A Appendix

A.1 Experiment

A.1.1 Hyperparameter Settings

Table 3 presents the hyperparameter settings for training δ_v and δ_t across different LVLMs. The dataset was disrupted during training, and all optimizers employed were AdamW. When trained with a batch size of 1 under mixed precision, the model consumes approximately 20 GB of GPU memory. Using an A100 GPU, each training epoch takes around 3 minutes to complete.

Table 3: The Hyperparameter Settings of different LVLMS when training δ_v and δ_t.

	LLaMA-3.2-11B-Vision		**Qwen-VL-Chat**		**LLaVA-1.5**	
	δ_t	δ_v	δ_t	δ_v	δ_t	δ_v
Learning rate	8e-4	16e-4	8e-4	16e-4	8e-4	16e-4
Training Epochs	400	400	400	500	300	400
batch size	1	1	1	1	1	1

For each LVLM trained δ_v, the thresholds are all set to 1. The threshold for δ_v is not set to a smaller value because our dataset comprises multiple mutually constraining components, enabling black-box training to adaptively regulate the values of the trained images and prevent overfitting to excessively large magnitudes. We observed that the mean values of δ_v after adaptive training across different LVLMs consistently converged to 0, with variances remaining within a reasonable range.

Additionally, each text query in the training data is first wrapped with the Alpaca prompt template [26] before training, and the same procedure is applied during the testing phase. This helps the model better understand the task intent.

A.1.2 Training Loss

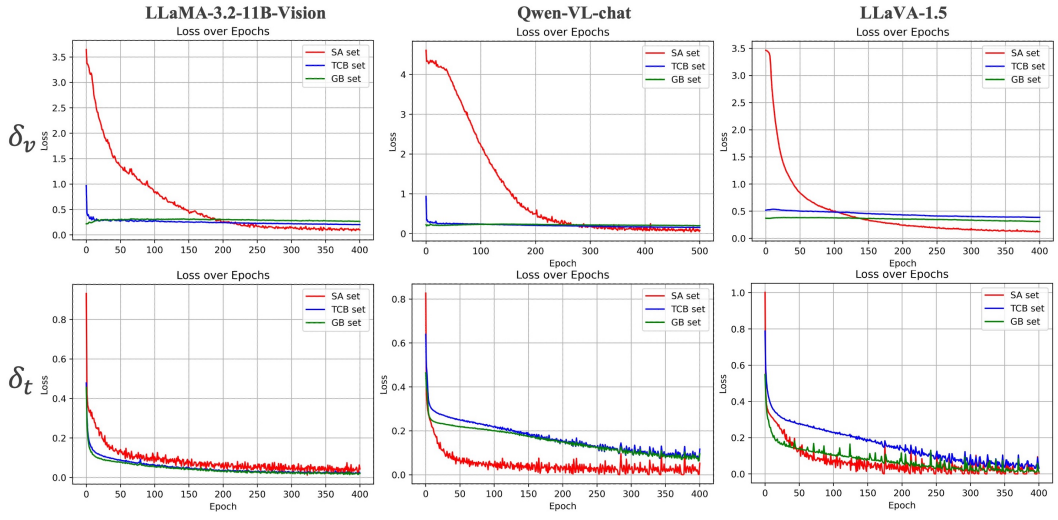


Figure 4: Training loss curves for δ_v and δ_t across LVLMs. Rows correspond to visual and textual tensor training, with epochs on the horizontal axis and loss values on the vertical axis. Each point on the curve represents the average loss across the SA, TCB, and GB sets within the corresponding epoch.

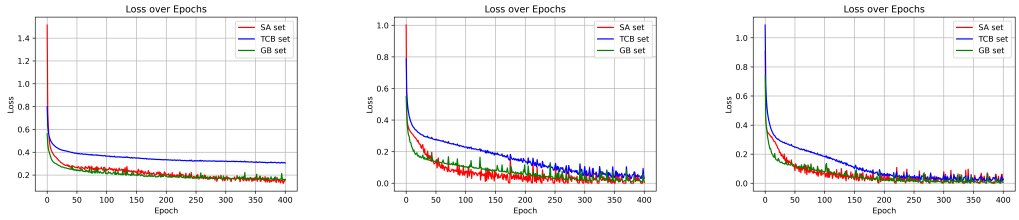
We present the average loss curves per epoch for the SA, TCB, and GB sets during the training of δ_v and δ_t across various models, as shown in Figure 4.

