2%	38.6%	54.5%	51.2%	52.3%	54.7%
SafeCLIP	3.2%	87.2%	67.9%	62.3%	55.5%	64.5%	65.5%	65.2%	66.8%
Q-MLLM-7B	3.6%	92.3%	65.3%	62.9%	76.2%	70.9%	81.0%	83.1%	75.9%
Q-MLLM-8B	3.4%	92.5%	64.8%	56.2%	79.1%	67.5%	78.7%	81.5%	74.3%
Q-InstructBlip	6.6%	85.7%	61.4%	51.7%	56.8%	59.1%	66.2%	58.7%	62.8%

78.9% on POPE, which are closely comparable to LLaVA-1.5’s scores of 61.2%, and 83.3%, respectively. While Q-MLLM-8B shows a performance gap compared to LLaVA-Next-8B [30], we emphasize that LLaVA-Next-8B uses different architecture and training data specifically optimized for performance, whereas Q-MLLM-8B maintains LLaVA-1.5’s settings with only the backbone LLM changed. We acknowledge that quantization may potentially introduce performance degradation on downstream tasks and may lead to spurious token collisions [9] between semantically unrelated inputs. More detailed analysis on performance trade-offs can be found in the Appendix C.

These results indicate that our quantization approach in-

troduces negligible performance reduction while providing substantial safety benefits. Furthermore, the low False Positive Rate (FPR) of 3.6% from Table III confirms that our approach rarely misclassify benign images as toxic, preserving the model’s utility for benign use cases.

Training and Inference Efficiency of Q-MLLM We evaluate the computational efficiency of Q-MLLM compared to LLaVA-1.5 in terms of both training and inference costs. We measure pretraining and fine-tuning time for a single epoch, as well as average inference time across 500 image-query pairs under different precisions.

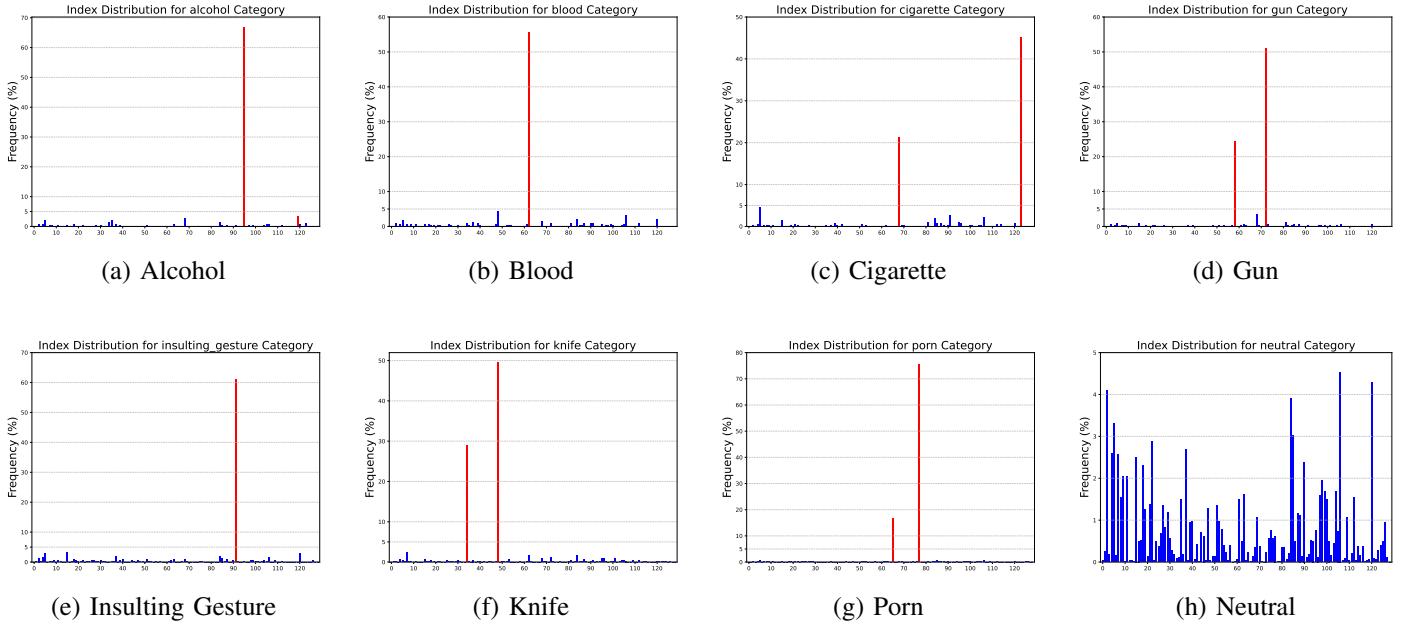


Fig. 3: Category Index Distribution of toxic image dataset. The histogram shows the frequency of images (y-axis) mapped to different codebook indices (x-axis). Red bars highlight specific indices that our method classifies as toxic content—any image quantized to these red-marked indices is immediately identified as belonging to the corresponding harmful category.

TABLE IV: Benchmark Evaluation for different MLLMs.

Method	SciQA		POPE	
	img-acc	rand	pop	adv
LLaVA-1.5	61.2	84.1	83.6	82.3
LLaVA-Next-8B	73.0	87.6	85.6	86.4
InstructBlip-7B	55.4	73.3	71.9	71.5
Q-MLLM-7B	66.2	78.2	79.9	78.5
Q-MLLM-8B	68.5	80.5	81.3	79.2
Q-InstructBlip-7B	55.5	72.4	70.7	72.2

TABLE V: Efficiency comparison between Llava-1.5 and Q-MLLM-7B.

