Alignment-aware Data Selection for Unlearning in Contrastive Vision-Language Models

Dongjun Hwang [1] **Yejin Kim** [2] **Beomyun Kwon** [1] **Junsuk Choe** [1]

Abstract

Recent advances in contrastive vision-language models have increased the need to selectively remove knowledge of specific data, drawing attention to *machine unlearning*. In this paper, we observe that unlearning performance in contrastive VLMs largely depends on the composition of the forget set. Based on this insight, we propose **ALISE**, a data selection framework that measures each forget sample's alignment with both the retain set and the full forget set, and selects samples accordingly. Extensive experiments across diverse downstream applications demonstrate that ALISE facilitates removing target knowledge in contrastive VLMs while preserving model utility.

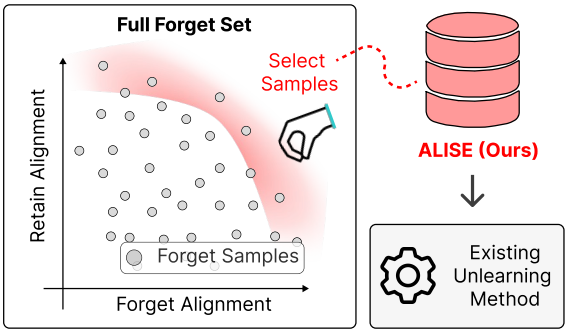

Figure 1. ALISE pipeline. It constructs a new forget set by selecting samples from the full forget set based on sample-level alignment.

1. Introduction

Contrastive Vision-Language Models (VLMs), such as CLIP (Radford et al., 2021), serve as core components in a wide range of systems, including Text-to-Image (T2I) generation models (Rombach et al., 2022; Yang et al., 2023) and Multimodal Large Language Models (MLLMs) (Liu et al., 2023). However, models trained on large-scale web data may contain information that poses privacy or copyright risks, raising safety concerns (Thiel, 2023).

The most direct solution to this problem is *exact unlearning*, which removes the target data from the training dataset and retrains the model. However, the computational cost of retraining is extremely high. For this reason, recent studies have focused on *approximate unlearning* (Nguyen et al., 2025), which aims to remove target knowledge from a pre-trained model without retraining it from scratch.

Existing approximate unlearning methods mainly differ in how they update the model parameters to remove knowledge in a forget set while preserving utility on a retain set. For example, Cai et al. (2025) selectively updates only the parameters that are more strongly associated with the forget set, while Zhong et al. (2025) optimizes the losses on the forget set and the retain set using separate optimizers.

Despite these efforts, approximate unlearning methods still struggles to remove only the target knowledge (Moon et al., 2025; Amara et al., 2025). This is because model parameters are highly entangled in a high-dimensional space, so suppressing the target knowledge through post hoc parameter updates can unintentionally affect other unrelated knowledge. Therefore, approximate unlearning inherently involves a trade-off between target removal and utility preservation (Ilharco et al., 2022).

In our experiments, we observe that unlearning performance under this trade-off strongly depends on the composition of the forget set. As shown in Figure G.1 in the Appendix, even with the same unlearning algorithm, different subsets randomly sampled from the full forget set lead to substantially different results. This observation suggests that the choice of samples for the forget set is a key factor in unlearning.

Motivated by this finding, we propose an ALIgnment-aware data SElection framework (**ALISE**) for unlearning in contrastive VLMs (Figure 1). For each forget sample, we quantify its alignment with both the retain set and the full forget set. Based on these alignment scores, ALISE selects samples to construct a new forget set for unlearning.

[1]Sogang University [2]KAIST. Correspondence to: Junsuk Choe <jschoe@sogang.ac.kr>.

Published at ICML 2026 Workshop on the Impact of Memorization on Trustworthy Foundation Models, Seoul, South Korea. PMLR 306, 2026.

We demonstrate the effectiveness of ALISE through extensive experiments. Specifically, we evaluate ALISE across four downstream applications of contrastive VLMs: open-vocabulary classification, image-text retrieval, MLLMs, and T2I generation, and assess its generality across two contrastive VLM backbones and five unlearning baselines. Experimental results show that, across all experimental settings, using the forget subset selected by ALISE yields better performance than using the full forget set. These results show that ALISE consistently improves the unlearning performance of existing methods without requiring modifications to the pipeline or architecture.

2. Method

In this section, we propose ALISE, a sample selection framework for unlearning in contrastive VLMs. We first define how to quantify the alignment between each forget sample and both the retain set and the full forget set, and then introduce a method for selecting samples to construct a new forget set based on these scores. The preliminaries required for understanding the proposed methodology are provided in Appendix B.

2.1. Sample-Level Alignment Scores

Retain Alignment Score $\mathbf{s_r}$. We quantify the alignment between a sample x_i and the retain set $\mathcal{D}_r$ as the cosine similarity between their representation gradients, g_i^{both} and g_R^{both}, derived from $\ell_f(x_i;\theta)$ and $\mathcal{L}_r(\theta;\mathcal{D}_r)$, respectively. Specifically, we define the retain alignment score as

$$s_r(x_i) = \frac{g_i^{\text{both}} \cdot g_R^{\text{both}}}{\|g_i^{\text{both}}\| \, \|g_R^{\text{both}}\|}. \tag{1}$$

This formulation follows prior data attribution studies (Li et al., 2026; Pruthi et al., 2020), which measure how a specific sample influences the model's output on other samples by leveraging the cosine similarity or inner product between gradients derived from their respective losses.

We now describe each term in the formulation. Let a sample from the forget set be denoted as $x_i = (x_i^{\text{img}}, x_i^{\text{text}}) \in \mathcal{D}_f$. The image and text encoders, including their projection heads, are denoted by f_{img} and f_{text}, respectively. The resulting embedding vectors are referred to as representation vectors, given by $f_{\text{img}}(x_i^{\text{img}}) = z_i^{\text{img}}$ and $f_{\text{text}}(x_i^{\text{text}}) = z_i^{\text{text}}$, where $z_i^{\text{img}}, z_i^{\text{text}} \in \mathbb{R}^d$.

For each modality $m \in \{\text{img}, \text{text}\}$, the representation gradient of a sample x_i is given by

$$g_i^m = \nabla_{z_i^m} \ell_f(x_i;\theta). \tag{2}$$

Similarly, based on the retain loss in Eq. (8), we define the representation gradient of each modality for the retain set as

$$g_R^m = \mathbb{E}_{\mathcal{B}\sim\mathcal{D}_r}\left[\frac{1}{|\mathcal{B}|}\sum_{x_j\in\mathcal{B}} \nabla_{z_j^m} \ell_r(x_j;\theta,\mathcal{B})\right]. \tag{3}$$

The gradients from the two modalities are then concatenated as follows:

$$g_i^{\text{both}} = [g_i^{\text{img}}; g_i^{\text{text}}], \quad g_R^{\text{both}} = [g_R^{\text{img}}; g_R^{\text{text}}].$$

Following prior studies on gradient interactions (Yu et al., 2020; Cai et al., 2025), we interpret that when $s_r(x_i) \approx 0$, the gradients are nearly orthogonal, and the sample is *weakly entangled* with the retain set. In contrast, when it is close to -1 or 1, the sample is *strongly entangled* with the retain set.

While s_r captures the alignment with the retain set, it does not capture how well the sample is aligned with the full forget set, which contains the target knowledge to be removed. Therefore, we define the forget alignment score s_f in the next paragraph to measure this alignment.

Forget Alignment Score $\mathbf{s_f}$. We quantify the alignment between a sample x_i and the full forget set $\mathcal{D}_f$ using the cosine similarity between their representation gradients, g_i^{both} and g_F^{both}. Specifically, we define the forget alignment score as

$$s_f(x_i) = \frac{g_i^{\text{both}} \cdot g_F^{\text{both}}}{\|g_i^{\text{both}}\| \, \|g_F^{\text{both}}\|}. \tag{4}$$

Here, following Eq. (9), we define the representation gradients of $\mathcal{L}_f(\theta;\mathcal{D}_f)$ for each modality as

$$g_F^m = \mathbb{E}_{x_i\sim\mathcal{D}_f}\left[\nabla_{z_i^m} \ell_f(x_i;\theta)\right], \qquad m \in \{\text{img}, \text{text}\}. \tag{5}$$

We then concatenate the gradients from the two modalities as $g_F^{\text{both}} = [g_F^{\text{img}}; g_F^{\text{text}}]$.

