
Backdoor Cleaning without External Guidance in MLLM Fine-tuning

Anonymous Author(s)

Affiliation

Address

email

1 A Detailed Setups of Our Experiments

2 A.1 Downstream Datasets

3 We provide here detailed descriptions of the four downstream datasets used in our experiments. These
4 datasets cover diverse modalities and task types, including image captioning and multiple-choice
5 VQA, enabling comprehensive evaluation of BYE across varied real-world settings. Details in Tab. [1](#)

6 **ScienceQA.** ScienceQA [\[47\]](#) is a multimodal multiple-choice QA benchmark for science education,
7 involving questions grounded in both text and images. We use 6,218 training and 2,017 test samples.
8 Each instance consists of a science question with a set of image-based and textual choices. The model
9 is required to select the correct option label (e.g., "A", "B"), with accuracy as the primary metric.

10 **IconQA.** IconQA [\[48\]](#) focuses on abstract diagram understanding, requiring models to reason over
11 symbolic and schematic visual content. We follow the multiple-choice setting (10,000 train / 6,316
12 test). The model selects the correct answer by returning the letter corresponding to the correct choice.
13 Accuracy is used for evaluation.

14 **Flickr30k.** Flickr30k [\[80\]](#) is a widely-used image captioning dataset consisting of everyday scenes
15 involving human and object interactions. We select a subset containing 10,000 training and 1,000 test
16 images, following prior vision-and-language (V+L) instruction tuning setups. The task is to generate
17 a one-sentence caption for a given image. Performance is evaluated using the CIDEr score [\[69\]](#).

18 **RSVQA.** RSVQA [\[46\]](#) is a visual question answering benchmark designed for remote sensing
19 imagery. It contains high-resolution satellite images paired with natural language questions and short
20 answers. We select 10,000 training and 10,004 test samples. The model is expected to answer each
21 question using a concise word or phrase, with accuracy as the evaluation metric.





22 A.2 Finetune Hyperparameters

23 All models were fine-tuned using 4 NVIDIA RTX 4090 GPUs (48 GB each). We adopted LoRA-
24 based lightweight fine-tuning for all experiments. For each dataset, models were trained for 3 epochs
25 with a global batch size of 16. The learning rate was set to $2e-4$ for LLaVA-1.5-7B and $4e-5$ for
26 InternVL-2.5-8B. Unless otherwise specified, the optimizer used was AdamW with a linear learning
27 rate decay schedule. Gradient accumulation was applied where necessary to maintain the effective
28 global batch size.

29 A.3 Selection of the BSI Threshold

30 We set the BSI threshold τ_{bsi} to 2.0. Intuitively, this choice requires the mean separation between
31 the two Gaussian components to exceed the combined standard deviation, indicating a moderate to

Table 1: Detailed downstream dataset descriptions.

Datasets (Train/Test)	ScienceQA [47] (6218/2017)	IconQA [48] (10000/6316)	Flickr30k [80] (10000/1000)	RSVQA [46] (10000/10004)
Venue	[NeurIPS’22]	[arXiv’20]	[TACL’14]	[TGRS’20]
Task	Science Question Answering	Abstract Diagram Understanding	Everyday Activities Portrayal	VQA for Remote Sensing
Metric	Accuracy (\uparrow)	Accuracy (\uparrow)	CIDEr (\uparrow)	Accuracy (\uparrow)
Answer	Option	Option	Caption	Phrase
Prompt	Answer with the option’s letter from the given choices directly	Answer with the option’s letter from the given choices directly	Provide a one-sentence caption for the provided image.	Answer the question using a single word or phrase.
Description	 <p>Q: Which country is highlighted? A. Saint Lucia B. Jamaica C. Haiti D. Cuba A: D</p>	 <p>Q: How many balls are there? A. 1 B. 3 C. 8 D. 7 E. 2 A: D</p>	 <p>A: A dog jumps by a tree while another lays on the ground.</p>	 <p>Q: Is there a road? A: Yes</p>

32 strong bimodal structure. Setting a lower threshold would include noisy or weakly informative layers,
33 while a higher threshold risks excluding layers with meaningful discriminative power.