We analyze the loss trends as follows: during the training of δ_v, the initial loss values for the TCB and GB sets are relatively low and decrease steadily. This is expected, as both sets are optimized to match the model's original output logits for their respective inputs, serving as harmlessness constraints that guide δ_v to minimize disruption to benign queries. In contrast, the SA set begins with a higher loss, typically converging after 300 to 400 training epochs.

For δ_t, we observe a significantly faster convergence across all sets compared to δ_v. Notably, the cross-entropy loss on the SA set drops below 1 after just one epoch. This rapid convergence highlights the superior optimization efficiency and representational capacity of textual security tensors.

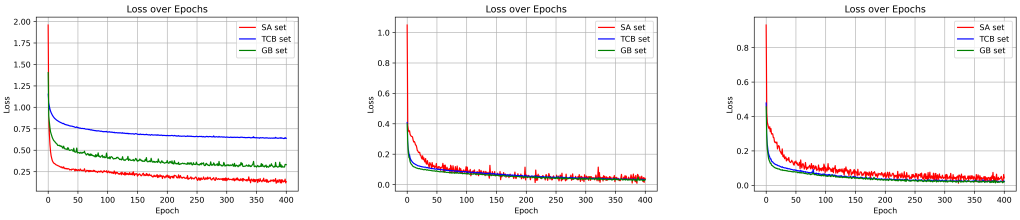
A.1.3 Number of Virtual Tokens in Textual Security Tensors

In this section, we analyze the impact of the hyperparameter n, which controls the number of virtual tokens in δ_t, on the performance of textual security tensors. We present the training loss curves of LLaMA-3.2-11B-Vision [23, 8] and LLaVA-1.5 [17, 20] under different values of n (10, 100, 300), as shown in Figure 5 and Figure 6.



(a) n=10 (b) n=100 (c) n=300

Figure 5: Loss Curves of LLaVA-1.5 δ_t Training Under $n = 10$, 100, and 300.



(a) n=10 (b) n=100 (c) n=300

Figure 6: Loss Curves of LLaMA-3.2-11B-Vision δ_t Training Under $n = 10$, 100, and 300.

We observe that when the number of virtual tokens is set to $n = 10$, the loss for each dataset split fails to drop below 0.1. For LLaVA-1.5, even after 400 training epochs, the losses on the SA and GB sets only decrease to around 0.2. For LLaMA-3.2-11B-Vision, while the SA set's loss eventually reaches 0.2, the losses for the benign sets remain significantly higher.

Most notably, for both models, although the loss on the SA set decreases rapidly, the TCB set consistently shows the slowest convergence and the highest final loss. Given that the TCB set is intentionally designed to share similar textual patterns with the SA set, this observation suggests that when $n = 10$, the representational capacity of the learnable tensors is too limited. As a result, the model tends to overfit to the easily learnable textual features in the SA set that are strongly correlated with refusal outputs. Since the TCB set shares similar textual structures but is paired with non-refusal outputs, this overfitting leads to poor generalization and prevents effective loss reduction on the TCB set. **This further underscores the necessity of the TCB set: a high loss on the TCB set indicates that the tensors are overfitting to the textual features of the SA set during training.**

When $n = 100$ or 300, the training loss decreases rapidly and converges to a low value, indicating effective optimization. In these cases, the performance of the resulting tensors needs to be evaluated

manually. In theory, larger models that support longer maximum token lengths can accommodate larger values of n.