When $s_f(x_i)$ is close to 1, the sample x_i is *strongly aligned* with the full forget set $\mathcal{D}_f$, indicating that it effectively facilitates forgetting knowledge associated with $\mathcal{D}_f$. To jointly account for this alignment and the entanglement with the retain set, we introduce a framework in the next subsection that leverages both s_r and s_f for sample selection.

Detailed discussions on the alignment scores are provided in Appendix C.

2.2. ALISE: Alignment-aware Data Selection

We propose **ALISE**, a method that selects samples to construct a new forget set based on s_r and s_f. Specifically, we formulate sample selection as the following multi-objective optimization problem:

$$\text{find} \quad x \in \mathcal{D}_f \quad \text{to maximize} \quad \begin{aligned} f_1(x) &= 1 - |s_r(x)|, \\ f_2(x) &= s_f(x). \end{aligned} \tag{6}$$

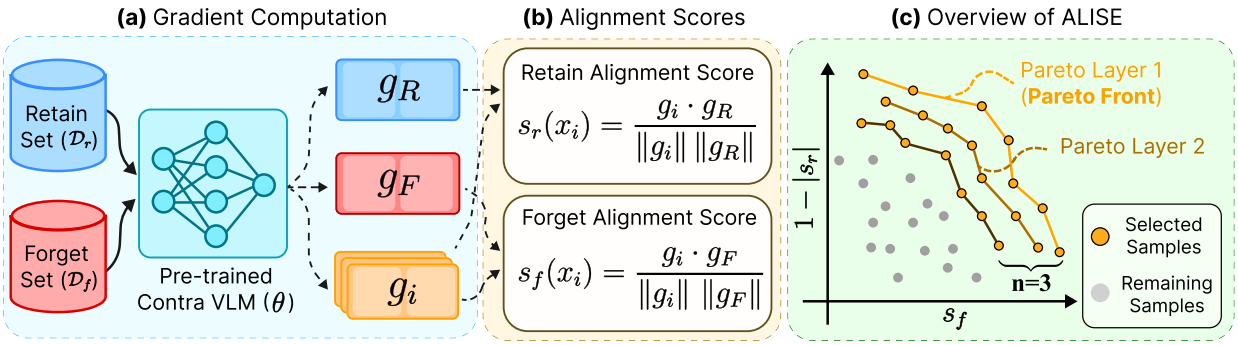

Figure 2. Overview of alignment score computation and ALISE. The procedure follows (a), (b), and (c). Here, n is a hyperparameter of ALISE, and the value shown in the figure is for illustration.

This formulation follows the interpretation of the alignment scores in Section 2.1. Samples with $s_r(x)$ close to 0 are weakly entangled with the retain set, whereas samples with $s_f(x)$ close to 1 facilitate forgetting of the target knowledge. Therefore, desirable forget samples should exhibit low $|s_r(x)|$ and high $s_f(x)$.

To jointly optimize these two objectives, we select samples using the Pareto front (Marler and Arora, 2010). The Pareto front consists of samples such that $1 - |s_r(x)|$ cannot be increased without decreasing $s_f(x)$, and vice versa, as illustrated in Figure 2c. By selecting samples from this front, we obtain a forget set that achieves a trade-off between the two objectives.

In addition, ALISE uses Non-Dominated Sorting (NDS) (Deb et al., 2002) to select a sufficient number of forget samples. Selecting only samples on the Pareto front may not provide enough samples in practice. NDS addresses this issue by iteratively extracting Pareto fronts while removing previously selected samples, generating multiple Pareto layers (*i.e.*, nondomination levels). ALISE then constructs a new forget set by selecting samples from the top n layers, where n is a hyperparameter controlling the number of selected Pareto layers.

3. Experiments

In this section, we validate ALISE on four downstream applications of contrastive VLMs: (1) Open-Vocabulary Classification (OVC), (2) Image-Text Retrieval, (3) Multimodal Large Language Models (MLLMs), and (4) Text-to-Image (T2I) Generation. Due to space limitations, the results for T2I Generation are provided in Appendix E. In addition, all ablation studies on ALISE are presented in Appendix F. For all experiments, we report only the performance on the test split of the forget set while constraining the performance drop on the test split of the retain set to within 5%. Detailed experimental settings are provided in Appendix D.

3.1. Main Results

Unlearning for OVC. In this subsection, we compare the unlearning performance of different data selection methods on the Open-Vocabulary Classification (OVC) task. All experiments are conducted on CLIP and EVA-CLIP using five unlearning methods. Results are reported in Table 1.

ALISE consistently achieves the lowest $\text{Acc}_{\mathcal{F}}$ across all settings compared to existing data selection methods. For example, when using samples selected by ALISE, the $\text{Acc}_{\mathcal{F}}$ of GAFT decreases from 15.0 to 1.0, which is the largest reduction among all data selection methods. This trend holds on both IU dataset and StanfordCars. In contrast, random sampling often yields higher $\text{Acc}_{\mathcal{F}}$ than using the full forget set, indicating that simply reducing the number of forget samples does not improve unlearning performance. RUM and UPCORE generally achieve lower $\text{Acc}_{\mathcal{F}}$ than using the full forget set, but remain higher than ALISE. These results show that ALISE more effectively removes target knowledge than other selection methods while maintaining utility across different contrastive VLMs, unlearning baselines, and datasets.

Unlearning for Image-Text Retrieval. We compare the unlearning performance of different data selection methods on the image-text retrieval task. All experiments use CLIPErase to unlearn CLIP. The results are presented in Table 2a.

The results show that ALISE consistently achieves superior unlearning performance compared to other data selection methods in the retrieval task. Specifically, ALISE yields lower R@5 and R@10 scores for both I2T and T2I retrieval than using the full forget set. Other methods exhibit R@5 values comparable to or slightly lower than those of the full forget set (*e.g.*, RUM: 51.4, UPCORE: 47.8 in I2T), but ALISE achieves a lower value (46.0 in I2T). In addition, other methods produce higher R@10 scores than when using the full forget set. These results suggest that ALISE is more effective at selectively removing target knowledge than existing methods in the retrieval task.

Table 1. Unlearning results on the IU dataset and StanfordCars for CLIP and EVA-CLIP. $\text{Acc}_{\mathcal{F}}$ is reported (lower is better), with the $\text{Acc}_{\mathcal{R}}$ drop on the Neighbor Set and ImageNet limited to 5%.