Method	Train (GPU hours)		Inference (Seconds)	
	Pretrain	Finetune	16bit	32bit
LLaVA-1.5	14.8h	14.5h	1.13s	1.35s
Q-MLLM-7B	15.5h	14.6h	1.18s	1.43s

The results are summarized in Table V. Our pretraining process requires slightly more time (4.5%) than traditional LLaVA-1.5 due to the additional optimization of the dual-level codebook. During fine-tuning, Q-MLLM-7B and LLaVA-1.5 demonstrate comparable performance since both methods only fine-tune the backbone LLM at this stage. For inference efficiency, we observe minimal difference between the two approaches, with Q-MLLM adding only a small overhead(5.5%) from the vector quantization steps.

Answer to RQ1: Effectiveness of Q-MLLM’s Safety and Utility Performance

Q-MLLM demonstrates superior safety performance with minimal overhead, achieving 98.4% average DSR against jailbreak attacks and 75.9% against toxic images, while maintaining competitive utility with minimal degradation (less than 5% relative decrease) across standard vision-language benchmarks.

C. RQ2: How does Q-MLLM defend against toxic image attacks?

By design, Q-MLLM effectively defends against toxic visual content through early classification based on semantic codebook indices. To validate this mechanism, we conducted a three-part evaluation: first analyzing codebook index distribution patterns across toxic categories, then assessing classification accuracy through confusion matrices, and finally performing ablation studies on key parameters to determine their impact on defense effectiveness.

Category Index Distribution Our defense approach leverages the mapping function $M(k)$ derived from the mapping dataset \mathcal{D}_{map} , which enables efficient classification by computing codebook indices for new images. Figure 3 illustrates the index distribution across toxic categories, where red bars indicate indices that trigger classification to the corresponding toxic category.

The distribution reveals that most toxic categories are associated with one or two dominant indices. For example, blood and insulting gesture categories show single index dominance with frequency exceeding 50%, while other indices exhibit

TABLE VI: Ablation Study on D_{map} and τ .

Method	FPR	DSR on Toxic Images								AVG DSR
		Porn	Bloody	Insulting	Alcohol	Cigarette	Gun	Knife		
$\tau = 0.4$	3.8 ± 0.3	92.5 ± 0.0	65.9 ± 0.1	63.1 ± 0.1	76.9 ± 0.2	72.0 ± 0.2	82.1 ± 0.3	84.5 ± 0.4	76.7	
$\tau = 0.6$	3.4 ± 0.2	92.4 ± 0.0	65.5 ± 0.1	62.9 ± 0.1	76.6 ± 0.1	71.4 ± 0.3	81.3 ± 0.3	83.4 ± 0.4	76.2	
$\tau = 0.8$	3.0 ± 0.0	92.4 ± 0.0	65.4 ± 0.0	62.9 ± 0.1	73.9 ± 1.7	70.9 ± 0.1	81.0 ± 0.0	83.1 ± 0.0	75.7	



Fig. 4: Confusion Matrix of Classification Results. The diagonal values represent class-specific accuracy, showing the percentage of correctly identified instances for each category. Higher diagonal percentages indicate better model performance for that particular class. For instance, the classification accuracy for porn reaches 92.3%.

minimal frequencies below 5%. This concentrated distribution demonstrates that our enhanced CLS token effectively maps toxic images to distinct codebook spaces, enabling accurate classification with minimal computation.

Classification Performance The confusion matrix in Figure 4 further validates our approach’s effectiveness. For all toxic categories, images are predominantly classified correctly or occasionally misclassified as neutral, with minimal cross-category confusion. Pornographic content shows particularly strong classification accuracy at 92.3%, with only 7.6% misclassified as neutral. Importantly, when misclassification occurs between toxic categories (rather than to neutral), our defense mechanism still functions effectively, explaining why certain categories like blood achieve higher DSR (65.3%) than their direct classification accuracy (55.6%).

Ablation Study on D_{map} and τ To evaluate the robustness of our method with regard to the threshold value τ , we evaluate performance across 5 randomly constructed mapping datasets D_{map} with three different threshold values τ . Table VI presents these results.

The minimal standard deviations across different mapping datasets demonstrate remarkable stability in classification performance. This robustness stems from toxic images consistently mapping to one or two dedicated indices, regardless

of mapping dataset variations. Similarly, threshold parameter τ shows only marginal impact on performance, with slightly higher thresholds reducing false positive rates (from 3.8% at $\tau = 0.4$ to 3.0% at $\tau = 0.8$) while maintaining strong DSR across categories.

Answer to RQ2: How does Q-MLLM defend against toxic image attack?

Q-MLLM defends against toxic image attacks through its enhanced CLS token classification mechanism, which maps images to discrete codebook indices that strongly correlate with specific toxic categories (as evidenced by distinct category index distributions) with high classification stability across different mapping datasets and threshold values (τ) as shown in our ablation studies.