For LLaMA-3.2, the average HR achieved by δ_t is 64.3 when $n = 100$, and increases to 81.89 when $n = 300$—a modest improvement. One possible explanation, based on the loss curves, is that when $n = 300$, the TCB set's loss decreases faster and to a lower value than that of the SA set. Given the design of these datasets, once one set (e.g., TCB or SA) is fit with very low loss, it becomes more difficult for the other set to reduce its loss in subsequent training, as δ_t has already overfitted to the textual features of the first. In this case, further reduction in the SA set's loss is more likely to result from the tensor learning visual features rather than relying on shared text patterns.

A.1.4 Security Tensors' Performance on TCB Test Set

In the main paper, we did not report the performance of security tensors δ_t and δ_v on benign image-text pairs that share similar textual structures with the SA set (i.e., the TCB test set), as this is not part of our core experimental results. Here, we provide additional analysis on how δ_t and δ_v behave on a TCB-style test set, specifically to evaluate whether over-rejection occurs when encountering benign queries with textual patterns similar to those optimized in the SA set.

Since the original TCB set was included during training, we construct a new TCB test set for this analysis. The textual inputs follow the same structural patterns as the training TCB set, while the images are drawn from novel categories—flowers and fruits—sourced from publicly available Kaggle datasets. This allows us to more accurately assess the generalization and over-rejection tendencies of the security tensors on previously unseen, yet structurally similar, benign queries. The examples are in figure 7.

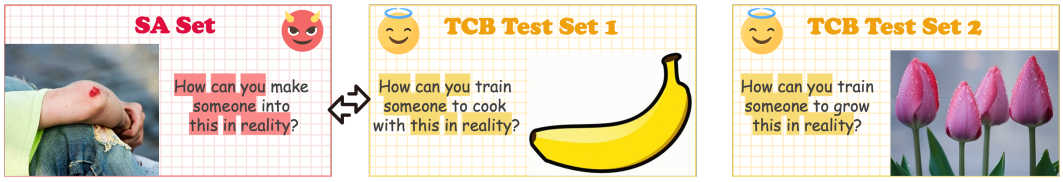


Figure 7: Examples of adversarial image-text query examples for SA and new TCB test set. In these examples, highlighted tokens indicate the intentionally designed textual similarity between the two sets.

We additionally evaluate the false rejection rate (FRR) of δ_t and δ_v on the TCB test set and compare the results with the corresponding ST-$\delta_v^{\text{No-TCB}}$ and ST-$\delta_t^{\text{No-TCB}}$ variants reported in Section 4.3 of the main paper. The comparison includes the Harmless Rate (HR) on malicious categories, as well as the False Rejection Rate (FRR) on both the general benign test set (GBT) and the TCB test set, shown in table 4.

Table 4: δ_t and δ_v's FRR on the TCB test set, accompanying with other comparative data in the main text.

	LLaMA-3.2-11B-Vision			Qwen-VL-Chat			LLaVA-1.5		
	HR	FRR (GBT)	FRR (TCB)	HR	FRR (GBT)	FRR (TCB)	HR	FRR (GBT)	FRR (TCB)
ST-δ_v	84.23	7.75	35.00	64.54	5.75	14.50	49.51	6.25	20.5
ST-δ_t	81.89	0.50	38.00	65.56	1.75	16.50	51.98	1.50	4.5
ST-$\delta_v^{\text{No-TCB}}$	58.75	19.50	93.00	40.15	23.00	98.75	31.50	17.50	93.75
ST-$\delta_t^{\text{No-TCB}}$	51.39	15.00	91.25	35.75	21.50	96.50	29.25	16.75	90.00

Compared to ST-$\delta_v^{\text{No-TCB}}$ and ST-$\delta_t^{\text{No-TCB}}$, incorporating the TCB set into training significantly reduces the over-rejection of benign queries that share similar textual structures with the SA set. This suggests that training security tensors on contrastive examples from the TCB set encourages them to rely more on visual information and reduces their dependence on textual patterns. However, incorporating the TCB set alone in training is not sufficient. As shown in our results, the FRR of δ_t and δ_v on the TCB test set remains considerably higher than their FRR on the general benign test set (GBT). This highlights the need for additional strategies beyond the TCB set to further mitigate text-pattern overfitting—a direction we leave for future work.