Unlearning Method	Data Selection	IU Dataset $\text{Acc}_{\mathcal{F}}$ (↓)		StanfordCars $\text{Acc}_{\mathcal{F}}$ (↓)	
		CLIP	EVA-CLIP	CLIP	EVA-CLIP
Original Model		86.6 ± 9.5	68.2 ± 15.3	85.4 ± 2.6	82.7 ± 9.6
GAFT	Baseline	15.0 ± 13.6	48.5 ± 16.7	6.9 ± 2.6	5.5 ± 9.5
	w/ Random	31.2 ± 38.2	48.8 ± 16.1	8.6 ± 6.4	5.5 ± 9.5
	w/ RUM	6.7 ± 5.8	47.5 ± 15.1	8.6 ± 4.2	4.9 ± 8.4
	w/ UPCORE	7.5 ± 5.8	47.2 ± 15.2	6.9 ± 5.6	6.1 ± 10.6
	w/ ALISE	**1.0** ± 1.7	**34.2** ± 11.9	**4.5** ± 3.4	**3.7** ± 6.3
SaLUN	Baseline	29.1 ± 34.6	51.5 ± 18.9	12.2 ± 2.7	25.2 ± 27.0
	w/ Random	39.0 ± 37.6	56.1 ± 21.1	11.3 ± 4.0	25.7 ± 26.4
	w/ RUM	18.6 ± 15.6	48.8 ± 16.1	10.7 ± 6.3	25.1 ± 26.4
	w/ UPCORE	22.6 ± 25.7	49.5 ± 16.1	12.2 ± 6.5	25.8 ± 27.2
	w/ ALISE	**15.4** ± 13.7	**38.7** ± 19.3	**8.3** ± 2.2	**24.0** ± 27.4
SLUG	Baseline	46.1 ± 44.0	45.7 ± 13.6	3.6 ± 5.0	19.6 ± 18.5
	w/ Random	25.0 ± 43.3	58.7 ± 22.4	2.5 ± 3.0	12.3 ± 12.4
	w/ RUM	25.0 ± 43.3	41.4 ± 11.9	3.7 ± 1.0	8.5 ± 14.8
	w/ UPCORE	45.7 ± 44.4	49.6 ± 16.9	3.6 ± 5.0	11.1 ± 16.1
	w/ ALISE	**20.8** ± 36.1	**31.3** ± 18.0	**0.0** ± 0.0	**8.1** ± 10.9
DualOptim	Baseline	36.1 ± 37.8	35.7 ± 20.1	17.3 ± 5.1	13.7 ± 13.7
	w/ Random	43.4 ± 36.7	36.3 ± 20.4	16.6 ± 7.9	13.0 ± 13.1
	w/ RUM	33.6 ± 37.0	33.7 ± 14.9	15.1 ± 6.3	25.1 ± 26.9
	w/ UPCORE	36.7 ± 38.6	34.0 ± 18.2	17.3 ± 4.4	10.9 ± 11.5
	w/ ALISE	**23.7** ± 16.6	**32.7** ± 19.9	**11.1** ± 7.1	**10.3** ± 10.6
CLIPErase	Baseline	33.6 ± 37.0	51.1 ± 27.9	21.7 ± 2.9	11.1 ± 17.6
	w/ Random	37.6 ± 37.0	40.8 ± 20.7	11.7 ± 11.6	11.8 ± 16.1
	w/ RUM	34.5 ± 35.1	42.4 ± 20.3	16.8 ± 9.0	12.4 ± 17.1
	w/ UPCORE	35.5 ± 35.2	49.8 ± 30.1	21.7 ± 2.9	11.1 ± 17.6
	w/ ALISE	**23.4** ± 22.6	**39.0** ± 18.8	**10.2** ± 1.6	**7.0** ± 7.4

Table 2. **(a)** Image-text retrieval unlearning: We report R@5 and R@10 on the forget set, with the R@5 drop on the Neighbor Set limited to 5%. **(b)** MLLM unlearning: We report $\text{Acc}_{\mathcal{F}}$ and MMBench performance, with the $\text{Acc}_{\mathcal{R}}$ drop on the Neighbor Set (validation split) limited to 5%.

(a)

Method	Image-to-Text (I2T)		Text-to-Image (T2I)	
	R@5 (↓)	R@10 (↓)	R@5 (↓)	R@10 (↓)
Original	63.5 ± 10.6	70.1 ± 7.0	72.5 ± 8.4	78.7 ± 5.5
Baseline	48.3 ± 7.9	56.6 ± 6.6	69.4 ± 9.0	75.9 ± 5.8
w/ Random	49.7 ± 7.0	57.9 ± 6.5	69.6 ± 9.0	76.0 ± 6.8
w/ RUM	51.4 ± 5.4	61.9 ± 6.8	69.1 ± 5.6	77.7 ± 4.6
w/ UPCORE	47.8 ± 5.8	56.9 ± 6.6	69.2 ± 9.1	76.1 ± 5.6
w/ ALISE	**46.0** ± 5.8	**56.2** ± 6.0	**68.2** ± 10.6	**75.4** ± 6.8

(b)

Method	$\text{Acc}_{\mathcal{F}}$ (↓)	MMBench (↑)
Original	93.3 ± 7.7	62.1 ± 0.0
Baseline	46.8 ± 37.3	59.4 ± 3.5
w/ Random	58.9 ± 35.1	58.1 ± 5.1
w/ RUM	52.9 ± 40.4	60.0 ± 3.5
w/ UPCORE	71.1 ± 29.0	59.6 ± 3.8
w/ ALISE	**33.1** ± 30.8	59.3 ± 5.1

Unlearning for MLLM. We compare ALISE with other data selection methods on MLLMs (Liu et al., 2023). All experiments use SLUG to unlearn CLIP on the IU dataset. The results are presented in Table 2b.

ALISE improves unlearning performance in MLLMs compared to using the full forget set. Specifically, ALISE reduces $\text{Acc}_{\mathcal{F}}$ from 46.8 to 33.1 while preserving performance on MMBench. In contrast, Random, RUM, and UPCORE yield higher $\text{Acc}_{\mathcal{F}}$ than the full forget set. These results suggest that ALISE facilitates removing only the target knowledge and prevents its generation, whereas other methods fail to do so.

The qualitative results for the responses generated by MLLMs and the images produced in T2I generation are provided in Appendix J.

4. Conclusion

In this paper, we show that the unlearning performance of contrastive VLMs depends not only on parameter update strategies but also on the composition of the forget set. Based on this observation, we propose ALISE, an alignment-aware data selection method for unlearning. The method quantifies each sample's alignment with both the retain set and the full forget set, and selects samples based on these two scores. Experiments demonstrate that ALISE more effectively removes target knowledge while preserving model utility than using the full forget set or existing selection methods. These findings indicate that explicitly considering sample-level alignment with both the retain and forget sets is crucial for unlearning in contrastive VLMs.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024.

Ibtihel Amara, Ahmed Imtiaz Humayun, Ivana Kajic, Zarana Parekh, Natalie Harris, Sarah Young, Chirag Nagpal, Najoung Kim, Junfeng He, Cristina Nader Vasconcelos, et al. Erasing more than intended? how concept erasure degrades the generation of non-target concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Zikui Cai, Yaoteng Tan, and M. Salman Asif. Targeted unlearning with single layer unlearning gradient. In *International Conference on Machine Learning (ICML)*, 2025.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2): 182–197, 2002.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *International Conference on Learning Representations (ICLR)*, 2024.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *International Conference on Learning Representations (ICLR)*, 2022.

Hyo Seo Kim, Dongyoon Han, and Junsuk Choe. Negmerge: Sign-consensual weight merging for machine unlearning. In *International Conference on Machine Learning (ICML)*, 2025.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, Dec 2013. doi: 10.1109/ICCVW.2013.77.

Alexey Kravets and Vinay Namboodiri. Zero-shot class unlearning in clip with synthetic samples. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 6456–6464, 2025.

Zhe Li, Wei Zhao, Yige Li, and Jun Sun. Where did it go wrong? attributing undesirable llm behaviors via representation gradient tracing. In *International Conference on Learning Representations (ICLR)*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, 2024.

R Timothy Marler and Jasbir S Arora. The weighted sum method for multi-objective optimization: new insights. *Structural and multidisciplinary optimization*, 41(6):853–862, 2010.

Saemi Moon, Minjong Lee, Sangdon Park, and Dongwoo Kim. Holistic unlearning benchmark: A multi-faceted evaluation for text-to-image diffusion model unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Soumyadeep Pal, Changsheng Wang, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. Llm unlearning reveals a stronger-than-expected coreset effect in current benchmarks. In *Conference on Language Modeling (COLM)*, 2025.

Vaidehi Patil, Elias Stengel-Eskin, and Mohit Bansal. Upcore: Utility-preserving coreset selection for balanced unlearning. *arXiv preprint arXiv:2502.15082*, 2025.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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 *International Conference on Machine Learning (ICML)*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

David Thiel. Identifying and eliminating csam in generative ml training data and models. *Stanford Internet Observatory, Cyber Policy Center, December*, 23(3):131, 2023.

Rishabh Tiwari, Krishnateja Killamsetty, Rishabh Iyer, and Pradeep Shenoy. Gcr: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jing Wu and Mehrtash Harandi. Munba: Machine unlearning via nash bargaining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4):1–39, 2023.

Tianyu Yang, Lisen Dai, Xiangqi Wang, Minhao Cheng, Yapeng Tian, and Xiangliang Zhang. Cliperase: Efficient unlearning of visual-textual associations in clip. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30438–30452, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:5824–5836, 2020.

Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Rao Kompella, Xiaoming Liu, et al. Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models. *arXiv preprint arXiv:2402.11846*, 2024.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xuyang Zhong, Haochen Luo, and Chen Liu. Dualoptim: Enhancing efficacy and stability in machine unlearning with dual optimizers. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

A. Related Works

A.1. Unlearning in Contrastive VLMs

Machine unlearning (MU) aims to remove target knowledge from a model while preserving its utility. This is particularly important for contrastive vision-language models (VLMs), such as CLIP (Radford et al., 2021) and EVA-CLIP (Sun et al., 2023), because they are trained on large-scale web data that may contain copyrighted or inappropriate content (Schuhmann et al., 2022; Thiel, 2023).

Recent studies on unlearning for contrastive VLMs have mainly focused on how to update model parameters to selectively remove target knowledge. For example, some methods (Fan et al., 2024; Cai et al., 2025) identify and update only the parameters that are strongly associated with the target knowledge. Others add regularization terms to the loss to prevent abrupt parameter changes during unlearning, thereby preserving model utility (Kravets and Namboodiri, 2025; Yang et al., 2025). Another line of work removes target knowledge by subtracting parameter changes induced by fine-tuning on the forget set from the original model weights (Ilharco et al., 2022; Kim et al., 2025).

In this paper, we show that unlearning performance in contrastive VLMs strongly depends on the composition of the forget set, particularly on the alignment of each sample with the retain set and the full forget set. However, existing parameter-update-based methods do not consider this factor. To address this gap, we propose a data selection framework based on sample-level alignment.

A.2. Data Selection for Unlearning

Data selection refers to the process of selecting samples for training or evaluating a machine learning model from a pool of candidate data points (Albalak et al., 2024). Its goals include improving data efficiency (Paul et al., 2021), mitigating bias (Longpre et al., 2024), and supporting specific learning objectives (Tiwari et al., 2022).

Several studies have explored data selection for machine unlearning in image classification and LLMs. Zhao et al. (2024) partitions the forget set into multiple groups based on unlearning difficulty and performs unlearning sequentially across groups for image classification models. In LLM unlearning, Pal et al. (2025) shows that using only a small fraction of the forget set (*e.g.*, 5%) can achieve performance comparable to using the full forget set, while Patil et al. (2025) proposes removing outlier samples that may excessively harm model utility.

Existing studies mainly consider the unlearning difficulty or efficiency of individual forget samples. Although they analyze the relationship between forget samples and the retain set, they do not explicitly incorporate it into the selection process. In contrast, we observe that the alignment of each forget sample with the retain set and the full forget set is important in unlearning. We therefore quantify these alignments and propose a data selection framework based on them.

B. Preliminaries

Suppose θ denotes the parameters of a pre-trained contrastive VLM, and let $\mathcal{D}_r$ and $\mathcal{D}_f$ denote the retain set and forget set, respectively. Unlearning for contrastive VLMs typically optimizes the following objective (Cai et al., 2025; Yang et al., 2025):

$$\min_{\theta} \ \mathcal{L}_r(\theta; \mathcal{D}_r) + \alpha \mathcal{L}_f(\theta; \mathcal{D}_f), \tag{7}$$

where $\mathcal{L}_r$ and $\mathcal{L}_f$ denote the retain and forget losses, respectively, and α is a hyperparameter that controls the magnitude and sign of the forget loss. The retain loss $\mathcal{L}_r$ is the InfoNCE-based (Oord et al., 2018) contrastive loss in CLIP (Radford et al., 2021), defined as

$$\mathcal{L}_r(\theta; \mathcal{D}_r) = \mathbb{E}_{\mathcal{B} \sim \mathcal{D}_r} \left[\frac{1}{|\mathcal{B}|} \sum_{x_i \in \mathcal{B}} \ell_r(x_i; \theta, \mathcal{B}) \right], \tag{8}$$

where $\ell_r(x_i; \theta, \mathcal{B})$ denotes the loss of sample $x_i = (x_i^{\text{img}}, x_i^{\text{text}})$ in mini-batch $\mathcal{B}$. The forget loss $\mathcal{L}_f$ is determined by the unlearning method, *e.g.*, cosine embedding loss (Cai et al., 2025) or cross-entropy loss (Wu and Harandi, 2025; Ilharco et al., 2022). It is defined as

$$\mathcal{L}_f(\theta; \mathcal{D}_f) = \mathbb{E}_{x \sim \mathcal{D}_f} \left[\ell_f(x; \theta) \right], \tag{9}$$

where $x = (x^{\text{img}}, x^{\text{text}})$ is a forget sample and $\ell_f(x; \theta)$ denotes the loss of x.

In the next section, we define gradients computed from $\ell_f(x_i;\theta)$ for a single forget sample, and from $\mathcal{L}_r(\theta;\mathcal{D}_r)$ and $\mathcal{L}_f(\theta;\mathcal{D}_f)$. Based on these, we define scores that quantify the alignment of each sample with the retain set and the full forget set, and use them to select samples for a new forget set.

C. Discussion on the Alignment Scores

To measure alignment scores, we use the representation gradient $(\partial\mathcal{L}/\partial z)$ instead of the parameter gradient $(\partial\mathcal{L}/\partial\theta)$. This choice is motivated by recent work (Li et al., 2026), which shows that representation gradients are more computationally efficient due to their lower dimensionality and enable more precise data attribution.

In addition, we concatenate gradients from both modalities to compute a unified alignment score. Without concatenation, the alignment between a sample and the retain set can be quantified for each modality using the cosine similarity between gradients as follows:

$$s_r^m(x_i) = \frac{g_i^m \cdot g_R^m}{\|g_i^m\|\,\|g_R^m\|}, \qquad s_f^m(x_i) = \frac{g_i^m \cdot g_F^m}{\|g_i^m\|\,\|g_F^m\|}, \quad m \in \{\text{img}, \text{text}\}. \tag{10}$$

However, modality-wise scores reflect alignment only within each modality and fail to jointly capture alignment across both modalities in contrastive VLMs. Therefore, we concatenate the gradients and compute their cosine similarity to define the final retain alignment score.

Ablation studies comparing parameter and representation gradients, as well as different score designs, are provided in Appendix F.

D. Further Details on the Unlearning Configurations

D.1. Tasks

Open-vocabulary classification (OVC). Unlike conventional image classification, OVC is a text-grounded task in which free-form text, including text not seen during training, can be used as candidate class descriptions. Specifically, the input text is converted into text embeddings by the text encoder, and the model predicts the class whose text embedding is most similar to the image embedding. In this task, unlearning aims to remove knowledge associated with the target text so that the model can no longer correctly recognize it.

Image-text retrieval. Image-text retrieval is a task that retrieves the most relevant text given an input image (image-to-text, I2T), or the most relevant image given a text query (text-to-image, T2I). In this task, the goal of unlearning is to remove knowledge associated with a specific target so that the model can no longer retrieve images or text related to that target.

Multimodal large language models (MLLMs). MLLMs take both free-form text and images as input and generate textual responses. In this task, unlearning aims to suppress the model from generating information related to the deletion target. For example, when an image containing a specific identity is given, the model should no longer produce specific descriptions or related information about that identity.

Text-to-image (T2I) generation. T2I generation is also a text-grounded task that takes free-form text as input and generates a corresponding image. In this task, we update only the text encoder of the contrastive VLM. The goal of unlearning is to remove knowledge related to the target text so that the model can no longer generate images corresponding to it.

D.2. Models

Open-vocabulary classification. We use a ViT-B/16 architecture and unlearn both the vision encoder and the text encoder of CLIP (Radford et al., 2021) and EVA-CLIP (Sun et al., 2023). For some GAFT-based unlearning methods, we follow prior work (Ilharco et al., 2022; Ortiz-Jimenez et al., 2023) and unlearn only the vision encoder.

Image-text retrieval. We use CLIP with a ViT-B/16 architecture. Following CLIPErase (Yang et al., 2025), we unlearn both the vision and text encoders.