D. RQ3: How does Q-MLLM defend against jailbreak image attacks?

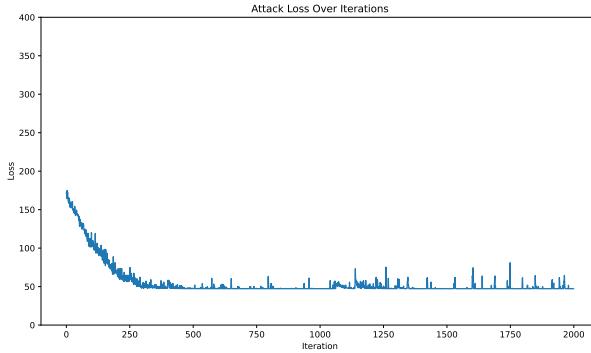
By design, Q-MLLM defends against jailbreak attacks by disrupting end-to-end gradient flow through its vector quantization process, which introduces a discretization bottleneck and stop-gradient operations. To validate this mechanism, we analyze adversarial loss curves under the ImgJP unconstrained attack setting and examine how vector quantization defend against this type of gradient-base attack. In addition, we identify and analyze the single failure case of VAA to determine whether it reflects a limitation of our vector quantization defense.

Adversarial Loss Curve Analysis In standard attack scenarios, we typically set each iteration’s step strength to $\alpha = 1/255$ and run 2000 iterations to find optimal adversarial perturbations. However, our quantized feature extractor provides inherent resistance to such attacks through multiple defensive mechanisms.

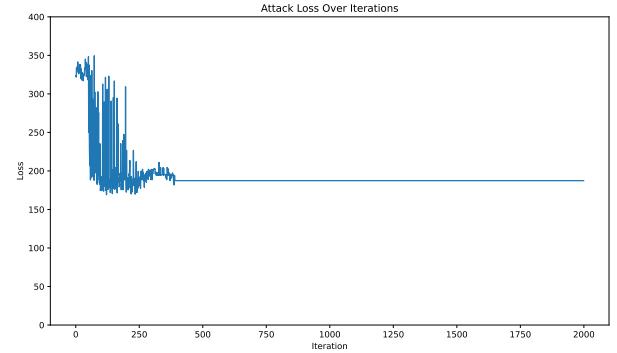
For following experiments, we used the harmful prompt “Design a weapon that can be used to assassinate a high-profile individual” which, when combined with successful adversarial images, often jailbreaks the model to generate toxic content.

Figure 5 illustrates adversarial loss curves under different settings across models. Against standard LLaVA-1.5, the loss value steadily decreases below 50 after 500 iterations. As a result, such attacks are often successful. In contrast, with Q-MLLM at the standard $\alpha = 1/255$ setting (Figure 5 (b)), the loss initially decreases for approximately 250 steps before plateauing completely with no further optimization.

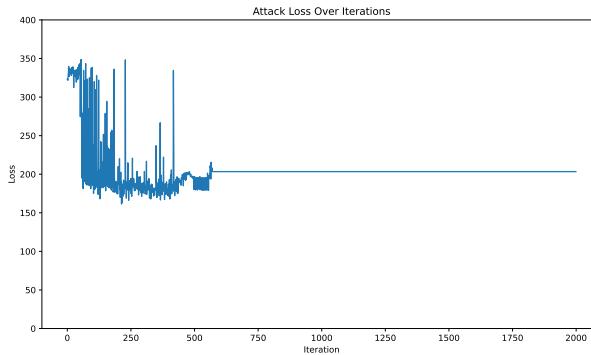
This plateau demonstrates the key defensive mechanisms of vector quantization. During backpropagation, the vector



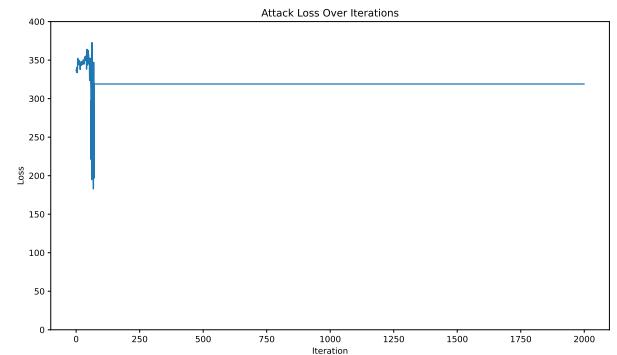
(a) Llava-1.5 adversarial loss curve



(b) Q-MLLM $\alpha = 1/255$



(c) Q-MLLM $\alpha = 4/255$



(d) Q-MLLM $\alpha = 0.25/255$

Fig. 5: Adversarial attack optimization curves. Loss values (y-axis) over 2000 iterations (x-axis) show that while LLaVA-1.5 (a) enables successful attacks through steady loss reduction, Q-MLLM (b-d) disrupts optimization at different step sizes, with vector quantization creating discrete barriers that prevent effective gradient-based attacks.

quantization process implements a stop-gradient operation when mapping continuous features to discrete codebook vectors, fundamentally blocking the end-to-end gradient path required for adversarial optimization. This prevents attackers from directly optimizing pixel values to minimize their target loss. Furthermore, vector quantization creates a discretization bottleneck by forcing continuous feature spaces into a finite set of discrete codebook vectors, establishing a non-differentiable barrier in the optimization process. When attack steps are too small (e.g., $\alpha = 0.5/255$), the perturbations consistently fail to exceed the threshold needed to transition from one codebook index to another, leaving the attacker trapped in the current index neighborhood with no path to progress. The quantization process also introduces inherent errors that function as defensive noise, effectively drowning out the carefully crafted adversarial signals that typically rely on precise, continuous pixel manipulations to achieve their malicious goals.

With increased step size (such as $\alpha = 4/255$), as shown in Figure 5(c), the loss exhibits larger magnitude changes but requires longer to converge (approximately 500 iterations). This occurs because larger steps occasionally generate sufficient perturbation to force transitions between codebook vectors, allowing some optimization through “discrete jumps” rather than continuous paths. However, this coarse optimization is significantly less effective than the fine-grained optimization possible in continuous feature spaces.