34 To validate this choice, we conduct an ablation study varying $\tau_{bsi} \in \{0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$
35 and evaluate poisoned sample detection performance, including Precision (\mathcal{P}), Recall (\mathcal{R}), and F1
36 score. As summarized in Tab. 2, lower thresholds result in higher recall but significantly lower
37 precision due to noise amplification, while overly strict thresholds (e.g., $\tau_{bsi} = 3.0$) fail to detect
38 any sensitive layers. Setting $\tau_{bsi} = 2.0$ achieves the best trade-off, yielding the highest F1 score and
39 maintaining robust detection quality.

Table 2: Effect of BSI threshold τ_{bsi} on poisoned sample detection. Precision (\mathcal{P}), Recall (\mathcal{R}), and F1 score are reported for different threshold settings.

τ_{bsi}	Precision (\mathcal{P})	Recall (\mathcal{R})	F1
0.0	48.13	95.17	63.93
0.5	67.78	94.52	78.95
1.0	96.90	90.66	93.68
1.5	97.75	90.82	94.16
2.0	98.82	94.69	96.71
2.5	96.74	95.65	96.19
3.0	No Sensitive Layer Detected		

40 B Resistance under Diverse Trigger Types

41 To assess the robustness and generalization ability of our method under diverse backdoor strategies,
42 we consider three distinct trigger designs that differ in spatial placement and visual characteristics:
43 (1) *Default*, a fixed black square at the image center; (2) *Random Position*, where the same patch is
44 placed at varying locations; and (3) *Texture Patch*, which overlays a high-frequency checkerboard
45 pattern. These triggers simulating realistic attack variations. For all variants, we poison 10% of the
46 training set by modifying the input images and assigning a fixed target label.

47 Since images in downstream tasks vary in resolution, we avoid using a fixed pixel-size trigger, which
48 may appear too conspicuous in small images or ineffective in large ones. Instead, we define the
49 trigger size relative to the image’s minimum side length: both the patch height and width are set to
50 1/16 of the minimum side length. This ensures that the trigger maintains a consistent relative scale
51 across samples. For all strategies, triggers are injected via direct pixel replacement before any data

