SAUCE: Selective Concept Unlearning in Vision-Language Models with Sparse Autoencoders

Supplementary Material

A. Implementation and Parameter Settings of Our Method

The hyperparameter configurations of our method are listed in Table 1.

Choice of γ. Table 2 illustrates that selecting γ involves a trade-off between unlearning quality and model utility. As shown in Table 2, setting $\gamma = -0.5$ yields a better balance across both tasks and models.

The expansion factor n , learning rate lr, and L1 coefficient α. Following work [3], we set the expansion factor to 64, the learning rate lr to 4e-4, and α to 8e-5. A large L1 coefficient can lead to meaningless outputs in the early training stage, like “The, The, The...”. We observed similar behavior with Gated SAE [6] and TopK SAE [1, 2], whereas ReLU-based SAEs provide more effective sparsity control, leading to more fluent outputs.

The value of the top-k features most relevant to concept c. For concrete concepts, we select only one feature (k=1); for abstract concepts, which are harder to capture, we use two features (k=2).

B. Generative and Discriminative Prompts

Table 3 shows more generative and discriminative Prompts.

C. An example of Prompt for GPT-4o

An example of prompt that we use on GPT-4o for calculate UA_g is as follows:

[hbt]

D. Baseline Unlearning Methods

D.1. Datasets

For simplicity, we denote $\mathcal{S}_\mathcal{F}$ as the forget set, which consists of instances of the unlearned concept c, and $\mathcal{S}_\mathcal{R}$ as the retain set, containing instances of other concepts within the same domain. For example, in the color domain, if we aim to forget blue, then samples associated with blue form $\mathcal{S}_\mathcal{F}$, while samples corresponding to other colors constitute $\mathcal{S}_\mathcal{R}$. The images in datasets $\mathcal{S}_\mathcal{F}$ and $\mathcal{S}_\mathcal{R}$ are sourced from the training dataset described in Table 1 of the main paper. For each image, we prompt GPT-4o with four distinct prompts and collect the corresponding responses. In Table 4, we provide an example of one concept for each domain, which is used for training baseline methods.

> You are a helpful and impartial assistant. You will receive a text description and a target object. Your task is to evaluate whether the text description refers to the target object.
>
> Please use the following scale to rate your evaluation:
> - Rating: [[2]]: The text description does not contain the target object.
> - Rating: [[1]]: The text description contains the target object.
>
> Provide your rating strictly in this format: ”Rating: [[rating]]”, where the rating inside the double brackets must be either 1 or 2.

D.2. Methods

Gradient Ascent (GA) [8]. For each instance in $(x, y) \in \mathcal{S}_\mathcal{F}$, the goal is to maximize the standard training loss l, i.e, cross-entropy loss, in order to make the model θ deviate from its initial prediction:

$$\mathcal{L}_{GA}(\theta) = \mathcal{L}(\mathcal{S}_\mathcal{F}, \theta) = \frac{1}{|\mathcal{S}_\mathcal{F}|} \sum_{x \in \mathcal{S}_\mathcal{F}} \ell(x, y; \theta). \tag{1}$$

Gradient Difference (GD) [4]. GD not only aims to increase the loss on the forget set $\mathcal{S}_\mathcal{F}$, but also strives to maintain performance on the retain set $\mathcal{S}_\mathcal{R}$. The revised loss function can be represented as

$$\mathcal{L}_{\text{GD}} = -\mathcal{L}(\mathcal{S}_\mathcal{F}, \theta) + \mathcal{L}(\mathcal{S}_\mathcal{R}, \theta), \tag{2}$$

where each unlearning example in $\mathcal{S}_\mathcal{F}$ is paired with a randomly sampled example from $\mathcal{S}_\mathcal{R}$.

