Semantic Shield: Defending Vision-Language Models Against Backdooring and Poisoning via Fine-grained Knowledge Alignment

Supplementary Material

1. Introduction

We offer supplementary results complementing those presented in our main paper. Initially, we describe the process of generating knowledge elements from Vicuna [2] in Sec. 2. Subsequently, we conduct a comparison between our model and a new model, RoCLIP [6], as outlined in Sec. 3. Finally, we provide additional qualitative analysis comparing the poisoned model to our model, demonstrating attention maps in Sec. 4.

2. Knowledge Elements Generation from Vicuna

Additionally, our approach incorporates external visual information related to the image to disrupt any spurious correlation between a backdoor trigger and caption. We employ an open-source large language model, *Vicuna-13b* [2], for this purpose. Vicuna is fine-tuned using approximately 70,000 user-shared conversations collected from ShareGPT.com via public APIs and has demonstrated strong performance in following specific instructions. In this task, we formulate a generic prompt template for each image-caption pair, aligning it with the paired caption. Alongside the prompt, 2-3 sample output formats are provided to ensure that the generated knowledge elements adhere to the same format. For instance, the input prompt for Vicuna is as follows:

```
   Given an image caption, extract a
list of distinct, low-level visual
attributes or sub-object elements
present in the image.  The goal is to
identify specific visual properties
or components that may characterize
the depicted scene.  Please generate
at least 4-5 descriptive elements that
could be associated with the visual
content in the image.

Example Caption:  A golden retriever
playing fetch on a green field under
the sun.
Expected Output:  Golden fur, Fetching
object (ball/stick), Green grass,
Bright sunlight, Sharp teeth/paws

Example Caption:  A man in hitting a
ball with a baseball bat.
Example Output:  White ball with red
stitching, Wooden bat, Man wearing a
helmet, Green grass/ Infield dirt

Example Caption:  {Input caption}
Example Output:
```

Some sample examples of image, caption, and corresponding knowledge elements are shown in Tab. 1. We use visual encoder (ViT) and text encoder (BERT) to compute the similarity between image and KEs and take top 5 KEs per sample. The ViT and text encoders are finetuned on some small proxy image text pair dataset for computing the similarity.

3. Dataset & Additional Experiments

3.1. Dataset

We evaluate our models and all other baselines on COCO [3] and Flickr30k [8]. Here, one image has five corresponding captions. For COCO, we have 5000 test images and for Flickr30k, train, validation, and test split are 80%, 10%, 10%. Dataset statistics are shown in Tab. 2

3.2. Comparison with RoCLIP [6]

RoCLIP [6] is concurrent work (unpublished as of date of submission), which proposes to make CLIP more robust to adversarial attacks. RoCLIP works by performing augmentation on images and captions and performing contrastive learning between augmented modalities. RoCLIP's augmentation technique for captions involves replacing words in captions with synonyms, randomly swapping words, and deletion of words. For images, they use random crops, grayscale conversion, color jittering, etc. to augment images. In contrast, our approach does not rely on augmenting images or captions, but instead encourages our transformer to attend to regions that align with external knowledge related to captions or objects by imposing a penalty term on the model's attention. Our approach relies on looking for lower-level knowledge related to captions or objects.

As of the date of our submission, RoCLIP's code is not publicly available and can not be directly compared with ours. However, we reimplemented RoCLIP's method in order to compare it with our approach. Following [1], we poisoned RoCLIP model similar to ours (0.01% of sampled data is backdoored or poisoned). For these experiments, we follow the same settings that we followed in main paper. We tested the models on randomly selected 100 backdoored

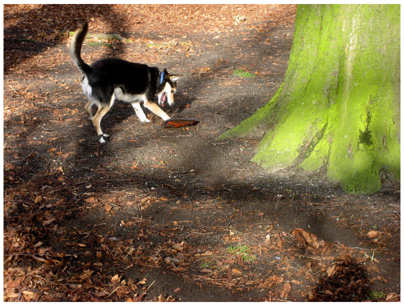
Image	Caption	Knowledge Elements
	A passenger bus pulling up to the side of a street	• Multiple wheels, usually in pairs • Entrance and exit doors • Headlights, turn lights • Bus logo, signs, pedestrian sidewalk
	A black and brown dog digging at an object on a dirt ground.	• Sharp paw • Long tail • Fur coat • Sharp nails and teeth
	A red fire hydrant is sitting in the woods near fallen leaves.	• Red color • Cylindrical body and valves on top • Color of leaves (e.g., shades of brown, red, or orange) • Trees, woods
	A para sailor with his board with sail in the surf.	• Clear or murky water • Parasailing harness and safety gear • Man wearing costume