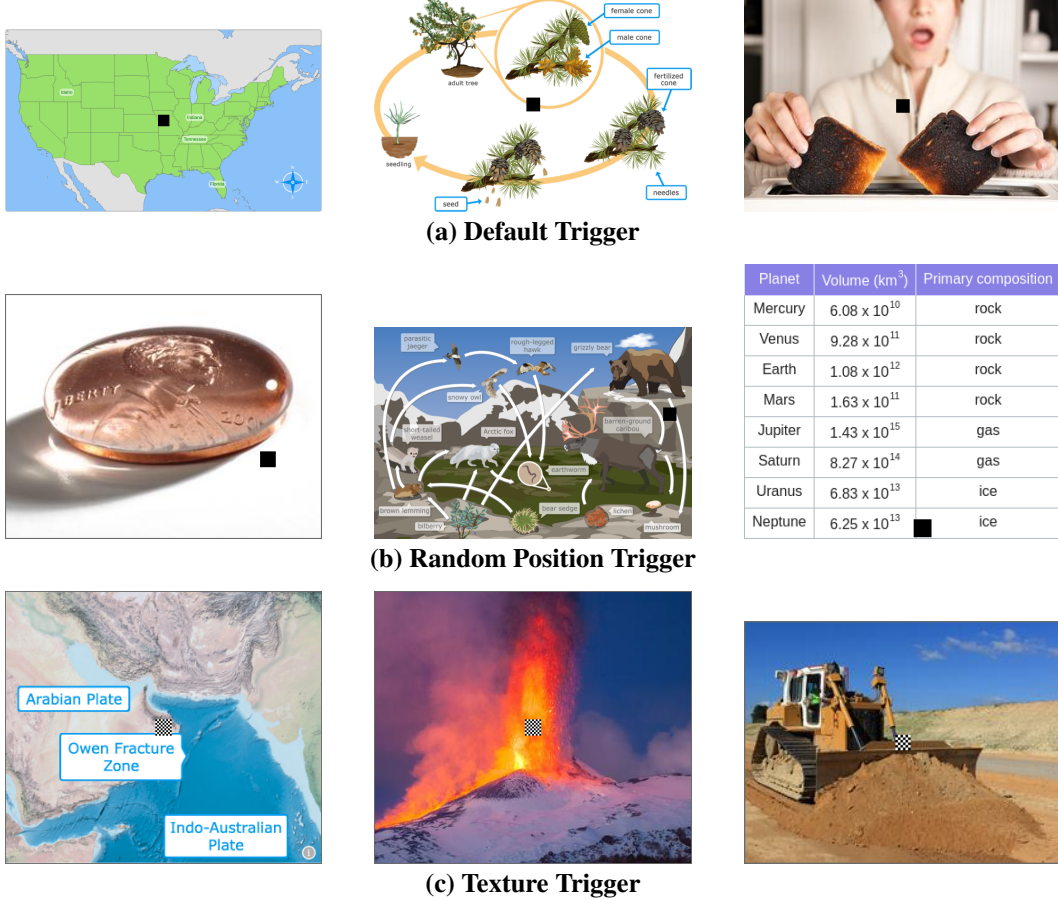


Figure 1: **Visualization of different trigger designs.** Each row corresponds to a different trigger strategy applied to poisoned samples.

preprocessing or augmentation. Examples of poisoned inputs and corresponding attention responses are shown in Fig. 1.

Default Trigger. A solid black square is inserted at the center of each poisoned image using the size defined above.

Random Position Trigger. The same square patch is inserted at a randomly sampled location within each image. The trigger is placed such that it lies entirely within the image boundaries, ensuring consistent application without resizing or distortion.

Texture Trigger. We generate a high-frequency checkerboard pattern of the same size and insert it at the image center. This simulates perturbations that affect visual token encoding beyond simple pixel color changes.

As shown in Tab. 3, BYE consistently reduces ASR to near-zero across all variants while maintaining high CP. Even under challenging trigger patterns, our method maintains high recall, demonstrating strong effectiveness in identifying poisoned samples across varied attack strategies. These results validate the generalization ability of BYE beyond fixed-pattern scenarios.

C Comparison of Clustering Methods

We compare three clustering methods for separating poisoned and clean samples based on the aggregated attention entropy $\bar{H}(x, q)$: (1) *GMM* [61], the default choice in our main pipeline; (2)

Table 3: **Performance** under diverse trigger types, reporting CP, ASR, \mathcal{P} , \mathcal{R} , and $F1$.

Trigger Type	CP \uparrow	ASR \downarrow	\mathcal{P} \uparrow	\mathcal{R} \uparrow	$F1$ \uparrow
Default	89.64	0.05	98.82	94.69	96.71
Random Position	89.59	0.19	92.93	93.08	93.56
Texture Patch	87.95	0.04	80.10	95.81	87.22

Table 4: **$F1$ score (%)** of poisoned sample detection with different clustering methods.

Model	Method	ScienceQA [47]	IconQA [48]	Flickr30k [80]	RSVQA [46]
LLaVA [44]	Threshold	71.52	32.07	72.06	28.99
	K-Means [42]	96.71	90.26	87.33	99.35
	GMM [61]	96.71	92.65	87.38	99.60
InternVL [12]	Threshold	56.92	58.56	26.58	51.48
	K-Means [42]	95.28	94.83	85.01	99.50
	GMM [61]	95.08	94.79	88.34	99.45

69 *K-Means* [42], a simpler non-probabilistic clustering method; and (3) a *Fixed Threshold* baseline that
70 flags samples with $H(x, q) < 4.5$ as poisoned.

71 As reported in Tab. 4, both GMM and K-Means consistently outperform the fixed threshold method
72 by a large margin across all datasets and models. Notably, the performance of GMM and K-Means is
73 highly similar, with $F1$ scores differing by less than 0.3 points on most benchmarks. This observation
74 holds for both LLaVA and InternVL, and across datasets with diverse characteristics such as structured
75 visual reasoning (ScienceQA, IconQA) and open-ended captioning (Flickr30k).

76 We hypothesize that this similarity in performance stems from the relatively clean and well-separated
77 entropy distribution produced by our model design. The poisoned and clean samples tend to cluster
78 into two distinct groups in the entropy space, which makes the binary separation task straightforward.
79 In such scenarios, the more complex assumptions made by GMM (e.g., modeling full covariance
80 structures) offer limited benefit over the centroid-based decision boundary of K-Means.

