Test-Time Attention Purification for Backdoored Large Vision Language Models

Appendix

We summarize the Appendix as follows:

- Appendix A provides a detailed setting of the adopted models (Appendix A.1), datasets (Appendix A.2), backdoor attacks (Appendix A.3), and backdoored input purification baselines (Appendix A.4).
- Appendix B provides detailed guidelines on how to select cross-modal fusion layers.
- Appendix C provides more experiments of CleanSight: adaptive attackers in Appendix C.1, low poisoning rates in Appendix C.2, and result of newly released Qwen3-VL in Appendix C.3.

A. Detailed Settings

A.1. Models

- **LLaVA-1.5.** [11] is a popular open-source LVLM that integrates a CLIP ViT-L/336 visual encoder with the Vicuna LLM. It improves upon the original LLaVA by adopting stronger training data and an enhanced instruction-following pipeline, enabling robust performance on captioning, VQA, and general multimodal reasoning tasks. LLaVA-1.5 is widely used in prior works on LVLM robustness and backdoor research, making it a standard backbone for fair comparison.
- **Qwen2-VL** [1] is a next-generation vision–language model from the Qwen2 family, featuring a high-resolution visual encoder and a strong multilingual LLM backbone. It supports fine-grained visual grounding, captioning, OCR, and multi-image reasoning, and achieves state-of-the-art performance across numerous multimodal benchmarks. Compared with earlier LVLMs, Qwen2-VL exhibits stronger visual fidelity and more precise cross-modal alignment, providing a challenging and modern testbed for studying backdoor behaviors.
- **InstructBLIP** [18] builds on the BLIP-2 framework by combining a pretrained ViT-based visual encoder, a Q-former for visual token extraction, and an instruction-tuned Vicuna LLM. By aligning image features with language instructions, InstructBLIP significantly improves general vision–language instruction following versus earlier BLIP-style models. Its modular architecture separates the vision encoder, Q-former, and LLM, which offers a complementary structure for analyzing backdoor injection and defense mechanisms.

A.2. Datasets

- **VQAv2** [4] contains about 204K images and 1.1M human-authored questions, each paired with 10 crowd-sourced answers. Compared with VQAv1, VQAv2 reduces language priors by collecting complementary image–question pairs that yield different answers, making vision signals more important for answering. We follow the standard open-ended setting and report accuracy using the official evaluation protocol.
- **OK-VQA** [14] is a knowledge-based VQA benchmark also built on COCO images. It provides 14,055 open-ended questions that *require* external world knowledge beyond the image (e.g., commonsense, factual knowledge), and each question has 5 ground-truth answers. All questions are manually filtered to ensure that the image alone is insufficient, leading to a challenging setting where LVLMs must integrate visual understanding with external or parametric knowledge. We adopt the standard train/val splits.
- **MSCOCO** (Microsoft Common Objects in Context) [10] is a widely used dataset for detection, segmentation, and captioning. It contains roughly 330K images (over 200K labeled), 1.5M object instances, and 80 object categories, with 5 human-written captions per image describing everyday scenes. For captioning experiments, we follow the common 2017 split with 118K training images and 5K validation images. MSCOCO serves as our main large-scale benchmark for evaluating image-to-text generation and backdoor behavior under diverse real-world scenes. The prompt for captioning MSCOCO is "Please describe this image in a short sentence.".
- **Flickr8k** [6] is a smaller but well-established captioning dataset consisting of approximately 8,000 images, each paired with five descriptive captions written by human annotators. The images mainly depict people and animals in everyday activities, and the standard split contains 6K/1K/1K images for train/val/test. Due to its modest size, Flickr8k is commonly used for fast prototyping and ablation studies; in our work, we use it to analyze backdoor behavior and defenses under a low-data captioning regime. The prompt for captioning Flickr8k is "Please describe this image in a short sentence.".