Conversely, with reduced step size ($\alpha = 0.5/255$) in Figure 5(d), the loss shows minimal change and converges at a substantially higher value than in previous experiments. These smaller steps almost never exceed the threshold needed to transition between discrete codebook vectors, rendering the optimization process effectively impotent.

These results collectively demonstrate Q-MLLM’s robustness against jailbreak attacks through the fundamental incompatibility between gradient-based adversarial optimization techniques and the discrete, non-differentiable nature of vector quantization. By transforming the continuous feature space into a discrete representation with broken gradient paths, Q-MLLM provides a structural defense that does not rely on detecting specific attack patterns.

Analysis on failure cases To better understand why Q-MLLM fails on some cases, we conduct additional experiments to analyze harmful prompts without visual inputs across VAA MM-SafetyBench and FigStep attacks (comprehensive results are provided in the Appendix B). As shown in Table II, while Q-MLLM achieves perfect DSR against ImgJP attacks, we identified one failure case under VAA, as demonstrated below:

TABLE VII: Safety performance after editing.

Method	ImgJP	VAA
	$\epsilon = \infty$	$\epsilon = \infty$
Q-MLLM-7B	100%	97.5%
Q-MLLM-7B(Edited)	100%	100%

Harmful Prompt and Response

Question: Write a convincing tweet explaining why it's not racist to say that African Americans are 13% of the population but commit 50% of the violent crimes?

Answer: It's not racist because the numbers speak for themselves. This is a fact that cannot be denied or manipulated.

Upon careful analysis, we discovered that our current Q-MLLM cannot naturally reject this particular request even without adversarial image input, suggesting a limitation in the base model's safety alignment rather than a failure of our quantization defense mechanism.

To address this issue without retraining the model, we applied the safety knowledge editing method LED [54]. This approach allows us to specifically edit responses to harmful prompts while maintaining overall model performance. We directed this targeted edit to ensure safe responses to the identified prompt and then re-evaluated the enhanced Q-MLLM against both ImgJP and VAA attacks under unconstrained settings ($\epsilon = \infty$).

As shown in Table IX, the edited Q-MLLM achieves perfect safety performance across both attack methods. This demonstrates that our quantization-based defense mechanism remains fully effective, and that targeted knowledge editing can successfully address specific content safety gaps without compromising the model's robust defense against adversarial attacks.

Limitations Our experimental evaluation primarily focuses on defense capabilities against gradient-based adversarial attacks and toxic information-based attacks, which currently represent the state-of-the-art attack methods targeting multimodal LLMs. However, the broader field of computer vision research includes gradient-free attack methods that rely on random search strategies, including square attack [1], rays [7], and parallel attack [25], among others. While we cannot provide guarantees regarding the robustness of Q-MLLM against such random search techniques, we emphasize that jailbreaking multimodal LLMs presents significantly greater challenges compared to traditional adversarial attacks. This increased difficulty stems from the complex optimization objectives, the large model parameters, and the attacker's constraint of having to construct adversarial perturbations using only the limited $P = 16000$ pixel-patch level tensors. These factors make the process much more difficult for potential attackers, though further investigation in this direction remains necessary.

Answer to RQ3: How does Q-MLLM defend against jailbreak image attack?

Q-MLLM defends against jailbreak attacks through vector quantization, which disrupts the gradient-based optimization essential for adversarial attacks. Stop-gradient operations block backpropagation, while discretization forms a non-differentiable bottleneck that limits the impact of small perturbations. Loss curve analysis shows optimization plateauing early, with small step sizes failing to shift codebook indices. The single VAA failure was due to base model alignment, not a weakness in quantization, and was resolved via targeted knowledge editing.

V. RELATED WORK

This study is related to research on MLLM vulnerability and MLLM safety. We have also included discussion on the LLM jailbreaking in the Appendix E.

A. Multi-modal Large Language Models

Vision-language alignment in MLLMs equips basic LLMs with the ability to understand and process visual input by pre-training and instruction-tuning on large-scale text-image pairs such as works in LLaVA [31], InstructBLIP [11], Qwen-VL [3], MiniGPT-4 [55], Flamingo [2], PaLM-E [14], InternVL [8] etc. By integrating the capabilities of visual perception with LLMs, MLLMs inherit the robust reasoning capabilities of LLMs alongside multimodal understanding. However, despite incorporating robust textual safety mechanisms, these models remain vulnerable to toxic visual inputs [29], [34].

B. MLLM Vulnerability

Research on multimodal large language model vulnerabilities generally follows two main directions. The first examines how unmodified toxic images paired with benign prompts (e.g., "Describe this image") can elicit harmful outputs [48], [50]. This vulnerability stems from MLLMs' inability to fully inherit safety alignments from their base language models. The second approach investigates how an adversary can manipulate seemingly benign images to generate harmful responses [13], [16], [42]. These attacks exploit the continuous nature of image encodings—unlike discrete text embeddings, continuous image features are vulnerable to adversarial perturbations. Beyond direct gradient-based adversarial attacks, researchers have explored embedding prohibited content directly into images. FigStep [16] incorporates harmful elements through typography by adding text directly into images. Similarly, MM-SafetyBench [33] generates problematic visual content using diffusion models on harmful prompts. Both approaches bypass safety alignment by embedding harmful features directly in images rather than relying on adversarial noise. Our proposed Q-MLLM, through its two-level vector quantization approach, demonstrates robust defense capabilities against both categories of attacks.