Text-to-image generation. We use Stable Diffusion v1.5 (Rombach et al., 2022), which employs the text encoder of a ViT-H/14 CLIP. After unlearning the CLIP text encoder, we replace the original text backbone of the T2I model with the unlearned encoder to construct the unlearned T2I model.

Multimodal large language model. We use LLaVA-v1.5-7B (Liu et al., 2023), which adopts a ViT-L/14 CLIP vision encoder as its visual backbone. After unlearning the CLIP vision encoder, we replace the original vision backbone of the MLLM with the unlearned encoder to construct the unlearned MLLM.

For OVC, we use publicly available weights from the `open_clip` repository (Ilharco et al., 2021). For T2I and MLLM, we use publicly released pretrained weights from Hugging Face.

D.3. Datasets

For all tasks, we use a subset of LAION-400M as the retain set, following prior work (Cai et al., 2025). Specifically, the retain set consists of approximately 7,900 image-text pairs, constructed by randomly selecting a single shard from LAION-400M (Schuhmann et al., 2022) and filtering out expired URLs.

Open-vocabulary classification. We evaluate unlearning performance on two benchmarks: Identity Unlearning (IU) dataset (Cai et al., 2025), which assesses how effectively the model removes identity-related information that may raise privacy concerns, and StanfordCars (Krause et al., 2013), which evaluates whether unlearning remains effective in a fine-grained setting where the classes in the forget set are highly similar to those in the retain set. For each benchmark, the forget set consists of samples corresponding to four target identities. In the IU dataset, the forget set for unlearning is extracted from LAION-400M, while the evaluation set is extracted from CelebA (Liu et al., 2015). In StanfordCars, we use the train split as the forget set for unlearning and the test split for evaluation. To evaluate whether the model preserves performance on semantically similar concepts, we additionally use a **Neighbor Set (Neigh)** (Patil et al., 2025). Prior work has shown that unlearning a target can degrade performance on semantically related samples (Amara et al., 2025). This suggests that unlearning should be evaluated not only by the removal of the targets but also by the preservation of semantically related concepts. Accordingly, for each OVC benchmark, we construct a Neighbor Set composed of samples corresponding to concepts semantically similar to the targets. For the IU dataset, we use samples from CelebA corresponding to 100 celebrities excluding the target identities. For Stanford Cars, we use test samples from all classes except the target classes.

We also use the ImageNet (Deng et al., 2009) test set to evaluate model utility on a broader domain.

Image-text retrieval. We use the Flickr30k (Plummer et al., 2015) dataset for both unlearning and evaluation in the image-text retrieval task. Flickr30k is a widely used benchmark consisting of approximately 30,000 images, each paired with five human-annotated captions describing the scene. It is commonly used to evaluate image-to-text and text-to-image retrieval performance. The forget set is constructed by selecting samples from Flickr30k whose text contains four targets. We further split the forget set into 70% for unlearning and 30% for evaluation. As in the OVC task, we define and use a Neighbor Set, which is constructed by randomly sampling 30% of the remaining dataset excluding the forget set.

Text-to-image generation. We evaluate unlearning performance on the UnlearnCanvas dataset (Zhang et al., 2024). This benchmark is designed to quantitatively evaluate whether the model can remove the ability to generate images of copyright-related artistic styles or object concepts while preserving the model's generation ability on non-target concepts. Specifically, we evaluate 50 artistic styles and 20 object concepts as targets. Each style contains 400 images, and each object contains 1,200 images.

Multimodal large language models. We use the IU dataset proposed in (Cai et al., 2025). This benchmark evaluates how effectively unlearning suppresses the generation of information about specific identities. The forget set consists of 10 identities in total. As in the IU dataset for the OVC task, the forget set used for unlearning is extracted from LAION-400M, while the forget set used for evaluation is extracted from CelebA. To evaluate whether model utility is preserved after unlearning, we use both a Neighbor Set and MMBench (Liu et al., 2024). The Neighbor Set is an evaluation dataset composed of samples corresponding to identities in the forget set that are not forgetting targets. However, since we impose a constraint on the drop in performance on the retain set ($\text{Acc}_{\mathcal{R}}$), evaluating the entire Neighbor Set at every unlearning iteration in MLLMs is computationally expensive. Therefore, following (Cai et al., 2025), we use a **Neighbor Validation Set** instead of the full Neighbor Set, which consists of 100 celebrity images that do not contain any forgetting targets.

D.4. Evaluation Metrics

Open-vocabulary classification. For the OVC task, we use the following evaluation metrics to assess unlearning performance: 1) Forget Accuracy ($\text{Acc}_{\mathcal{F}}$) denotes the classification accuracy on the targets to forget. A lower $\text{Acc}_{\mathcal{F}}$ indicates better unlearning of the target. 2) Retain Accuracy ($\text{Acc}_{\mathcal{R}}$) denotes the classification accuracy on datasets that evaluate

model utility, such as the Neighbor Set and ImageNet. We report $\mathrm{Acc}_{\mathcal{F}}$ under the constraint that the drop in $\mathrm{Acc}_{\mathcal{R}}$ on both the Neighbor Set and ImageNet does not exceed 5%. We independently perform unlearning for each of the four targets to forget and report the mean and standard deviation across them.

Image-text retrieval. To evaluate the image-text retrieval task, we use the Recall@K (R@K) metric for both image-to-text (I2T) and text-to-image (T2I) retrieval. Recall@K measures whether the ground-truth item is included within the top-K retrieved results. It can be expressed in terms of true positives (TP) and false negatives (FN) as $\mathrm{Recall@K} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$, where a true positive indicates that the correct match appears within the top-K results, and a false negative indicates that it does not. We also evaluate performance on the Neighbor Set using the same Recall@K metric. We report the retrieval performance on the forget set under the constraint that the drop in R@5 for both I2T and T2I on the Neighbor Set does not exceed 5%.

Text-to-image generation. We use the four standard evaluation metrics provided by UnlearnCanvas. To assess whether the generated images correspond to the input text, we use the style and object classifiers provided by UnlearnCanvas. Specifically, UA (Unlearn Accuracy) measures the classification accuracy of generated images when the forgetting target is given as input text, and the final value is computed as $1 -$ accuracy. Thus, higher UA indicates better removal of the target knowledge. IRA (In-domain Retain Accuracy) measures the classification accuracy of generated images that belong to the same domain (style or object) as the forgetting target but correspond to different concepts, and is used to evaluate model utility. CRA (Cross-domain Retain Accuracy) measures the classification accuracy of generated images corresponding to concepts from different domains than the forgetting target, also evaluating model utility. FID (Fréchet Inception Distance) measures the distance between generated and real image distributions in the embedding space of the Inception network, where lower FID indicates better generation quality.

Multimodal large language models. Following Cai et al. (2025), we use the following metrics: (1) Forget Accuracy ($\mathrm{Acc}_{\mathcal{F}}$) measures whether the generated text contains the target identity name. A lower $\mathrm{Acc}_{\mathcal{F}}$ indicates that the model is less likely to generate the target. (2) Retain Accuracy ($\mathrm{Acc}_{\mathcal{R}}$) measures whether the model correctly generates non-target identity names, and is evaluated on both the Neighbor Set and the Neighbor Validation Set.

In MLLM experiments, we perform unlearning under the constraint that the drop in $\mathrm{Acc}_{\mathcal{R}}$ on the Neighbor Validation Set does not exceed 5%. For models satisfying this constraint, we report $\mathrm{Acc}_{\mathcal{F}}$, $\mathrm{Acc}_{\mathcal{R}}$ on the Neighbor Set, and MMBench performance. The $\mathrm{Acc}_{\mathcal{R}}$ on the Neighbor Set is omitted from the main paper due to page limits and is provided in Appendix G.2.

We independently perform unlearning for each of the 10 targets and report the mean and standard deviation. During evaluation, we prompt the model with *"What is the name of the person in the image?"*.

D.5. Implementation Details

ALISE computes gradients for samples in both the full forget set and the retain set and uses them for data selection. Since all unlearning tasks in this paper use the same retain set, we precompute and store the gradients of the retain set and reuse them across all experiments. An ablation study on the batch composition of the retain set is provided in Section G.3. We set the batch size to 64 for both the forget and retain sets. Experiments using CLIPErase and those with larger batch sizes in Section G.3 are conducted on NVIDIA A6000. Experiments on MLLMs and T2I generation are conducted on NVIDIA RTX 4090, while all remaining experiments are conducted on NVIDIA A5000 GPUs.