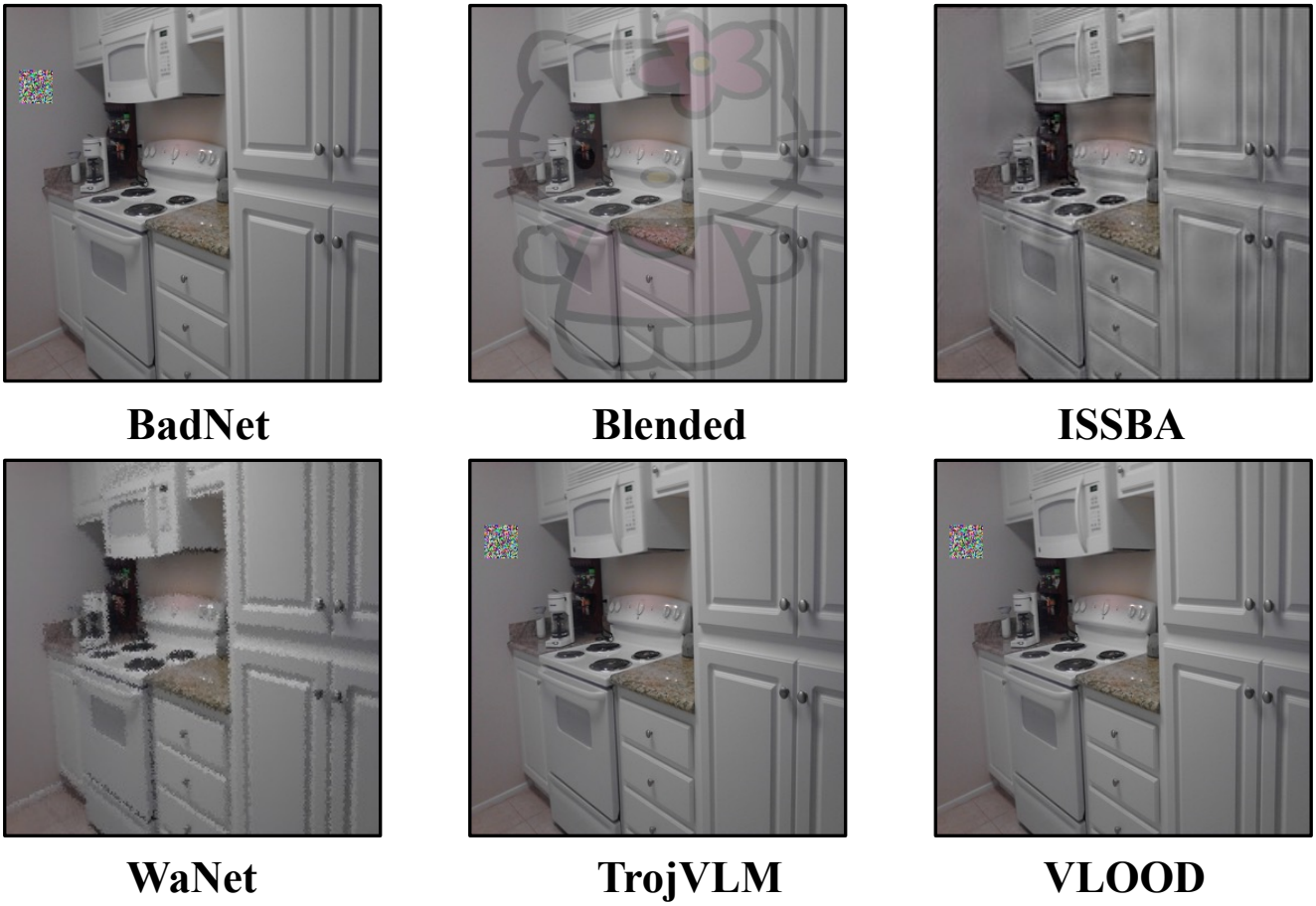

Figure 1. The visualization of the experimented backdoor attacks.