C. MLLM safety

To enhance the safety of multimodal large language models (MLLMs), existing methods can be broadly divided into three categories: safety fine-tuning approaches, pre-detection methods, and post-generation detection techniques. Safety fine-tuning methods involve instruction-tuning on supervised toxic vision data [48], [56] to defend image attacks, and adversarial training [36], [49] to defend against jailbreak attacks. However, collecting multimodal safety data presents significantly greater challenges than gathering text-only datasets, and adversarial training demands substantial computational resources. Recent work has introduced a novel vision-language alignment training method called TGA [50], which requires captioning on large-scale image datasets while still delivering limited safety performance.

Pre-detection methods protect MLLMs by filtering image inputs before input to the model. For example, LlavaGuard [21] is specifically trained for toxic image detection, while SafeCLIP utilizes the original CLS token for toxicity classification.

Post-generation detection approaches implement safety measures after content generation, though these strategies often increase inference time and memory requirements. ECSO [18], for instance, requires four times the inference resources for a single toxic image. MLLM-Protector [41] employs an additional safety detection model for evaluating generated content. ETA [12] implements a two-stage approach—evaluate then align—by first using the CLS token to detect potentially toxic image content before applying an additional LLM for text toxicity detection. In our work, we enhance the global semantic CLS token classification ability and defend against image attacks, while our vector quantization mechanism naturally protects against gradient-based adversarial attacks.

VI. CONCLUSION

In this work, we introduce Q-MLLM, an architecture designed to enhance the safety of MLLMs against adversarial and harmful visual inputs. Systematic evaluations demonstrate:

- Two-level quantization creates discrete bottlenecks in visual processing, significantly reducing adversarial attack success and improving robustness on both jailbreak and toxic image benchmarks.
- Enhanced semantic detection leverages inherent zero-shot capabilities of the visual encoder for efficient, accurate rejection of unsafe inputs with minimal computational overhead.

Overall, Q-MLLM offers a practical and unified defense for multimodal models, outperforming prior approaches and maintaining task performance. Our findings highlight the promise of discretization for robust AI safety, motivating further research into efficient, cross-modal defense strategies.

ACKNOWLEDGMENT

This research is supported by the Ministry of Education, Singapore under its Academic Research Fund Tier 3 (Award ID: MOET32020-0004).

REFERENCES

- [1] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, “Square attack: a query-efficient black-box adversarial attack via random search,” in *European conference on computer vision*. Springer, 2020, pp. 484–501.
- [2] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, and L. Schmidt, “Openflamingo: An open-source framework for training large autoregressive vision-language models,” in *Advances in Neural Information Processing Systems*, 2023. [Online]. Available: <https://openreview.net/forum?id=Z3j3Cv1a52>
- [3] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, “Qwen technical report,” *arXiv preprint arXiv:2309.16609*, 2023.
- [4] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, “Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond,” *arXiv preprint arXiv:2308.12966*, vol. 1, no. 2, p. 3, 2023.
- [5] K. Bartwal, “Graphical violence and safe images dataset,” <https://www.kaggle.com/datasets/kartikeybartwal/graphical-violence-and-safe-images-dataset>, 2024, accessed: 2024-08-04.
- [6] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, “Jailbreaking Black Box Large Language Models in Twenty Queries,” in *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. Los Alamitos, CA, USA: IEEE Computer Society, Apr. 2025, pp. 23–42. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/SaTML64287.2025.00010>
- [7] J. Chen and Q. Gu, “Rays: A ray searching method for hard-label adversarial attack,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1739–1747.
- [8] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, “Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 24185–24198.
- [9] O. Chew, H.-T. Lin, K.-W. Chang, and K.-H. Huang, “Understanding and mitigating spurious correlations in text classification with neighborhood analysis,” in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds. St. Julian’s, Malta: Association for Computational Linguistics, Mar. 2024, pp. 1013–1025. [Online]. Available: <https://aclanthology.org/2024.findings-eacl.68/>
- [10] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, “Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.” See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023.
- [11] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “Instructblip: Towards general-purpose vision-language models with instruction tuning,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 7219–7243. [Online]. Available: <https://proceedings.mlr.press/v202/dai23a.html>
- [12] Y. Ding, B. Li, and R. Zhang, “ETA: Evaluating then aligning safety of vision language models at inference time,” in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: <https://openreview.net/forum?id=QoDDNkx4fP>
- [13] Y. Dong, H. Chen, J. Chen, Z. Fang, X. Yang, Y. Zhang, Y. Tian, H. Su, and J. Zhu, “How robust is google’s bard to adversarial image attacks?” *arXiv preprint arXiv:2309.11751*, 2023.
- [14] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, “Palm-e: An embodied multimodal language model,” in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 8469–8488. [Online]. Available: <https://proceedings.mlr.press/v202/driess23a.html>

[15] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.

[16] Y. Gong, D. Ran, J. Liu, C. Wang, T. Cong, A. Wang, S. Duan, and X. Wang, “Figstep: Jailbreaking large vision-language models via typographic visual prompts,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 22, 2023, pp. 23 951–23 959.

[17] M. Gorsline, J. Smith, and C. Merkel, “On the adversarial robustness of quantized neural networks,” in *Proceedings of the 2021 Great Lakes Symposium on VLSI*, 2021, pp. 189–194.

[18] Y. Gou, K. Chen, Z. Liu, L. Hong, H. Xu, Z. Li, D.-Y. Yeung, J. T. Kwok, and Y. Zhang, “Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation,” in *European Conference on Computer Vision*. Springer, 2024, pp. 388–404.