D.6. Compared Methods

Unlearning Baselines. To evaluate ALISE, we consider five unlearning baselines: GAFT, SaLUN (Fan et al., 2024), SLUG (Cai et al., 2025), DualOptim (Zhong et al., 2025), and CLIPErase (Yang et al., 2025). GAFT (Gradient Ascent Fine-Tuning) removes target knowledge by applying gradient ascent on the forget set, while performing standard fine-tuning on the retain set to preserve model utility. SaLUN estimates the importance of each parameter based on the Fisher Information Matrix and selectively updates a subset of parameters through saliency masking. SLUG performs class-wise unlearning on a single layer rather than the entire model. Specifically, it selects one layer that best satisfies both the forget and retain objectives and applies unlearning only to that layer. DualOptim optimizes the forget loss and the retain loss with separate optimizers, thereby explicitly disentangling the two objectives during optimization. CLIPErase defines the forget loss as the cosine similarity between image and text embeddings, encouraging orthogonality between the two embeddings. In addition, it introduces a consistency loss to preserve the model outputs for non-target information.

Data Selection Methods. To evaluate ALISE, we compare it with three data selection strategies: Random, RUM (Zhao

Table A1. Unlearning results on UnlearnCanvas for Stable Diffusion v1.5, evaluated on both style and object unlearning scenarios.

Method	Style Unlearning				Object Unlearning				FID (↓)
	UA (↑)	IRA (↑)	CRA (↑)	Avg (↑)	UA (↑)	IRA (↑)	CRA (↑)	Avg (↑)	
Baseline	88.2	87.1	84.2	86.5	75.0	82.0	80.6	79.2	124.0
w/ Random	76.9	71.0	81.0	76.3	66.0	76.0	68.9	70.3	114.2
w/ RUM	88.6	85.5	86.1	86.7	59.0	70.0	77.4	68.8	146.1
w/ UPCORE	90.2	82.7	84.1	85.7	70.0	83.0	80.6	77.9	121.0
w/ ALISE	89.8	91.4	92.5	**91.2**	76.0	81.0	81.0	**79.3**	123.7

et al., 2024), and UPCORE (Patil et al., 2025). Random uniformly samples from the full forget set to construct a subset with the same size as ALISE, serving as a reference baseline. RUM is a meta-learning-based unlearning method that partitions the forget set into multiple subsets according to unlearning difficulty, identifies the most effective unlearning algorithm for each subset, and performs unlearning sequentially. A variant that applies a single unlearning algorithm across all subsets is denoted as RUM$^{\mathcal{F}}$. Since each experiment uses a single unlearning algorithm, we adopt RUM$^{\mathcal{F}}$ for comparison. UPCORE is a coreset-based method that removes outlier forget samples that may degrade utility and selects a core subset. By leveraging the dispersion of hidden representations, it aims to remove target knowledge while minimizing performance degradation on the retain set.

Hyperparameter Search. We split the forget set used for unlearning into 90% training and 10% validation for OVC and image-text retrieval tasks. Then, we use the validation set for hyperparameter search. For MLLMs and T2I generation, we follow the original validation settings of each benchmark.

For unlearning methods, we use the hyperparameters reported in the original papers and codebases, and perform random search only over the learning rate and weight decay for 10 trials using the full forget set. The selected hyperparameters are then fixed and reused across all data selection methods.

For data selection methods, we perform grid search. Since prior work does not report explicit hyperparameter ranges, we define the search space empirically. For RUM, the order of subsets used for sequential unlearning is treated as a hyperparameter; with three subsets, we evaluate all six possible orderings. For UPCORE, we vary the outlier ratio p from 5 to 50 in increments of 5. For ALISE, the Pareto layer n is the key hyperparameter. We search $n \in [1, 10]$ for OVC and image-text retrieval tasks, and $n \in \{1, 2, 4, 8, 16\}$ for the remaining tasks due to the larger size of the forget set.

E. Unlearning for T2I Generation

We compare the unlearning performance of ALISE with other data selection methods on T2I generation, a downstream task of contrastive VLMs. All experiments are conducted under a setting where CLIP is unlearned using SLUG. The results are presented in Table A1.

The results show that ALISE achieves the best overall performance across all evaluation metrics. In particular, in the style unlearning scenario, ALISE consistently improves UA($+1.6$), IRA($+4.3$), and CRA($+8.3$) over using the full forget set, whereas other selection methods perform worse. These results suggest that, compared to other methods, using the forget subset selected by ALISE for unlearning more effectively suppresses image generation related to the target knowledge while preserving model utility.

F. Ablation Study

In this section, we conduct ablation studies on the alignment scores s_r and s_f used in ALISE, and analyze sensitivity to the hyperparameter n. All experiments are performed on the OVC task using the IU dataset, under a setting where CLIP is unlearned with GAFT. In all ablation studies, each forget subset contains the same number of samples as the subset selected by ALISE.

Ablation on the two scores of ALISE. We compare ALISE with variants that use only one of the two scores, s_r or s_f, to assess the benefit of their joint use. Samples are ranked in descending order by $1 - |s_r(x_i)|$ when using s_r alone, and by $s_f(x_i)$ when using s_f alone. The top samples are selected with the same size as in ALISE. Results are reported in Table A2a.

Table A2. Four ablations on ALISE. All results report $\text{Acc}_{\mathcal{F}}$, with $\text{Acc}_{\mathcal{R}}$ drop on the Neighbor Set $\leq 5\%$.

(a) Two scores of ALISE.

	s_f	s_r	$\text{Acc}_{\mathcal{F}}$ (↓)
CLIP	-	-	86.6
+ GAFT	-	-	15.0
w/ selected forget subset	✗	✓	24.0
	✓	✗	6.0
	✓	✓	**1.0**

(b) s_r-based selection criterion of ALISE.

	Selection criterion	$\text{Acc}_{\mathcal{F}}$ (↓)
CLIP	–	86.6
+ GAFT	–	15.0
w/ selected forget subset	$s_r \approx -1$	50.0
	$s_r \approx 1$	30.7
	$s_r \approx -1,\ s_f \approx 1$	9.4
	$s_r \approx 1,\ s_f \approx 1$	4.8
	$\lvert s_r \rvert \approx 0,\ s_f \approx 1$	**1.0**

(c) Score for alignment using both modalities.

	Score Design	$\text{Acc}_{\mathcal{F}}$ (↓)
CLIP	–	86.6
+ GAFT	–	15.0
w/ selected forget subset	Average	3.0
	Multiplication	3.0
	Inner product	5.3
	Concat grads (Ours)	**1.0**

(d) Parameter vs. Representation gradients

	Gradient Type	Time (s)	$\text{Acc}_{\mathcal{F}}$ (↓)
CLIP	–	–	86.6
+ GAFT	–	–	15.0
w/ selected forget subset	Param (All Layers)	1128.9	30.7
	Param (Last Layer)	73.1	2.0
	Representation	**8.2**	**1.0**

The results show that jointly using s_r and s_f achieves the best unlearning performance compared with using either score alone. Specifically, using both scores together yields a $\text{Acc}_{\mathcal{F}}$ that is 23.0 and 5.0 lower than using s_r and s_f alone, respectively. These results indicate that jointly considering both scores is more effective for removing target knowledge while preserving model utility.

Ablation on the s_r-based selection criterion. We compare ALISE with alternatives that select samples whose $s_r(x_i)$ values are close to -1 or 1. The results are reported in Table A2b.

The results show that the selection criterion used in ALISE —selecting samples with $|s_r(x_i)|$ close to 0—yields the best unlearning performance. In contrast, constructing the forget set from samples with $s_r(x_i)$ close to -1 or 1 results in even higher $\text{Acc}_{\mathcal{F}}$ than using the full forget set. Moreover, even when combined with s_f, samples with $s_r(x_i)$ close to -1 or 1 still yield higher $\text{Acc}_{\mathcal{F}}$ than selecting samples with $|s_r(x_i)|$ close to 0. These results suggest that selecting samples that are weakly entangled with the retain set, as in ALISE, facilitates removing the target knowledge while preserving model utility.