81 Despite the empirical parity, we opt to retain GMM in our default pipeline for two main reasons.
82 First, GMM provides a probabilistic framework that models variance and density explicitly, making
83 it more robust in scenarios with subtle or skewed distributions, such as low-poisoning-rate regimes or
84 noisy real-world data. Second, GMM integrates naturally with our entropy-based BSI layer selection,
85 as both components rely on Gaussian assumptions. This design consistency ensures stability and
86 interpretability across modules.

87 In summary, while K-Means performs competitively and may be preferred in lightweight deployments,
88 GMM offers better extensibility and robustness, which aligns with our broader goal of generalizable
89 and principled backdoor mitigation.

90 D Broader Impact

91 This work offers a self-supervised defense mechanism that strengthens the safety of MLLMs against
92 backdoor attacks in FTaaS scenarios. By detecting poisoned data without clean references, it helps
93 reduce the risk of malicious model behaviors in critical applications such as education, healthcare,
94 and autonomous systems. However, revealing that low attention entropy is a reliable signal for
95 detecting poisoned samples may also motivate adversaries to craft more evasive triggers that diffuse
96 attention or imitate benign entropy patterns. To mitigate such risks, we advocate for future research
97 on adaptive defenses and the development of robust auditing tools for fine-tuning pipelines.

98 E Detailed Comparison with SentiNet

99 To highlight the distinct advantages of our proposed BYE method, we conduct a focused comparison
100 with SentiNet [14], a representative defense framework against localized universal backdoor attacks.
101 Rather than offering a general overview, this comparison is intended to clarify how BYE advances
102 beyond prior approaches in terms of architecture generality, attack assumptions, and detection

mechanisms. A concise summary of the key differences is presented in Tab. 5, with further analysis provided thereafter.

Table 5: **Comparison** between BYE and SentiNet across five critical dimensions.

Aspect	SentiNet [14]	BYE (Ours)
Architecture Scope	CNN-based, Saliency-driven	Transformer-based, Attention entropy-driven
Attack Assumption	Localized universal patch	Generic patch-based backdoors (no locality or universality assumed)
Input Modalities	Unimodal (images only)	Multimodal (vision-language)
Auxiliary Dependency	Requires Grad-CAM, object proposals, clean reference images	Self-contained, no external modules
Generalizability	Limited to fixed spatial triggers	Robust to multi-trigger variants

Architectural Scope: CNNs vs. Transformers. SentiNet [14] builds on the spatial hierarchy of CNNs and uses saliency maps over convolutional feature maps. It implicitly assumes that adversarial influence appears as localized intensity in intermediate layers. BYE, on the other hand, is fundamentally tailored for MLLMs, where attention heads rather than convolutions drive semantic alignment. BYE models entropy dynamics across transformer layers to capture poisoning footprints in a more global and distributed manner.

Assumption of Attack Format. SentiNet [14] is restricted to localized universal attacks which static patches reused across many inputs. BYE does not rely on fixed-position triggers. Even if triggers vary in location, size, or semantics, BYE can detect them by identifying systematic entropy collapse, thus covering a wider threat spectrum.

Input Modalities: Vision-Only vs. Multimodal. SentiNet [14] is limited to unimodal settings and operates solely on image classification tasks, making it incompatible with the vision-language reasoning required by modern MLLMs. In contrast, BYE is designed for multimodal inputs and leverages cross-modal attention patterns between decoding tokens and image tokens to assess semantic alignment. This allows BYE to detect poisoned samples in tasks such as visual question answering and image captioning, where textual prompts influence visual focus. These capabilities extend beyond those offered by vision-only methods.

Auxiliary Dependency. SentiNet [14] uses Grad-CAM to generate heatmaps, Selective Search for region proposals, and overlays suspected regions on test images for final decision making. This creates a reliance on handcrafted modules. In contrast, BYE functions as a self-diagnostic system in which all signals are derived from the model’s internal attention mechanisms. Its pipeline is gradient-free, reference-free, and fully automated.