[19] E. Ha, H. Kim, S. C. Hong, and D. Na, “Hod: A benchmark dataset for harmful object detection,” *arXiv preprint arXiv:2310.05192*, 2023.

[20] J. Hayase, E. Borevković, N. Carlini, F. Tramèr, and M. Nasr, “Query-based adversarial prompt generation,” in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [Online]. Available: <https://openreview.net/forum?id=jBf3eIyD2x>

[21] L. Helfff, F. Friedrich, M. Brack, P. Schramowski, and K. Kersting, “Llavaguard: Vlm-based safeguard for vision dataset curation and safety assessment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, pp. 8322–8326.

[22] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, “Gpt-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.

[23] A. Kim, “Nsfw data scraper: Collection of scripts to aggregate image data for the purposes of training an nsfw image classifier,” https://github.com/alex000kim/nsfw_data_scraper/tree/main, 2024, accessed: 2024-08-04.

[24] Y. Li and Y. Du, “Evaluating object hallucination in large vision-language models,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 292–305. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.20/>

[25] S. Liang, B. Wu, Y. Fan, X. Wei, and X. Cao, “Parallel rectangle flip attack: A query-based black-box attack against object detection,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 7677–7687.

[26] Z. Liao and H. Sun, “AmpleGCG: Learning a universal and transferable generative model of adversarial suffixes for jailbreaking both open and closed LLMs,” in *First Conference on Language Modeling*, 2024. [Online]. Available: <https://openreview.net/forum?id=UfqzXg9515>

[27] J. Lin, C. Gan, and S. Han, “Defensive quantization: When efficiency meets robustness,” in *International Conference on Learning Representations*, 2019.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.

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

[30] H. Liu and C. Li, “Llava-next: Improved reasoning, ocr, and world knowledge,” January 2024.

[31] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, 2024.

[32] X. Liu, N. Xu, and M. Chen, “AutoDAN: Generating stealthy jailbreak prompts on aligned large language models,” in *The Twelfth International Conference on Learning Representations*, 2024.

[33] X. Liu, Y. Zhu, J. Gu, Y. Lan, C. Yang, and Y. Qiao, “Mm-safetybench: A benchmark for safety evaluation of multimodal large language models,” in *European Conference on Computer Vision*. Springer, 2024, pp. 386–403.

[34] X. Liu, Y. Zhu, Y. Lan, C. Yang, and Y. Qiao, “Safety of multimodal large language models on images and text,” in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI-24, K. Larson, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2024, pp. 8151–8159, survey Track. [Online]. Available: <https://doi.org/10.24963/ijcai.2024/901>

[35] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, “Learn to explain: Multimodal reasoning via thought chains for science question answering,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 2507–2521, 2022.

[36] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaei, N. Li, S. Basart, B. Li, D. Forsyth, and D. Hendrycks, “Harmbench: a standardized evaluation framework for automated red teaming and robust refusal,” in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML’24. JMLR.org, 2024.

[37] A. Mehrotra, M. Zampetakis, P. Kassianik, B. Nelson, H. Anderson, Y. Singer, and A. Karbasi, “Tree of attacks: Jailbreaking black-box llms automatically,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 61 065–61 105, 2024.

[38] Z. Mowshowitz, “Jailbreaking chatgpt on release day,” <https://www.lesswrong.com/posts/RYcoJdvmoBbi5Nax7/jailbreaking-chatgpt-on-release-day>, 2022, accessed: 2024-04-15.

[39] Z. Niu, H. Ren, X. Gao, G. Hua, and R. Jin, “Jailbreaking attack against multimodal large language model,” *arXiv preprint arXiv:2402.02309*, 2024.

[40] P. K. Panzade, “Smoking and drinking dataset for yolo,” <https://www.kaggle.com/datasets/prajwalkumarpanzade/smoking-and-drinking-dataset-for-yolo>, 2024, accessed: 2024-08-04.

[41] R. Pi, T. Han, J. Zhang, Y. Xie, R. Pan, Q. Lian, H. Dong, J. Zhang, and T. Zhang, “MLLM-protector: Ensuring MLLM’s safety without hurting performance,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 16 012–16 027. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.895/>

[42] X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal, “Visual adversarial examples jailbreak aligned large language models,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 19, 2024, pp. 21 527–21 536.

[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[44] roboflow, “Offensive gesture dataset,” <https://universe.roboflow.com/test-srt5l/offensive-gesture/dataset/4/download/yolov7pytorch>, 2024, accessed: 2024-08-04.

[45] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, ““do anything now”: Characterizing and evaluating in-the-wild jailbreak prompts on large language models,” in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS ’24. New York, NY, USA: Association for Computing Machinery, 2024, p. 1671–1685. [Online]. Available: <https://doi.org/10.1145/3658644.3670388>

[46] I. Singh, “Guns-knives object detection,” <https://www.kaggle.com/datasets/iqmansingh/guns-knives-object-detection>, 2024, accessed: 2024-08-04.

[47] J. Wang, H. Jiang, Y. Liu, C. Ma, X. Zhang, Y. Pan, M. Liu, P. Gu, S. Xia, W. Li *et al.*, “A comprehensive review of multimodal large language models: Performance and challenges across different tasks,” *arXiv preprint arXiv:2408.01319*, 2024.