Score design for alignment using both modalities. We conduct ablation studies on score designs that jointly consider alignment measured from both modalities. ALISE reflects this by concatenating gradients from the two modalities and computing their cosine similarity. Alternative designs include using gradient inner products or aggregating modality-wise alignment scores through averaging or multiplication. Results are reported in Table A2c.

ALISE achieves the best unlearning performance among all score design choices. Specifically, ALISE yields a lower $\text{Acc}_{\mathcal{F}}$ of 1.0, compared to 5.3 for inner product and 3.0 for both addition and multiplication of modality-wise scores. These results indicate that computing cosine similarity on concatenated gradients is more effective for unlearning than alternative score designs.

Parameter vs. Representation gradients. We compare representation and parameter gradients in terms of data selection time and unlearning performance, where alignment scores are computed using each gradient type. As shown in Table A2d, representation gradients are both more efficient and more effective for unlearning. In contrast, parameter gradients incur substantially higher cost: using all layers is about $137\times$ slower, and even restricting to the last layer remains about $9\times$ slower. These results suggest that using representation gradients to compute alignment scores enables more efficient and effective sample selection for unlearning.

Ablation on the number of pareto layers. We analyze how the number of selected Pareto layers, denoted by n, affects unlearning performance. As shown in Figure A1a, $\text{Acc}_{\mathcal{F}}$ generally decreases as n increases in the range of $n \leq 6$ and reaches its lowest value at $n = 7$. In contrast, for $n \geq 8$, $\text{Acc}_{\mathcal{F}}$ increases as n becomes larger. This trend is explained by Figure A1b: as n increases beyond a certain point, the number of selected samples approaches that of the full forget set,

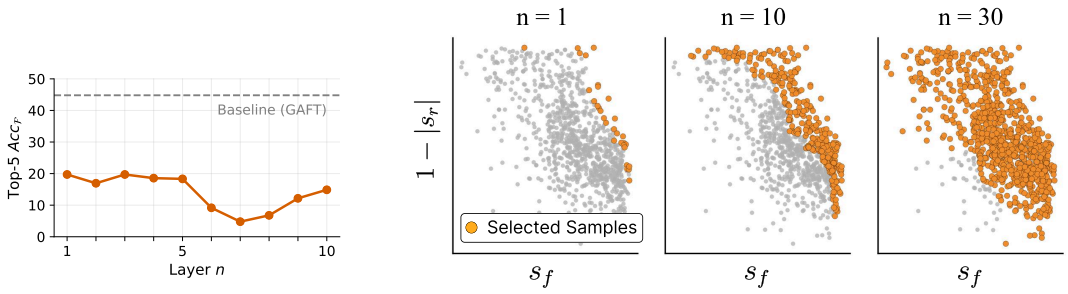

Figure A1. (a) Effect of the Pareto layer n on $\text{Acc}_{\mathcal{F}}$. Top-5 $\text{Acc}_{\mathcal{F}}$ is reported with the $\text{Acc}_{\mathcal{R}}$ drop on the Neighbor Set limited to 5%. (b) Visualization of selected samples across Pareto layers n.

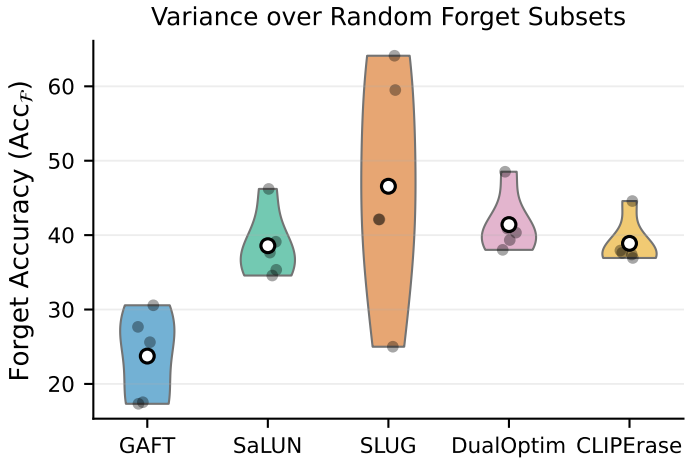

Figure A2. **Variance of $Acc_{\mathcal{F}}$ on random forget subsets.** The black dots indicate the performance of each unlearning method when using randomly sampled forget subsets, and the white dots represent the average performance. All experiments are conducted under the constraint that the decrease in retain accuracy ($Acc_{\mathcal{R}}$) is limited to within 5%.

diminishing the effect of data selection.

G. Additional Experiments

G.1. Performance Variance on Random Forget Subsets

In this section, we examine the performance variance across forget subsets constructed by randomly sampling from the full forget set for each unlearning method. Specifically, we use five different random seeds to sample 10% of the full forget set and construct five forget subsets. We then measure and compare the forget accuracy obtained after unlearning the contrastive VLM CLIP using each subset.

As shown in Figure A2, the performance of the unlearning methods varies substantially even when the forget set is constructed by simple random sampling. For example, in the case of SLUG, the forget accuracy ranges from 25.0% to 64.1% depending on the subset. This suggests that sample selection for constructing the forget set is an important factor in unlearning performance.

G.2. Unlearning for MLLM under Different Retain Constraints

We validate ALISE on MLLMs (Liu et al., 2023), where contrastive VLMs serve as core components. We use the Identity Unlearning (IU) dataset and compare data selection methods using SLUG as the unlearning baseline. Table A3 presents the results.

Table A3. Comparison of data selection methods under the MLLM unlearning setting. We report $\text{Acc}_{\mathcal{F}}$, $\text{Acc}_{\mathcal{R}}$ on the Neighbor Set, and MMBench performance, under constraints that limit the drop in $\text{Acc}_{\mathcal{R}}$ on the validation split of the Neighbor Set to 5%, 10%, or 20%.

Method	Retain Constraint = 5%			Retain Constraint = 10%			Retain Constraint = 20%		
	$\text{Acc}_{\mathcal{F}}$ (↓)	$\text{Acc}_{\mathcal{R}}$ (↑)	MMBench (↑)	$\text{Acc}_{\mathcal{F}}$ (↓)	$\text{Acc}_{\mathcal{R}}$ (↑)	MMBench (↑)	$\text{Acc}_{\mathcal{F}}$ (↓)	$\text{Acc}_{\mathcal{R}}$ (↑)	MMBench (↑)
CLIP	93.3 ± 7.7	93.3 ± 0.9	62.1 ± 0.0	93.3 ± 7.7	93.3 ± 0.9	62.1 ± 0.0	93.3 ± 7.7	93.3 ± 0.9	62.1 ± 0.0
Baseline	46.8 ± 37.3	75.9 ± 12.5	59.4 ± 3.5	46.8 ± 37.3	75.8 ± 12.5	59.4 ± 3.5	35.5 ± 38.8	73.9 ± 17.9	58.3 ± 5.6
w/ Random	58.9 ± 35.1	79.0 ± 11.5	58.1 ± 5.1	57.6 ± 35.4	78.7 ± 11.0	58.1 ± 4.8	46.1 ± 41.3	74.6 ± 13.5	57.6 ± 4.7
w/ RUM	52.9 ± 40.4	82.2 ± 13.4	60.0 ± 3.5	53.0 ± 34.3	81.8 ± 12.6	59.1 ± 3.9	46.4 ± 30.6	72.9 ± 15.2	58.9 ± 3.9
w/ UPCORE	71.1 ± 29.0	84.7 ± 9.6	59.6 ± 3.8	70.4 ± 30.1	78.3 ± 17.1	59.5 ± 3.2	37.0 ± 34.2	78.1 ± 12.3	57.4 ± 6.7
w/ ALISE	**33.1** ± 30.8	76.9 ± 14.8	59.3 ± 5.1	**31.7** ± 29.8	80.9 ± 13.0	59.2 ± 5.0	**21.1** ± 20.7	77.0 ± 10.3	60.7 ± 2.6

The results show that ALISE improves unlearning performance in MLLMs over using the full forget set. Specifically, with ALISE, $\text{Acc}_{\mathcal{F}}$ drops from 46.8 to 33.1 under the 5% retain constraint, from 46.8 to 31.7 under the 10% retain constraint, and from 35.5 to 21.1 under the 20% retain constraint. In particular, ALISE substantially lowers $\text{Acc}_{\mathcal{F}}$ while preserving performance on MMBench. This indicates that the forget subset selected by ALISE is more effective than the full forget set, as it removes target knowledge while better preserving model utility. In contrast, with forget sets selected by Random, RUM, and UPCORE, $\text{Acc}_{\mathcal{F}}$ increases compared with the baseline. Similar to the observations in T2I generation, this suggests that, in MLLMs, the samples selected by ALISE are more effective for unlearning than those selected by existing data selection methods.