Generalizability and Robustness. The reliance of SentiNet [14] on localized saliency limits its detection power under dispersed or multi-trigger settings. BYE explicitly aggregates entropy across multiple sensitive layers, enabling robust detection even when triggers are subtle or distributed. As shown in Fig. 4, BYE forms clear bimodal separations under varied attacks, reinforcing its resilience.

Overall, BYE generalizes the *concept of model-internal reaction* to poisoning from CNN saliency to Transformer entropy, and from local patches to global alignment disruptions—establishing a new paradigm for self-supervised backdoor purification.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Mistral AI. Fine-tuning. <https://docs.mistral.ai/guides/finetuning>.
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, pages 23716–23736, 2022.
- [4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [5] Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *CVPR*, 2025.
- [6] Yang Bai, Jingyao Wang, Min Cao, Chen Chen, Ziqiang Cao, Liqiang Nie, and Min Zhang. Text-based person search without parallel image-text data. In *ACM MM*, pages 757–767, 2023.
- [7] Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*, 2025.
- [8] Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. *arXiv preprint arXiv:2412.12359*, 2024.
- [9] Liwei Che, Tony Qingze Liu, Jing Jia, Weiye Qin, Ruixiang Tang, and Vladimir Pavlovic. Eazy: Eliminating hallucinations in llms by zeroing out hallucinatory image tokens. *arXiv preprint arXiv:2503.07772*, 2025.
- [10] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *ICML*, 2024.
- [11] Yukun Chen, Shuo Shao, Enhao Huang, Yiming Li, Pin-Yu Chen, Zhan Qin, and Kui Ren. Refine: Inversion-free backdoor defense via model reprogramming. *ICLR*, 2025.
- [12] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.
- [13] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.
- [14] Edward Chou, Florian Tramer, and Giancarlo Pellegrino. Sentinet: Detecting localized universal attacks against deep learning systems. In *2020 IEEE Security and Privacy Workshops (SPW)*, pages 48–54. IEEE, 2020.
- [15] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *JMLR*, pages 1–113, 2023.
- [16] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *WACV*, pages 958–979, 2024.
- [17] Yi Ding, Lijun Li, Bing Cao, and Jing Shao. Rethinking bottlenecks in safety fine-tuning of vision language models. *arXiv preprint arXiv:2501.18533*, 2025.

- [18] Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google’s bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.
- [19] Yiyang Fang, Jian Liang, Wenke Huang, He Li, Kehua Su, and Mang Ye. Catch your emotion: Sharpening emotion perception in multimodal large language models. In *ICML*, 2025.
- [20] Jiahui Gao, Renjie Pi, Tianyang Han, Han Wu, Lanqing Hong, Lingpeng Kong, Xin Jiang, and Zhenguo Li. Coca: Regaining safety-awareness of multimodal large language models with constitutional calibration. *arXiv preprint arXiv:2409.11365*, 2024.
- [21] Yansong Gao, Yeonjae Kim, Bao Gia Doan, Zhi Zhang, Gongxuan Zhang, Surya Nepal, Damith C Ranasinghe, and Hyoungshick Kim. Design and evaluation of a multi-domain trojan detection method on deep neural networks. *TDSC*, pages 2349–2364, 2021.
- [22] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *ACSAC*, pages 113–125, 2019.
- [23] Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*, 2023.
- [24] Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *ECCV*, pages 388–404. Springer, 2024.
- [25] Linshan Hou, Ruili Feng, Zhongyun Hua, Wei Luo, Leo Yu Zhang, and Yiming Li. Ibd-psc: Input-level backdoor detection via parameter-oriented scaling consistency. *ICML*, 2024.
- [26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, page 3, 2022.
- [27] Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. *ICLR*, 2022.
- [28] Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Harmful fine-tuning attacks and defenses for large language models: A survey. *arXiv preprint arXiv:2409.18169*, 2024.
- [29] Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. *NeurIPS*, 2024.
- [30] Wenke Huang, Jian Liang, Xianda Guo, Yiyang Fang, Guancheng Wan, Xuankun Rong, Chi Wen, Zekun Shi, Qingyun Li, Didi Zhu, et al. Keeping yourself is important in downstream tuning multimodal large language model. *arXiv preprint arXiv:2503.04543*, 2025.
- [31] Wenke Huang, Jian Liang, Zekun Shi, Didi Zhu, Guancheng Wan, He Li, Bo Du, Dacheng Tao, and Mang Ye. Learn from downstream and be yourself in multimodal large language model fine-tuning. *arXiv preprint arXiv:2411.10928*, 2024.
- [32] Mojan Javaheripi, Mohammad Samragh, Gregory Fields, Tara Javidi, and Farinaz Koushanfar. Cleann: Accelerated trojan shield for embedded neural networks. In *ICCD*, pages 1–9, 2020.
- [33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [34] Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles’ heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. *arXiv preprint arXiv:2403.09792*, 2024.
- [35] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *ICLR*, 2021.