Table 1. Sample image, caption, and corresponding knowledge elements

Dataset	# Pairs	# Images	# Labeled images	Category
COCO	616,767	123,287	122,218	80
Flickr30k	158,915	31,873	-	-

Table 2. Dataset statistics

images and poisoned samples.

We notice that from Tab. 3, our model outperforms in all settings except the backdoor patch, where the performance is very competitive (0.0 vs 0.9 in Hit@1). However, the utility of RoCLIP is not good (58.74 vs 74.22). Moreover, for single target label and multiple target label attack, our

Dataset	Models	Backdoor Patch			Backdoor BPP			Backdoor Wanet		
		Hit@1 ↓	Hit@5↓	Hit@10↓	Hit@1 ↓	Hit@5↓	Hit@10↓	Hit@1↓	Hit@5↓	Hit@10↓
COCO	CL (No Defense)	90.66	94.60	95.43	100.0	100.0	100.0	100.0	100.0	100.0
	CL+ RoCLIP [6]	0.7	1.57	1.94	0.0	0.0	0.0	0.0	0.0	0.0
	CL + KE	9.0	15.31	21.90	25.39	47.98	50.12	12.21	56.79	88.38
	CL + Attention	4.20	5.12	6.01	0.0	5.26	36.21	0.0	2.10	7.20
	Weighted CL + Attention	**0.9**	**1.22**	**1.57**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Flickr30k	CL (No Defense)	91.97	97.63	98.21	100.0	100.0	100.0	100.0	100.0	100.0
	CL+ RoCLIP [6]	1.2	1.32	1.56	2.23	9.21	9.91	1.21	3.13	9.21
	CL + KE	16.10	33.15	41.09	13.14	36.54	56.27	23.36	41.21	47.43
	CL + Attention	1.20	3.12	3.01	0.0	7.24	23.17	0.0	12.01	14.07
	Weighted CL + Attention	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**

Table 3. Backdoor attack and defense performance with baselines. The first row of the table shows an undefended model while other rows are baselines or variants of our method. CL+ KE, CL+ Attention are our baselines. The best results are shown in bold.

Dataset	Models	Single Target Label			Multiple Target Label					
		dog2boat			dog2boat			train2zebra		
		Hit@1 ↓	Hit@5↓	Hit@10↓	Hit@1 ↓	Hit@5↓	Hit@10↓	Hit@1 ↓	Hit@5↓	Hit@10↓
COCO	CL (No Defense)	18.0	57.20	82.0	77.12	99.23	99.56	55.32	95.76	97.98
	CL+ RoCLIP [6]	2.12	4.51	7.65	3.45	7.23	10.72	4.49	14.51	17.72
	CL + KE	4.56	5.32	5.95	54.45	64.21	85.52	65.12	70.92	86.12
	CL + Attention	0.56	3.38	4.51	**0.63**	65.60	69.42	2.25	6.77	12.99
	Weighted CL + Attention	**0.04**	**1.12**	**2.54**	2.23	**5.21**	**6.45**	**0.0**	**0.0**	**0.0**
		dog2boat			dog2boat			bird2sofa		
		Hit@1 ↓	Hit@5↓	Hit@10↓	Hit@1 ↓	Hit@5↓	Hit@10↓	Hit@1 ↓	Hit@5↓	Hit@10↓
Flickr30k	CL (No Defense)	29.0	57.20	82.23	28.12	82.39	93.76	55.32	90.62	100.0
	CL+ RoCLIP [6]	1.23	5.51	14.61	11.69	16.27	20.75	12.42	14.11	19.13
	CL + KE	7.34	28.09	32.21	21.12	45.32	47.67	12.77	42.34	54.21
	CL + Attention	4.56	21.81	34.11	**1.63**	16.70	29.21	3.25	18.43	32.22
	Weighted CL + Attention	**0.32**	**1.21**	**2.54**	1.78	**4.56**	**5.67**	**0.0**	**0.0**	**0.0**

Table 4. Poisoning attack and defense performance with baselines. First row of the table shows how good the attack, and other rows are baselines along with our proposed models. CL+ KE, CL+ Attention are our baselines. The best results are highlighted.