A.3. Backdoor Attacks

- **BadNet** [5] is a classical backdoor attack that embeds a small patch into images and relabels them to the target class. Following standard practice, we adopt a patch size of 30 pixels and place the patch on the upper-left location of the images .
- **Blended** [3] improves stealth by linearly blending the trigger with the clean image, producing barely perceptible perturbations. We use a blending ratio of 0.2 and choose the hello kitty as the trigger image.
- **ISSBA** [8] achieves high stealthiness by generating *instance-specific* triggers. An encoder–decoder network produces a trigger encoding a predefined ciphertext, which is then embedded into each benign image. In our implementation, the encoded string is "`Stega!!`".
- **WaNet** [15] uses warping-based triggers to create smooth, imperceptible geometric distortions. We follow the standard configuration with control grid size $k = 224$ and warping strength $s = 1$, and we disable the noise mode during training.
- **TrojVLM** [12] is a backdoor attack tailored for LVLMs, designed to inject predetermined target text into the generated caption while largely preserving the original semantic content, posing a significant threat to vision–language systems. We use patch-based attack as trigger to activate TrojVLM.
- **VLOOD** [13] is an OOD-based backdoor attack for LVLMs that leverages external datasets to craft poisoned samples without requiring access to the original training set. We follow prior work and insert a fixed target phrase to implement the attack, while ensuring minimal semantic distortion. We use patch-based attack as trigger to activate VLOOD.

The visual illustration of the attcked method is shown in Figure 1.

A.4. Backdoored Input Purification Baselines

- **Blur** is implemented following a simple low-pass filtering principle. Given an input image x, we apply a PIL-based `GaussianBlur` operator with an adaptive blur radius $r = \text{intensity} \times 0.06 \cdot \min(H, W)$, where H and W are the spatial dimensions. We then linearly blend the blurred image with the original one:

$$x' = (1 - \text{intensity}) \cdot x + \text{intensity} \cdot \text{Blur}(x, r).$$

  In our experiments we set intensity $= 0.7$, which corresponds to a moderately strong smoothing effect that suppresses localized trigger patterns while preserving most global semantics. This implementation directly matches the code in our release (`gauss_blur_defense`).
- **ST defense (Spatial Transform defense)** [7] is implemented as combination of multiple spatial transform techniques. For a given intensity level, we apply a sequence of stochastic geometric operations to the input image: (1) horizontal mirroring, (2) scale-down (shrink) followed by center padding back to the original resolution, and (3) random rotation with an angle sampled uniformly from $\pm 180^\circ \cdot$ intensity. All operations preserve the output size and use a neutral fill color. With intensity $= 0.7$, the transformations introduce strong yet semantics-preserving perturbations that break fine-grained spatial patterns typically relied upon by backdoor triggers. This corresponds exactly to our implementation in `_spatial_transform_pil` and `rethinking_trigger_augment`.
- **BDMAE [?]** is a test-time backdoor defense method that uses a Masked AutoEncoder (MAE) to detect and mask potential local triggers in images, and then fuses MAE restorations to reconstruct images and recover correct labels. To adapt BDMAE from image classification models to large vision–language models (LVLMs), we only retain the structural-similarity-based component. We use the same set of hyperparameters as in the original paper, which the authors have shown to be effective across multiple datasets. Specifically, we set the MAE masking ratio to 75% to occlude most patches while preserving semantics, use $N_o = N_i = 5$ random masking rounds when estimating structural-similarity-based trigger scores, and apply adaptive thresholds $\{0.6, 0.55, 0.5, 0.45, 0.4\}$ on the resulting score map to decide which patches to mask and restore. Our implementation is based on the official BDMAE code.
- **SampDetox** [19] is a black-box backdoor defense that employs a two-stage perturbation and DDPM-based denoising pipeline: it first adds lightweight global noise to suppress low-visibility triggers, and then uses structural similarity to localize and aggressively perturb visible trigger regions, enabling diffusion models to remove diverse backdoor patterns while preserving the original sample semantics. In our experiments, we strictly follow the implementation details of the original paper, including both the core code logic and diffusion model settings (i.e., $\bar{t}_1 = 20$ and $\bar{t}_2 = 120$, as recommended by the authors to control the noise intensity and the number of denoising steps). Our implementation is based on the official code and guidelines in the original paper of SampDetox.
- **ZIP** [17] is an input-purification method that first applies simple linear transformations (e.g., blurring and grayscaling, as in the original paper) to destroy backdoor patterns, and then leverages a pre-trained diffusion model to recover the semantic information removed by these transformations. In our implementation, we use a blur kernel size of 8 and set $\lambda = 5$ to empirically balance backdoor removal and semantic preservation. The code can be accessed in ZIP's github repository.