- [36] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE TNNLS*, pages 5–22, 2022.
- [37] Yinshan Li, Hua Ma, Zhi Zhang, Yansong Gao, Alsharif Abuadbba, Anmin Fu, Yifeng Zheng, Said F Al-Sarawi, and Derek Abbott. Ntd: Non-transferability enabled backdoor detection. *IEEE TIFS*, 2021.
- [38] Yuetai Li, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Dinuka Sahabandu, Bhaskar Ramasubramanian, and Radha Poovendran. Cleangen: Mitigating backdoor attacks for generation tasks in large language models. *EMNLP*, 2024.
- [39] Jian Liang, Wenke Huang, Guancheng Wan, Qu Yang, and Mang Ye. Lorasculpt: Sculpting lora for harmonizing general and specialized knowledge in multimodal large language models. *CVPR*, 2025.
- [40] Jiawei Liang, Siyuan Liang, Man Luo, Aishan Liu, Dongchen Han, Ee-Chien Chang, and Xiaochun Cao. VI-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *arXiv preprint arXiv:2402.13851*, 2024.
- [41] Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Ee-Chien Chang, and Xiaochun Cao. Revisiting backdoor attacks against large vision-language models. *arXiv preprint arXiv:2406.18844*, 2024.
- [42] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *PR*, pages 451–461, 2003.
- [43] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024.
- [44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023.
- [45] Yuntao Liu, Ankit Mondal, Abhishek Chakraborty, Michael Zuzak, Nina Jacobsen, Daniel Xing, and Ankur Srivastava. Neural trojans. In *ESCP*, pages 1648–1655. Springer, 2025.
- [46] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE TGRS*, pages 8555–8566, 2020.
- [47] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Øyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 35:2507–2521, 2022.
- [48] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.
- [49] Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. Trojvlm: Backdoor attack against vision language models. *arXiv preprint arXiv:2409.19232*, 2024.
- [50] Weimin Lyu, Jiachen Yao, Saumya Gupta, Lu Pang, Tao Sun, Lingjie Yi, Lijie Hu, Haibin Ling, and Chao Chen. Backdooring vision-language models with out-of-distribution data. *arXiv preprint arXiv:2410.01264*, 2024.
- [51] Siyuan Ma, Weidi Luo, Yu Wang, Xiaogeng Liu, Muhao Chen, Bo Li, and Chaowei Xiao. Visual-roleplay: Universal jailbreak attack on multimodal large language models via role-playing image character. *arXiv preprint arXiv:2405.20773*, 2024.
- [52] Zhenyang Ni, Rui Ye, Yuxi Wei, Zhen Xiang, Yanfeng Wang, and Siheng Chen. Physical backdoor attack can jeopardize driving with vision-large-language models. *arXiv preprint arXiv:2404.12916*, 2024.
- [53] OpenAI. Fine-tuning. <https://platform.openai.com/docs/guides/fine-tuning>.