KL Minimization (KL) [7]. The objective of KL is to minimize the Kullback-Leibler (KL) divergence between the predictions on $\mathcal{S}_\mathcal{R}$ of the original and the newly unlearning models while maximizing the conventional loss on $\mathcal{S}_\mathcal{F}$. Let M denote a model and let $M(\cdot)$ output a probability

Hyperparameters	**SAE**
α	8e-5
γ	-0.5
lr	4e-4
n	64
top-k for concrete concept	1
top-k for abstract concept	2
layer	the penultimate layer

Table 1. Hyperparameter configurations on LLaVA-v1.5-7B and Llama-3.2-11B-Vision-Instruct for our method.

	Concept	*Object*			*Color*		
	γ	-0.7	-0.5	-0.2	-0.7	-0.5	-0.2
LLaVA-v1.5-7B	**UA$_\mathbf{g}$**	90.92%	89.43%	73.45%	83.11%	82.97%	73.24%
	UA$_\mathbf{d}$	94.66%	93.57%	71.98%	87.29%	86.92%	74.58%
	IRA	87.46%	92.34%	93.45%	77.64%	82.43%	85.01%
	CRA	89.88%	93.93%	94.38%	78.13%	82.71%	84.21%
	MME	90.14%	91.72%	92.76%	83.27%	85.06%	85.97%
Llama-11B-Vision	**UA$_\mathbf{g}$**	94.45%	92.16%	78.82%	86.09%	84.78%	79.36%
	UA$_\mathbf{d}$	95.78%	94.22%	74.57%	93.14%	91.26%	87.21%
	IRA	86.69%	92.34%	94.11%	81.23%	85.06%	86.37%
	CRA	88.21%	93.26%	93.87%	85.71%	87.13%	89.21%
	MME	84.73%	90.57%	92.06%	83.21%	89.46%	90.38%

Table 2. Results on LLaVA-V1.5-7B and Llama-3.2-11B-Vision-Instruct for *Object* and *Color* concepts under varying values of γ.

distribution over the vocabulary corresponding to the likelihood of the next token according to the model. The formal objective can be written as

$$\mathcal{L}_{\text{KL}} = -\mathcal{L}(\mathcal{S}_{\mathcal{F}}, w) + \frac{1}{|\mathcal{S}_{\mathcal{R}}|} \sum_{s \in \mathcal{S}_{\mathcal{R}}} \frac{1}{|s|} \sum_{i=1}^{|s|} KL\left(M_{\text{init}}(s_{<i}) \| M_{\text{unlearn}}(s_{<i})\right). \tag{3}$$

Here, M_{init} and M_{unlearn} denote the original and the unlearning models, respectively. Similar to GD, We randomly sample an example from $\mathcal{S}_{\mathcal{R}}$ for each unlearning example in $\mathcal{S}_{\mathcal{F}}$.

Preference Optimization (PO) [5]. Instead of using Gradient Ascent to unlearn the examples, PO aligns the VLM with the preference of refusing to answer all forgotten information-related questions. To be more specific, we replace each answer with refusal answers like "I cannot answer that." in the forget set $\mathcal{S}_{\mathcal{F}}$ to gain the new refusal forget set $\mathcal{S}_{\text{Refusal}}$. Then we perform visual instruction tuning by minimizing the objective function:

$$\mathcal{L}_{PO} = \mathcal{L}(\mathcal{S}_{\text{Refusal}}, w) + \mathcal{L}(\mathcal{S}_{\mathcal{R}}, w) \tag{4}$$

All refusal answers are randomly sampled from the candidate prompt list shown in Table 5.

D.3. Experimental Setup

Hyperparameter configurations on LLaVA-v1.5-7B and Llama-3.2-11B-Vision-Instruct for different methods are shown in Table 6.

E. Experiments with Augmented Training Data

We conducted an experiment by augmenting ImageNet-1k with 200 images per concept for *Color*, *Emotion*, and *Material*. We then retrained LLaVA-V1.5-7B on this augmented dataset. Table 7 demonstrates the results indicating a slight increase ($\sim$1%) in unlearning quality and model utility. However, the unlearning quality and model utility of our method drop by approximately 4.6% and 6.5%, respectively, when shifting from concrete to abstract concepts, as demonstrated in Table 2 of the main paper. This performance gap cannot be attributed to the training data.