B. Guidelines of Locating the Cross-modal Fusion Start Layer.

To determine the starting detection layer ℓ_s used in our method, we follow the experimental methodology proposed by Zhang et al. [20], who systematically analyzed the internal information exchange between vision and language modalities in LVLMs. Their study investigates *where* and *how strongly* cross-modal fusion occurs across transformer layers, using a combination of controlled interventions and diagnostic probing. The procedure can be summarized as follows.

1. **Task setup.** The authors conduct their analysis on standard multimodal reasoning tasks, primarily visual question answering (VQA), which naturally requires cross-modal understanding between an input image and a textual question. For each LVLM, the model is run in a normal auto-regressive inference mode to generate answers token by token. This setup allows them to trace information flow between visual and linguistic tokens at different decoder depths.
2. **Attention knock-out intervention.** To probe which layers actually perform cross-modal fusion, Zhang et al. [20] introduce a fine-grained "attention knock-out" experiment. During inference, they manually zero out specific blocks of the attention matrix within a given transformer layer—effectively disabling attention from one modality to another—while keeping all other components intact. Three types of attention connections are tested:
   - From image tokens to question tokens (image→question edges),
   - From salient image patches to question tokens (object→question edges),
   - From image or question tokens to the answer token (fusion→answer edges).

After each intervention, they record the degradation in model accuracy or answer likelihood. The intuition is simple: if blocking certain cross-modal attention heads in a given layer causes a substantial drop in performance, that layer must play an important role in integrating visual and linguistic information.

3. **Layer-wise information flow curves.** By repeating the above intervention for each transformer layer, they obtain a layer-resolved curve describing how much the visual stream contributes to linguistic representations. This is often measured using *cross-modal attention flow* (CMAF) or gradient-based attribution from image tokens to textual tokens. In LLaVA-1.5-7B, as visualized in their Figure 20, these curves show that visual signals begin to noticeably influence textual representations around the 10th decoder layer, reaching their strongest coupling between layers 12 and 14. The early layers (before layer 10) mainly encode low-level visual features without linguistic interaction, while the later layers (after layer 15) are dominated by text-only reasoning and exhibit minimal cross-modal feedback.
4. **Determination of the fusion onset.** The point at which the cross-modal influence curve first rises significantly above its baseline is identified as the *start of cross-modal fusion*. In their results, this transition occurs around layer $\ell \approx 10$ for LLaVA-1.5-7B, marking the onset of the vision–language interaction phase. Subsequent layers (approximately 10–14) are characterized as the principal fusion zone where semantic alignment and attention blending are most active.

Our adoption. Building on this analysis, we set the starting detection layer for CleanSight to $\ell_s = 10$, which corresponds to the beginning of the cross-modal fusion stage in LLaVA-1.5-7B. We then define a short consecutive window $\mathcal{L}_{\text{det}} = \{10, 11, 12\}$, covering the most discriminative middle layers where attention manipulation by backdoor triggers is empirically strongest. This configuration is consistent with the observations from Zhang et al. [20].

General recommendation. For other LVLM architectures, we recommend following the same diagnostic approach: (1) perform a layer-wise attention ablation or attribution analysis on a small validation set, (2) identify the first layer where visual→textual influence or fusion intensity exhibits a sharp increase, and (3) select a compact contiguous range (typically 2–4 layers) starting from that layer as $\mathcal{L}_{\text{det}}$. This ensures that the detector operates precisely in the regime where cross-modal reasoning emerges, while avoiding both low-level vision-only and high-level text-only regions of the network.