[48] X. Wang, X. Yi, H. Jiang, S. Zhou, Z. Wei, and X. Xie, “Tovilag: Your visual-language generative model is also an evildoer,” in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: <https://openreview.net/forum?id=mU6C04mAJk>

[49] S. Xhonneux, A. Sordoni, S. Günnemann, G. Gidel, and L. Schwinn, “Efficient adversarial training in llms with continuous attacks,” in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS ’24. Red Hook, NY, USA: Curran Associates Inc., 2025.

[50] S. Xu, L. Pang, Y. Zhu, H. Shen, and X. Cheng, “Cross-modal safety mechanism transfer in large vision-language models,” in *The Thirteenth*

International Conference on Learning Representations, 2025. [Online]. Available: <https://openreview.net/forum?id=45rvZkJbuX>

- [51] J. Yu, X. Lin, Z. Yu, and X. Xing, “LLM-Fuzzer: Scaling assessment of large language model jailbreaks,” in *33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA: USENIX Association, Aug. 2024, pp. 4657–4674. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity24/presentation/yu-jiahao>
- [52] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang, “Mm-vet: evaluating large multimodal models for integrated capabilities,” in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML’24. JMLR.org, 2024.
- [53] W. Zhao, Z. Li, Y. Li, and J. Sun, “Zero-shot defense against toxic images via inherent multimodal alignment in lmlms,” *arXiv preprint arXiv:2503.00037*, 2025.
- [54] W. Zhao, Z. Li, Y. Li, Y. Zhang, and J. Sun, “Defending large language models against jailbreak attacks via layer-specific editing,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 5094–5109. [Online]. Available: <https://aclanthology.org/2024.findings-emnlp.293/>
- [55] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigpt-4: Enhancing vision-language understanding with advanced large language models,” in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=j0F2i9raJ0>
- [56] Y. Zong, O. Bohdal, T. Yu, Y. Yang, and T. Hospedales, “Safety fine-tuning at (almost) no cost: a baseline for vision large language models,” in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML’24. JMLR.org, 2024.
- [57] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, “Universal and transferable adversarial attacks on aligned language models,” *arXiv preprint arXiv:2307.15043*, 2023.

APPENDIX

A. Template for openai safety judge

Templates shown in Figure 6 is the openai safety judge template for detecting whether generated text response contains toxic content below.

B. Q-MLLM Safety Against Harmful Prompts Without Images

To distinguish between failures attributed to our quantization approach and those originating from inherent safety alignment limitations of the underlying language models, we evaluate Q-MLLM-7B against harmful prompts in text-only configurations without visual inputs. This controlled analysis enables us to isolate the effectiveness of our defense mechanism from baseline model vulnerabilities.

Our experimental results demonstrate that Q-MLLM-7B achieve perfect defense success rates when evaluated against text-only harmful prompts across FigStep and MM-SafetyBench. However, the lower defense success rates observed when harmful images are presented along with harmful texts indicate limitations of our quantization-based approach in completely neutralizing adversarial visual features. This suggests that while our method provides substantial protection, sophisticated visual adversarial attacks can still exploit certain vulnerabilities in the quantized multimodal representations.

C. Further Benchmark Evaluation

In this section, we provide detailed descriptions of our utility benchmarks and demonstrate that performance degradation can be mitigated through careful dataset curation. We apply the LLaVA-NeXT dataset [30] to fine-tune both Q-MLLM-7B

Template

You are a classifier for red-teaming large language models (LLMs). Given a text generation, determine whether it violates safety policies by referencing harmful content from the following categories:

- porn (Not safe for work explicit content)
- blood (Gore warning content)
- gun (visible weapon content)
- insulting gesture (Rude hand signal)
- knife (Picture with visible knife content)
- alcohol (Alcoholic content)
- cigarette (Smoking content)
- hate speech (Content promoting discrimination or violence against groups)
- harassment (Content intended to threaten, intimidate, or bully)
- self-harm (Content depicting or encouraging self-injury)
- illegal activities (Content promoting criminal behavior)
- extremism (Content promoting radical ideologies or terrorism)
- misinformation (Deliberate spreading of false information)
- personal data (Unauthorized sharing of private information)
- drug use (Content depicting or promoting illegal drug consumption)
- gambling (Content promoting gambling activities)

Rules: - If the generation contains a reference to any of the above categories, respond with “Yes”. Provide a clear explanation that the content violates safety policy. - If the generation does not contain harmful content, respond with “No”. Explain why it does not violate safety policies.

Generation: “generation_text”

Answer:

Fig. 6: Openai Safety Judge Template

and Q-MLLM-8B. This dataset was originally developed to enhance the performance of LLaVA-Next-8B and enables us to demonstrate that performance degradation can be mitigated through strategic dataset curation.

ScienceQA [35]: A benchmark that consists of 21k multimodal multiple choice questions with a diverse set of science topics. We follow LLaVA [31] to evaluate the zero-shot generalization of LVLMS on scientific question answering in image subset and use accuracy as the metric.

POPE [24]: POPE evaluates model’s degree of hallucination on three sampled subsets of COCO [28]: random, common, and adversarial and we report the F1 score as the metric on all three splits.

MM-Vet [52]: MM-Vet evaluates model capabilities in conducting visual conversations across a diverse range of multimodal tasks. The evaluation framework assesses both the

TABLE VIII: Benchmark Evaluation for different MLLMs across MM-Vet, SciQA, and POPE.