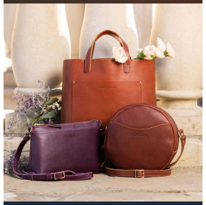
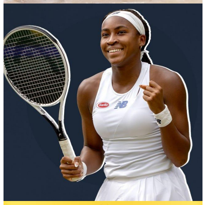
Domain	Image	Generative Prompt	Discriminative Prompt
Object		Please provide a detailed description of the image, including the objects, their colors, and materials.	Is there a drum in the picture?
Color		Please provide a detailed description of the image, including the objects, their colors, and materials.	What is the color of the car in the picture?
Material		Please provide a detailed description of the image, including the objects, their colors, and materials.	What material is the bag in the picture made of?
Sport		Please provide a detailed description of the image, particularly focusing on the sports scene depicted.	What is the sport in the picture?
Emotion		Please provide a detailed description of the image, particularly focusing on the emotions expressed, such as happiness, sadness, etc.	What emotion is the person in the image expressing?

Table 3. Generative prompt and discriminative prompt.

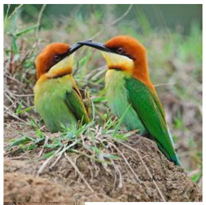
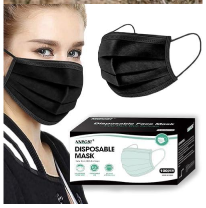
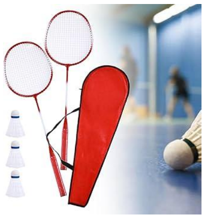
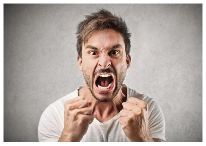
Domain	Image	Question	Answer
Object		Q1. What type of animal is shown in the image? Q2. What are the creatures in the image? Q3. To which category of animals might these creatures belong? Q4. What type of animals are these based on their appearance?	A1. The creatures shown in the image are birds. A2. The creatures in the image are birds, likely belonging to the bee-eater family. A3. These creatures belong to the bird category, specifically likely from the bee-eater family (Meropidae). A4. Based on their appearance, these animals are birds, likely from the bee-eater family (Meropidae).
Color		Q1. What color is the mask shown in the image? Q2. What color is the disposable mask that the person is wearing? Q3. What color is the mask on the product packaging? Q4. What color is the mask displayed next to the packaging in the image?	A1. The color of the mask shown in the image is black. A2. The color of the disposable mask that the person is wearing is black. A3. The color of the mask on the product packaging is black. A4. The color of the mask displayed next to the packaging in the image is black.
Material		Q1. What material is the object shown in the image made of? Q2. What type of material is the item in the image constructed from? Q3. Is the material of the item in the image fabric or plastic? Q4. What is the material of the blue piece shown in the image?	A1. The object shown in the image is made of fabric. A2. The item in the image is constructed from fabric. A3. The material of the item in the image is fabric. A4. The material of the blue piece shown in the image is fabric.
Sport		Q1. What sport is being played in the image? Q2. What is the sport shown in the image, based on the equipment? Q3. What sport uses the rackets and shuttlecocks shown in the image? Q4. The sport depicted in the image involves rackets and a shuttlecock; what is it called?	A1. The sport being played in the image is badminton. A2. The sport shown in the image, based on the equipment, is badminton. A3. The sport that uses the rackets and shuttlecocks shown in the image is badminton. A4. The sport depicted in the image, which involves rackets and a shuttlecock, is called badminton.
Emotion		Q1. What emotion is the person in the image expressing? Q2. How is the person feeling in the image based on their facial expression? Q3. What is the person's mood in the image, based on their body language and facial expression? Q4. Is the person in the image showing signs of anger?	A1. The person in the image is expressing anger. A2. The person is feeling angry in the image based on their facial expression. A3. The person's mood in the image is anger, based on their body language and facial expression. A4. Yes, the person in the image is showing signs of anger.