Dataset	Multiple target label attack			
	dog2boat	train2zebra	horse2sheep	chair2sandwich
COCO	2.23	0.0	1.1	0.9
Flickr30k	1.78	0.0	0.0	0.0

Table 5. Performance of Hit@1 of Weighted CL + Attention on COCO and Flickr30k in multiple target label attack. We experiment on two additional target label.

model outperforms RoCLIP by a large margin both in defense (Tab. 4, Tab. 6). Therefore, for image-text retrieval, our model is preferable compared to RoCLIP for defending all types of poisoning and backdooring attacks.

3.3. Experiments on different numbers of KE

We experiment with different numbers of KEs per sample (3, 5 (main paper) and 7) KEs in both the COCO and Flick30k datasets. In Tab. 7, we present the Hit@1 performance of our Weighted CL + Attention model (best proposed model). In main paper, we experimneted with 5 KE per sample. Additionally, we experimented with 3 and 7 KE per samples, and it shows the performance remains similar which indicates varying number of KEs has no significant impact of the model performance.

3.4. Different LLM for KE generation

We tried Mistral-7b-instruct-v0.1 for KE generation instead of Vicuna using the same prompt. The KEs generated are similar to those from Vicuna. We retrained Weighted CL + Attention using KEs from Mistral and gained similar performance in Tab. 8. Therefore, changing LLM for KE generation does not change model performance.

3.5. Poisoning attack with other attack targets

Following [7], we experiment on single target label `dog2boat` and multiple target label attack `dog2boat` and `train2zebra`. In multiple target label attack, instead of two target labels, we added two additional target labels (`horse2sheep`, `chair2sandwich`) in Tab. 5.

Dataset	Task	Models	Backdoor Patch	BPP	Wanet	Single Target Label	Multiple Target Label
COCO	IR	CL	74.99	73.94	74.54	74.68	74.72
		RoCLIP	58.74	61.68	58.38	58.66	54.43
		CL + KE	74.15	70.7	74.0	74.24	73.28
		CL + Attention	74.38	73.13	74.43	75.70	75.13
		Weighted CL + Attention	74.22	74.56	74.23	73.46	73.51
COCO	TR	CL	81.58	77.44	78.74	80.16	81.12
		RoCLIP	56.21	56.69	54.80	55.97	54.17
		CL + KE	78.40	75.54	77.86	79.08	81.20
		CL + Attention	79.20	77.36	78.04	80.05	81.06
		Weighted CL + Attention	79.46	77.78	78.45	79.67	80.0
Flickr30k	IR	CL	59.13	59.86	61.08	60.92	57.41
		RoCLIP	47.8	48.41	45.21	50.27	45.43
		CL + KE	60.34	61.85	61.13	58.12	58.18
		CL + Attention	61.32	55.96	59.14	58.97	58.16
		Weighted CL + Attention	61.07	56.32	60.16	59.76	58.78
Flickr30k	TR	CL	68.07	68.79	69.86	71.06	68.14
		RoCLIP	43.12	46.43	52.45	51.23	52.59
		CL + KE	69.67	70.65	69.62	66.98	62.20
		CL + Attention	70.0	64.46	68.0	68.13	62.97
		Weighted CL + Attention	70.23	65.66	68.87	68.45	62.12

Table 6. Comparison between CL, RoCLIP and our defended models' utility (Recall@10).