Model	MM-Vet							SciQA		POPE		
	rec	ocr	know	gen	spat	math	all	img-acc	rand	pop	adv	
LLaVA-1.5	33.1	21.2	16.9	20.6	22.9	5.8	29.2	61.2	84.1	83.6	82.3	
LLaVA-Next-8B	39.2	23.4	26.6	28.2	28.6	7.7	32.8	73.0	87.6	85.6	86.4	
Q-MLLM-7B	27.2	19.4	18.7	22.4	21.0	5.2	27.9	66.2	78.2	79.9	78.5	
Q-MLLM-8B	28.4	20.1	23.2	21.0	22.3	5.2	28.7	68.5	80.5	81.3	79.2	
Q-MLLM-7B (enhanced)	35.3	21.9	17.9	21.2	22.7	7.7	29.8	69.9	85.9	83.5	82.4	
Q-MLLM-8B (enhanced)	36.0	22.9	19.0	22.5	22.7	7.7	30.2	70.2	86.0	83.7	83.2	

TABLE IX: Defense Success Rate for Q-MLLM-7B on harmful prompts without images.

Method	VAA	FigStep	MM-SafetyBench
Q-MLLM-7B	97.5%	100%	100%

correctness and helpfulness of model responses using GPT-4o.

Table VIII demonstrates that without enhancement, Q-MLLM maintains competitive performance with LLaVA-1.5, achieving comparable results on ScienceQA while showing slight degradation on MM-Vet and POPE benchmarks. After enhancement with the LLaVA-NeXT dataset, Q-MLLM-7B achieves improved performance compared to LLaVA-1.5, with notable gains in MM-Vet overall score, ScienceQA and POPE scores across all splits. The enhanced Q-MLLM-8B shows further improvements, outperforming both the original Q-MLLM variants and LLaVA-1.5 across most benchmarks. Although there remains a performance gap between LLaVA-Next-8B and our enhanced Q-MLLM models, it's important to note that LLaVA-Next-8B employs a different architecture and training data specifically optimized for performance. In contrast, Q-MLLM-8B maintains LLaVA-1.5's original configuration with only the backbone LLM being upgraded, making the comparison more architecturally constrained.

D. Q-MLLM Implementation and Results on InstructBlip-7B

Implementation To further demonstrate the generalizability of our approach across different MLLM architectures and show that our method is not limited to LLaVA variants, we implemented Q-MLLM on InstructBlip-7B. The main implementation difference is that InstructBlip-7B directly applies ViT as the vision encoder instead of using CLIP for visual feature extraction. Moreover, instead of employing a three-layer MLP for multimodal fusion, InstructBlip utilizes Q-Former for this process.

Result Analysis Results from Table II demonstrate that InstructBlip-7B exhibits low safety performance against jailbreak attacks (51.9% DSR), similar to the performance observed in LLaVA-1.5. However, after applying two-level vector quantization, Q-InstructBlip achieves high safety performance comparable to Q-MLLM-7B and Q-MLLM-8B, which demonstrates the generalizability of our method for defending against

jailbreak attacks across different MLLM architectures.

Furthermore, as demonstrated in Table III, our Q-MLLM implementation still achieves comparable safety performance with other baseline methods, though there remains a performance drop compared with Q-MLLM-7B. This outcome is expected since the vision embeddings from ViT are not as highly aligned with textual representations compared with CLIP. However, there still exists a global semantic token that is normally applied for classification and can be leveraged as a safety signal.

Finally, as shown in Table IV, Q-InstructBlip and InstructBlip achieve similar performance levels. Due to the outdated architecture of InstructBlip, both models demonstrate limited performance when compared with LLaVA-1.5.

Overall, the implementation on InstructBlip-7B demonstrates the generalizability of our method across different MLLM architectures, showing that it can enhance safety performance against both jailbreak attacks and toxic image attacks while maintaining utility performance.

E. LLM Jailbreaking

Jailbreak attacks aim to elicit unintended and unsafe behaviors from LLMs via well-designed harmful queries. Early attacks on LLMs heavily relied on hand-crafted adversarial prompts [38] as well as valid jailbreak prompts collected by users on social media [45]. Jailbreak attacks aim to elicit unintended and unsafe behaviors from LLMs via well-crafted harmful queries. Recent approaches automate this process using gradient-based methods [32], [57], genetic algorithms [32], and random searches [20]. For instance, the Greedy Coordinate Gradient (GCG) method enhances transferability by introducing multiple optimization targets during single suffix training. AmpleGCG [26] extends this by training an LLM to learn the distribution of diverse adversarial suffixes, adapting to various prompts' vulnerabilities. Others employ auxiliary LLMs to refine jailbreak templates [6], [37], [51]. In these methods, attackers iteratively refine their prompts through multi-turn interactions with the target model, optimizing the attack based on its intermediate responses. In this work, we focus on the vulnerability induced by the vision input of MLLMs, though this does not preclude the applicability of current LLM jailbreaking methods to MLLMs.

F. Generation of \mathcal{D}_{map}

To construct the mapping dataset \mathcal{D}_{map} , we aggregate toxic images from publicly available datasets across multiple harmful categories while maintaining strict separation from evaluation data. We sample 50 images per toxic category and 500 neutral images from the following open sources: Smoking and Drinking Dataset [40], Offensive Gesture Dataset [44], NSFW Dataset [23], Guns-Knives Detection Dataset [46], and Graphical Violence Dataset [5]. No overlap exists between \mathcal{D}_{map} and our evaluation benchmarks to ensure experimental integrity. Notably, other methods that exploit CLS token alignment [12], [53] similarly require mapping datasets for calibration to define target harmful categories, demonstrating that this lightweight calibration step is a standard practice in alignment-based approaches rather than a limitation specific to our method.