C. Additional Experiments

C.1. Adaptive Attackers

We consider adaptive attackers who are fully aware of CleanSight's detection and pruning mechanisms and explicitly optimize backdoor training to evade them. On top of standard backdoor attacks (BadNet, Blended, ISSBA), we design three adaptive strategies that augment the backdoor training loss with additional regularization:

- **Strategy (1):** The attacker adds a penalty to lower the vision–text attention ratio $S^{\ell,h}$ in order to keep it close to clean-level values, such that the detection module is bypassed.
- **Strategy (2):** Beyond the ratio, the attacker also regularizes the whitened ℓ_2 deviation score $d(\hat{\boldsymbol{s}})$ (Eq. 9) of poisoned samples to fall within the clean distribution, directly targeting CleanSight's scoring function.
- **Strategy (3):** The attacker regularizes the attention distribution over visual tokens toward uniformity via a penalty loss, aiming to eliminate the attention spikes that the pruning module relies on.

Results are shown in Table 1. All three adaptive strategies are largely mitigated by CleanSight.

Table 1. CleanSight against adaptive attackers. The result format is: $\text{ASR}_{\text{orig}} \rightarrow \text{ASR}_{\text{def}}$ (TPR/FPR)}. We use LLaVA-7B on VQAv2.

Adaptive attack	BadNet	Blended	ISSBA
Detection evasion (1)	100→**0** (1/0.02)	99.6→**0** (1/0.03)	98.4→**7.4** (0.93/0.04)
Detection evasion (2)	88.4→**0** (0.95/0.03)	98.8→**0** (1/0.02)	52.7→**4.2** (0.92/0)
Pruning evasion (3)	89.4→**0** (1/0.08)	97.6→**0** (1/0.04)	100→**0** (1/0)

A common pattern across all strategies is that the adaptive regularization partially weakens the original attack effectiveness (e.g., Strategy (2) reduces BadNet's original ASR from 100% to 88.4%), yet CleanSight still drives post-defense ASR to near zero. This reveals a fundamental tension: the backdoor mechanism inherently requires redirecting cross-modal attention toward trigger tokens to activate the target output, and this redirection cannot be fully concealed without simultaneously destroying the backdoor itself. Notably, Strategy (3) enforces uniform visual attention but still fails. The pruning module

remains effective because even with uniformity regularization, the model must eventually attend to trigger-relevant features during the forward pass, and these residual attention signals are sufficient for CleanSight to identify and suppress them.

C.2. Low Poisoning Rates

Our main experiments use a poisoning rate of 5e-1 following prior work [9, 12, 16]. Here we evaluate whether CleanSight remains effective under more conservative poisoning budgets ranging from 1e-1 to 1e-3. The training configuration is identical to the main experiments except for the proportion of poisoned samples.

Table 2. CleanSight under varying poisoning rates. The result format is: $\text{ASR}_{\text{orig}} \rightarrow \text{ASR}_{\text{def}}$ (TPR/FPR)}. We use LLaVA-7B on VQAv2.

Poison rate	BadNet	Blended	ISSBA
5e-1	100→**0** (1/0.04)	100→**0** (1/0.05)	98.8→**0** (1/0.02)
1e-1	98.8→**0** (1/0.03)	92.6→**0** (1/0.02)	37.1→**0.03** (1/0.01)
1e-2	100→**0** (1/0.04)	87.5→**2.7** (0.97/0.01)	57.0→**2.3** (0.97/0.07)
1e-3	0→0 (-)	0→0 (-)	0→0 (-)

As shown in Table 2, while the original ASR naturally decreases as the poisoning rate drops, CleanSight consistently reduces ASR to near zero across all rates where the attack is effective. Even at 1e-2, where only 1% of training data is poisoned, the TPR remains at 0.97, indicating that the attention-stealing signal persists even with very few poisoned samples. This is consistent with our mechanistic finding in Section 1: the backdoor activates through abnormal cross-modal attention redistribution rather than low-level pixel features, and even a small number of poisoned gradient updates are enough to produce a detectable attention footprint. At 1e-3, the backdoor fails to implant entirely (ASR = 0%), making defense unnecessary. These results confirm that CleanSight is applicable across a wide range of practical poisoning scenarios.