Dataset	# KE	Patch	BPP	Wanet	SingleTL	MultiTL	
					dog2boat	dog2boat	train2zebra
COCO	3	0.9	0.0	0.0	0.05	2.10	0.0
	5	0.9	0.0	0.0	0.04	2.23	0.0
	7	0.1	0.10	0.0	0.05	2.30	0.2
Flickr30k	3	0.0	0.0	0.0	0.30	1.76	0.0
	5	0.0	0.0	0.0	0.32	1.78	0.0
	7	0.1	0.0	0.0	0.36	1.79	0.1

Table 7. Performance of Hit@1 of Weighted CL + Attention on COCO and Flickr30k. Here the number of KEs are 3, 5, and 7 per sample.

Dataset	LLM	Patch	BPP	Wanet	SingleTL	MultiTL	
					dog2boat	dog2boat	train2zebra
COCO	Mistral	0.9	0.0	0.0	0.05	2.25	0.0
	Vicuna	0.9	0.0	0.0	0.04	2.23	0.0
Flickr30k	Mitral	0.0	0.0	0.0	0.34	1.81	0.0
	Vicuna	0.0	0.0	0.0	0.32	1.78	0.0

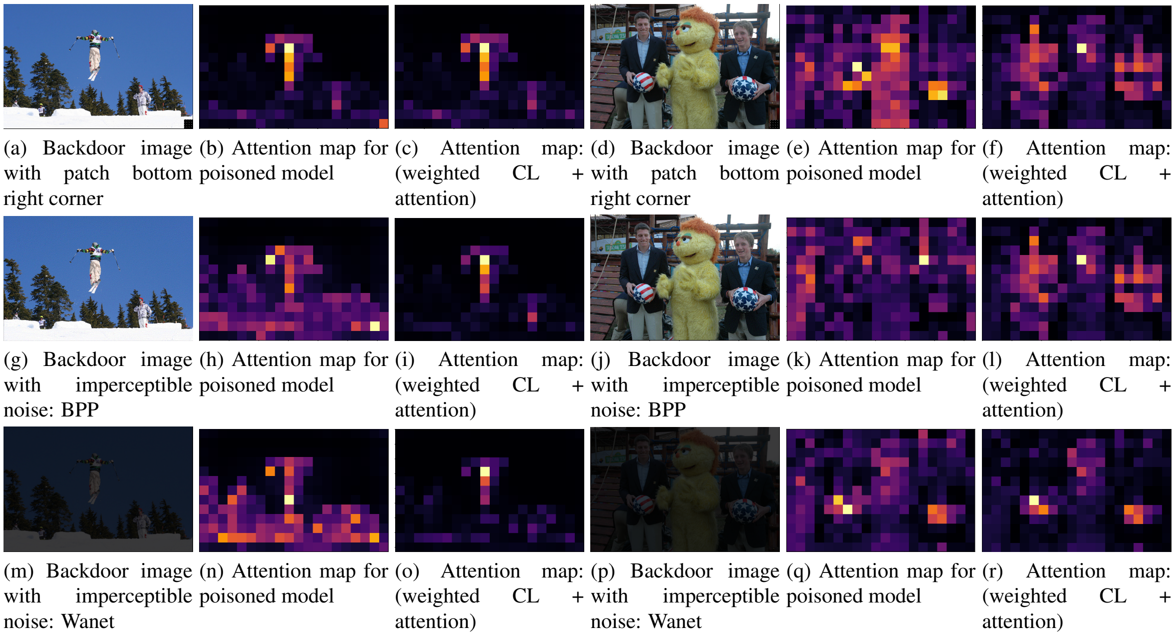
Table 8. Performance of Hit@1 of Weighted CL + Attention on COCO and Flickr30k. We compare Mistral with Vicuna. In main paper we reported KEs generated from Vicuna.