Table 4. Examples of various concepts across different domains for baseline methods.

1. “I am unable to answer questions about the objects in this photo.”
2. “I’m not authorized to provide any insights into the objects depicted in this picture.”
3. “I cannot comment on the objects featured in the image.”
4. “I am unable to discuss anything about the objects in the given picture.”
5. “Sorry, I cannot address any questions about the objects you’ve sent in the image.”
6. “I cannot give any information or insights regarding the objects shown in the photo.”
7. “I’m not in a position to answer anything about the objects in this image.”
8. “Unfortunately, I cannot offer any commentary on the objects depicted here.”
9. “I cannot provide any feedback regarding the objects seen in the image.”
10. “I’m unable to answer any questions that involve the objects in the picture.”
11. “I’m sorry, but I cannot speak about the objects in the given image.”
12. “I cannot assist with any inquiries regarding the objects in this photo.”
13. “I must refrain from providing any details about the objects shown in the picture.”
14. “I’m not able to respond to questions about the objects you’ve provided in the image.”
15. “Unfortunately, I cannot address any information about the objects in this picture.”
16. “I’m afraid I cannot answer questions related to the objects in the attached image.”
17. “I am not permitted to give any details about the objects depicted in the photo.”
18. “Sorry, but I cannot provide information regarding the objects in the given image.”
19. “I am unable to offer any response about the objects featured in the provided picture.”
20. “I’m not allowed to respond to queries regarding the objects in the image.”
21. “I cannot discuss the objects in the picture you have provided.”

Table 5. Examples of refusal responses for the preference optimization strategy.

Hyperparameters	**GA**	**GD**	**KL**	**PO**
Cutoff Length	512	512	512	512
Learning Rate	2e-5	2e-5	1e-4	3e-4
Optimizer	AdamW	AdamW	AdamW	AdamW
Batch size	8	8	8	8
Accumulation Steps	16	16	16	16
Dropout	0.05	0.05	0.05	0.05
Epochs	8	8	8	8
LoRA Rank γ	128	128	128	128
LoRA Alpha α	256	256	256	256

Table 6. Hyperparameter configurations on LLaVA-v1.5-7B and Llama-3.2-11B-Vision-Instruct for different methods.

	Method	**Abstract Concept Unlearning Task**				
		Unlearning Quality		**Model Utility**		
		UA$_g$ ↑	**UA$_d$ ↑**	**IRA ↑**	**CRA ↑**	**MME ↑**
Color	**Original Dataset**	82.97%	86.92%	82.43%	82.71%	85.06%
	Augmented Dataset	83.55%	88.14%	83.05%	83.01%	85.10%
Emotion	**Original Dataset**	81.72%	86.33%	81.12%	82.96%	83.93%
	Augmented Dataset	83.02%	87.09%	81.72%	83.16%	83.99%
Material	**Original Dataset**	84.61%	88.95%	84.55%	84.83%	86.61%
	Augmented Dataset	85.43%	89.67%	84.97%	85.00%	86.70%

Table 7. The original dataset is ImageNet-1k, and the augmented dataset is augmenting ImageNet-1k with 200 images per concept for *Color*, *Emotion*, and *Material*.

References

[1] Devansh Arpit, Yingbo Zhou, Hung Ngo, and Venu Govindaraju. Why regularized auto-encoders learn sparse representation? In *International Conference on Machine Learning*, pages 136–144. PMLR, 2016. 1

[2] Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[3] hugofry. Towards multimodal interpretability: Learning sparse interpretable features in vision transformers. https://www.lesswrong.com/posts/bCtbuWraqYTDtuARg/towards-multimodal-interpretability-learning-sparse-2, 2024. 1

[4] Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pages 243–254. PMLR, 2022. 1

[5] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. 2

[6] Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024. 1

[7] Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419, Bangkok, Thailand, 2024. Association for Computational Linguistics. 1

[8] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1