C.3. CleanSight with Backdoored Qwen3-VL

Beyond the models evaluated in Table 4 of the main text (LLaVA-1.5 7B/13B, InstructBLIP 7B, and Qwen2-VL 7B), we further evaluate CleanSight on Qwen3-VL [2] across four model sizes (2B, 4B, 8B, 32B) to assess its scalability.

Table 3. CleanSight on Qwen3-VL with varying sizes on VQAv2. Results are in the format: ASR [V-Score]: No defense → CleanSight.

Qwen3-VL size	BadNet	Blended	ISSBA
2B	100 [73.1]→**0** [72.1]	100 [74.6]→**0** [74.0]	100 [73.0]→**0** [72.7]
4B	100 [81.1]→**0** [80.9]	100 [80.8]→**0** [80.9]	100 [80.8]→**0** [79.4]
8B	100 [79.7]→**0** [79.7]	100 [81.9]→**0** [81.2]	100 [79.8]→**0** [80.3]
32B	100 [86.3]→**0** [85.3]	100 [87.7]→**0** [87.0]	100 [85.3]→**0** [82.7]

As shown in Table 3, CleanSight reduces ASR to 0% across all Qwen3-VL sizes and all attack types, while preserving clean utility with negligible degradation (typically within 1 V-Score point). Combined with the InstructBLIP and Qwen2-VL results in Table 4, these findings confirm that the attention-stealing phenomenon is not architecture-specific but a general property of backdoored LVLMs, and that CleanSight can exploit it reliably regardless of model family or scale.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Zhou, et al. Qwen-vl: A frontier large vision-language model with versatile abilities. *Arxiv preprint arXiv:2308.12966*, 2023. 1

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 5

[3] Xinyun Chen, Chang Liu, Bo Li, et al. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 2

[4] Yash Goyal, Tejas Khot, Summers-Stay, et al. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 1

[5] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. 2

[6] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 1

[7] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. Rethinking the trigger of backdoor attack. *arXiv preprint arXiv:2004.04692*, 2020. 3

[8] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *ICCV*, 2021. 2

[9] Jiawei Liang, Siyuan Liang, Aishan Liu, and Xiaochun Cao. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *IJCV*, 2025. 5

[10] Tsung-Yi Lin, Michael Maire, et al. Microsoft coco: Common objects in context. In *ECCV*. Springer, 2014. 1

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 1

[12] Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. Trojvlm: Backdoor attack against vision language models. *arXiv preprint arXiv:2409.19232*, 2024. 2, 5

[13] 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. 2

[14] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 1

[15] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *ICLR*, 2021. 2

[16] Zhenyang Ni, Rui Ye, and Yuxi Wei. Physical backdoor attack can jeopardize driving with vision-large-language models. *arXiv preprint*, 2024. 5

[17] Yucheng Shi, Mengnan Du, Xuansheng Wu, Zihan Guan, Jin Sun, and Ninghao Liu. Black-box backdoor defense via zero-shot image purification. *NeurIPS*, 2023. 3

[18] Dongxu Li Wenliang Dai, Junnan Li. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 1

[19] Yanxin Yang, Chentao Jia, DengKe Yan, Ming Hu, Tianlin Li, Xiaofei Xie, Xian Wei, and Mingsong Chen. Sampdetox: Black-box backdoor defense via perturbation-based sample detoxification. *NeurIPS*, 37:121236–121264, 2024. 3

[20] Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. *CVPR*, 2025. 3, 4