3.6. Quality of KEs.

We empirically assess the quality of KEs generated from LLMs *e.g.* Vicuna, Mistral. We formulated two tasks for this, KE2caption and caption2KE. In the KE2caption task, we prompt LLMs with KEs to find the correct caption from 5 random captions for each sample data. Similarly, in caption2KE tsk, we prompt LLMs with a caption to retrieve the correct KE set from 5 randomly generated KE set. We obtain an accuracy of 91% for KE2caption and 93% for cap-

Dataset	Loss	Patch	BPP	Wanet	SingleTL	MultiTL	
					dog2boat	dog2boat	train2zebra
COCO	$\mathcal{L}_{WeightedCL+Attention}$	0.9	0.0	0.0	0.04	2.23	0.0
	$\mathcal{L}_{WeightedCL+KE}$	1.9	0.8	0.7	0.18	3.10	1.11
	$\mathcal{L}_{WeightedCL+Attention+KE}$	1.1	0.5	0.4	0.09	2.9	0.8
					dog2boat	dog2boat	bus2sofa
Flickr30k	$\mathcal{L}_{WeightedCL+Attention}$	0.0	0.0	0.0	0.32	1.78	0.0
	$\mathcal{L}_{WeightedCL+KE}$	0.95	1.5	1.4	1.2	2.21	1.1
	$\mathcal{L}_{WeightedCL+Attention+KE}$	1.2	0.5	0.7	0.80	1.98	0.90

Table 9. Perfromance of Hit@1 on COCO and Flickr30k in diffrent combination of loss functions. The top row is the best proposed model.



(a) Backdoor image with patch bottom right corner
(b) Attention map for poisoned model
(c) Attention map: (weighted CL + attention)
(d) Backdoor image with patch bottom right corner
(e) Attention map for poisoned model
(f) Attention map: (weighted CL + attention)
(g) Backdoor image with imperceptible noise: BPP
(h) Attention map for poisoned model
(i) Attention map: (weighted CL + attention)
(j) Backdoor image with imperceptible noise: BPP
(k) Attention map for poisoned model
(l) Attention map: (weighted CL + attention)
(m) Backdoor image with imperceptible noise: Wanet
(n) Attention map for poisoned model
(o) Attention map: (weighted CL + attention)
(p) Backdoor image with imperceptible noise: Wanet
(q) Attention map for poisoned model
(r) Attention map: (weighted CL + attention)

Figure 1. Attention map comparison between our model (weighted CL + attention) and poisoned models for three backdoor attacks.

tion2KE tasks. We experiment with 1000 randomly selected captions and a 1000 KE set, where each KE set contains 5 KEs.

3.7. Combination of loss function

In Tab. 9, we report different combinations of our proposed losses. Our best model is $\mathcal{L}_{WeightedCL+Attention}$. Adding $\mathcal{L}_{KE}$ slightly hurts the model performance since all patches are not aligned with KEs.

3.8. Computational overhead for KE generation

If we use Mistral-7b for KE generation, for Flickr30k it takes only 30 min (a one time process). For training, the KE overhead is only one additional forward pass for all KEs at once per iteration. In practice, train time is roughly similar to the baseline.

4. Qualitative Analysis

In Fig. 1, we present illustrative examples of attention maps for both backdoored models and our defended model (weighted CL + attention). In Fig. 1a and Fig. 1d, a small backdoor trigger is introduced to the bottom right corner of the image. It is evident that the backdoored models focus their attention on the backdoor trigger in Fig. 1b and Fig. 1e (highlighting the bottom right part of the image). Conversely, our defended model (weighted CL + attention) exhibits no attention in the bottom right part of the images, as observed in Fig. 1c and Fig. 1f.

Next, we conducted experiments on our model's performance with two types of visually imperceptible examples, namely BPP [5] and WANet [4] (refer to Fig. 1g, Fig. 1j, Fig. 1m, Fig. 1p). We introduced noise throughout the images to deceive traditional visual language models, as these models tend to pay attention all over the images (Fig. 1h, Fig. 1k for BPP, Fig. 1n, Fig. 1q for Wanet). In contrast, our models focus their attention on the expected regions, ignoring all noisy areas of the images (Fig. 1i, Fig. 1l for BPP, Fig. 1o, Fig. 1r for Wanet).

References

[1] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 1

[2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 1(2):3. 1

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[4] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021. 5

[5] Zhenting Wang, Juan Zhai, and Shiqing Ma. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15054–15063. IEEE, 2022. 5

[6] Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Robust contrastive language-image pretraining against data poisoning and backdoor attacks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 3

[7] Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. Data poisoning attacks against multimodal encoders. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 39299–39313. PMLR, 2023. 3

[8] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014. 1