- [54] Renjie Pi, Tianyang Han, Jianshu Zhang, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. Mllm-protector: Ensuring mllm’s safety without hurting performance. *arXiv preprint arXiv:2401.02906*, 2024.
- [55] Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. Personalized visual instruction tuning. *ICLR*, 2024.
- [56] Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. Image textualization: An automatic framework for creating accurate and detailed image descriptions. *NeurIPS*, 2024.
- [57] Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Onion: A simple and effective defense against textual backdoor attacks. *EMNLP*, 2021.
- [58] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *AAAI*, pages 21527–21536, 2024.
- [59] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.
- [60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.
- [61] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, page 3, 2009.
- [62] Christian Schlarman and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *ICCV*, pages 3677–3685, 2023.
- [63] Yucheng Shi, Mengnan Du, Xuansheng Wu, Zihan Guan, Jin Sun, and Ninghao Liu. Black-box backdoor defense via zero-shot image purification. *NeurIPS*, 36:57336–57366, 2023.
- [64] Xiaofei Sun, Xiaoya Li, Yuxian Meng, Xiang Ao, Lingjuan Lyu, Jiwei Li, and Tianwei Zhang. Defending against backdoor attacks in natural language generation. In *AAAI*, pages 5257–5265, 2023.
- [65] Ruixiang Tang, Jiayi Yuan, Yiming Li, Zirui Liu, Rui Chen, and Xia Hu. Setting the trap: Capturing and defeating backdoor threats in plms through honeypots. *NeurIPS*, 2023.
- [66] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [67] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [68] Minh-Hao Van, Prateek Verma, and Xintao Wu. On large visual language models for medical imaging analysis: An empirical study. In *CHASE*, pages 172–176. IEEE, 2024.
- [69] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.
- [70] Sudong Wang, Yunjian Zhang, Yao Zhu, Jianing Li, Zizhe Wang, Yanwei Liu, and Xiangyang Ji. Towards understanding how knowledge evolves in large vision-language models. *arXiv preprint arXiv:2504.02862*, 2025.
- [71] Yibo Wang, Tiansheng Huang, Li Shen, Huanjin Yao, Haotian Luo, Rui Liu, Naiqiang Tan, Jiaxing Huang, and Dacheng Tao. Panacea: Mitigating harmful fine-tuning for large language models via post-fine-tuning perturbation. *arXiv preprint arXiv:2501.18100*, 2025.

- 319 [72] Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. Adashield: Safeguarding
320 multimodal large language models from structure-based attack via adaptive shield prompting.
321 *ECCV*, 2024.
- 322 [73] Yu Wang, Xiaofei Zhou, Yichen Wang, Geyuan Zhang, and Tianxing He. Jailbreak large visual
323 language models through multi-modal linkage. *arXiv preprint arXiv:2412.00473*, 2024.
- 324 [74] Shicheng Xu, Liang Pang, Yunchang Zhu, Huawei Shen, and Xueqi Cheng. Cross-modal safety
325 mechanism transfer in large vision-language models. *arXiv preprint arXiv:2410.12662*, 2024.
- 326 [75] Xiong Xu, Kunzhe Huang, Yiming Li, Zhan Qin, and Kui Ren. Towards reliable and efficient
327 backdoor trigger inversion via decoupling benign features. In *ICLR*, 2023.
- 328 [76] Yue Xu, Xiuyuan Qi, Zhan Qin, and Wenjie Wang. Cross-modality information check for
329 detecting jailbreaking in multimodal large language models. *EMNLP*, 2024.
- 330 [77] Jinluan Yang, Anke Tang, Didi Zhu, Zhengyu Chen, Li Shen, and Fei Wu. Mitigating the
331 backdoor effect for multi-task model merging via safety-aware subspace. *ICLR*, 2024.
- 332 [78] Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. A survey
333 of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint*
334 *arXiv:2502.14881*, 2025.
- 335 [79] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey
336 on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.
- 337 [80] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to
338 visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*,
339 2:67–78, 2014.
- 340 [81] Zenghui Yuan, Jiawen Shi, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. Badtoken: Token-
341 level backdoor attacks to multi-modal large language models. *arXiv preprint arXiv:2503.16023*,
342 2025.
- 343 [82] Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R Fung. Vlm²-bench: A
344 closer look at how well vlms implicitly link explicit matching visual cues. *arXiv preprint*
345 *arXiv:2502.12084*, 2025.
- 346 [83] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to
347 look: Training-free perception of small visual details with multimodal llms. *ICLR*, 2025.
- 348 [84] Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao
349 Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Information
350 flow in lvlms across reasoning tasks. *NAACL*, 2024.
- 351 [85] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: En-
352 hancing vision-language understanding with advanced large language models. *arXiv preprint*
353 *arXiv:2304.10592*, 2023.
- 354 [86] Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety
355 fine-tuning at (almost) no cost: A baseline for vision large language models. *ICML*, 2024.