G.3. Batch Composition and Size Robustness

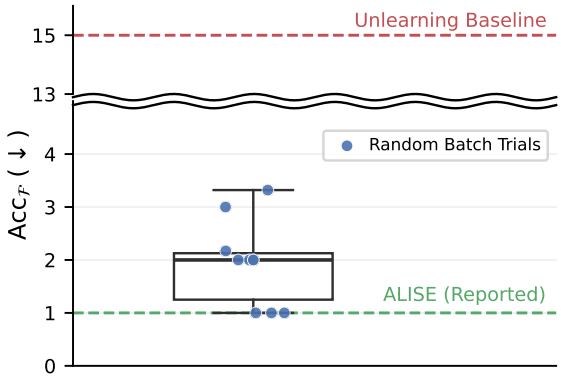

Figure A3. **Variance of performance under random batch compositions.** All experiments use GAFT to unlearn CLIP, with the decrease in retain accuracy ($Acc_{\mathcal{R}}$) constrained within 5%.

According to Eq. (8), the retain gradient g_R is derived from the batch-wise contrastive loss. Thus, both batch composition and batch size can affect the gradient, which may influence the performance of ALISE. We analyze its robustness to these factors.

We construct 10 random batch compositions and report the variance of $\text{Acc}_{\mathcal{F}}$ in Figure A3. ALISE remains robust to batch composition: the average $\text{Acc}_{\mathcal{F}}$ is 1.95, close to the reported ALISE performance (1.0) in Table 1. Even in the worst case (3.32), $\text{Acc}_{\mathcal{F}}$ is still significantly lower than using the full forget set (15.0).

We further vary the batch size (16, 32, 128) from the default 64 and report the results in Table A4. Larger batch sizes yield better unlearning performance. However, even with a small batch size (16), ALISE still achieves substantially lower $\text{Acc}_{\mathcal{F}}$ than using the full forget set.

These results indicate that ALISE is effective for unlearning regardless of batch composition and batch size.

Table A4. **Performance of ALISE under different batch sizes used for data selection.** All experiments use GAFT to unlearn CLIP, with $Acc_{\mathcal{R}}$ drop limited to 5%.

Method	Batch size	Acc$_{\mathcal{F}}$ (↓)
CLIP	-	86.6
+ GAFT	64	15.0
w/ ALISE	16	7.3
	32	3.0
	64	1.0
	128	0.0

Table A5. Comparison of Acc$_{\mathcal{F}}$ between modality-specific scores and ALISE using both modalities. All experiments are conducted on the IU dataset for the OVC task, with the drop in $Acc_{\mathcal{R}}$ limited to 5%.

	Identity Dataset Acc$_{\mathcal{F}}$ (↓)	
Method	CLIP	EVA-CLIP
Original Model	86.6 ± 9.5	68.2 ± 15.3
Baseline (GAFT)	15.0 ± 13.6	48.5 ± 16.7
$(s_r^{\text{img}}, s_f^{\text{img}})$	6.4 ± 5.9	41.8 ± 17.7
$(s_r^{\text{text}}, s_f^{\text{text}})$	2.0 ± 3.5	38.2 ± 19.6
(s_r, s_f) (ALISE)	**1.0** ± 1.7	**34.2** ± 11.9

G.4. Ablation Study on Modality-Specific Alignment Scores

We compare unlearning performance when alignment scores are computed using gradients from a single modality versus concatenated gradients from both modalities. As shown in Table A5, using both modalities consistently yields the lowest Acc$_{\mathcal{F}}$. While single-modality scores still outperform using the full forget set, they underperform compared to using both modalities. These results indicate that measuring alignment by jointly considering all modalities in contrastive VLMs is important for selecting samples that better support unlearning.

G.5. Unlearning Time Comparison

Table A6. **Unlearning Time Comparison.** We separately report data selection time and unlearning time, with all values measured in seconds. All experiments are conducted under the SLUG-based CLIP unlearning setting. The evaluation is performed on the OVC task, where the target is to unlearn information related to "Elon Musk".

Method	Selection Time (s)	Unlearn Time (s)	Total Time (s)
Baseline	0.0	1050.1	1050.1
w/ RUM	4.6	1045.1	1049.7
w/ UPCORE	4.3	3301.5	3305.8
w/ ALISE	8.2	1033.6	1041.8

We compare the unlearning time across different data selection methods under the SLUG-based CLIP unlearning setting. As shown in Table A6, ALISE maintains a total unlearning time comparable to the baseline, even when accounting for data selection overhead. UPCORE also shows a similar overall runtime to the baseline; however, it exhibits shorter selection time but longer unlearning time. This difference arises from the amount of data used during unlearning. Specifically, UPCORE excludes only a small subset and uses most of the forget set, resulting in an unlearning time close to the baseline. In contrast, ALISE performs unlearning using only a selected small subset, which reduces the data volume and thus shortens the unlearning time. RUM divides the dataset into three groups and performs unlearning sequentially on each group. As a result, the unlearning process is repeated multiple times, leading to approximately three times longer runtime compared to other methods.

H. Limitations

ALISE does not explicitly determine the number of samples to select (*i.e.*, the budget). The method constructs a Pareto front based on two alignment scores and selects samples that lie on this front. The number of samples on the Pareto front can vary depending on factors such as the model, the unlearning algorithm, and the dataset. While ALISE can increase the number of selected samples by including those from the top n Pareto layers obtained via non-dominated sorting (NDS) (Deb et al., 2002), it remains difficult to match a predefined budget exactly.

One way to address this issue is to adopt a budget-constrained selection strategy inspired by NSGA-II (Deb et al., 2002). Specifically, let $L_1, L_2, \ldots$ denote the Pareto layers obtained via NDS. We construct the subset by sequentially including samples from $L_1, L_2, \ldots$ in order. If including all samples in a layer L_k exceeds the budget, we perform partial selection within L_k by selecting the required number of samples based on the crowding distance. This strategy enables precise control over the subset size while preserving diversity among the selected samples.

I. Future Work

This work focuses on constructing a *new forget set* by selecting samples based on their alignment with the retain set and the full forget set. A natural extension is to select samples within the retain set that better preserve model utility while remaining disentangled from the full forget set, and to use them to construct a *new retain set* for unlearning.

J. Qualitative Results

We qualitatively analyze the generation outputs of MLLM and T2I models after unlearning target identities. For MLLMs, unlearning is performed under a constraint that limits the performance drop on the Neighbor Set of the IU dataset to within 5%. For T2I generation, we consider two scenarios where each identity ("Elon Musk" and "Mark Zuckerberg") is unlearned separately.

For MLLMs (Table A4), models unlearned with ALISE fail to generate the target identities, compared to those using the full forget set or existing data selection methods. For example, when unlearning Lady Gaga, other methods still generate the identity, whereas ALISE produces a generic description such as "a woman with blonde hair." Similarly, when unlearning Kanye West, other methods continue to generate the identity, while ALISE generates responses that focus instead on contextual attributes, such as "Donald Trump." These results show that ALISE removes target-related knowledge in MLLMs.

For T2I generation (Table A5), ALISE removes the target identity while maintaining the ability to generate other identities. In contrast, using the full forget set also removes the target identity but degrades knowledge of similar identities. This shows that ALISE removes only the target knowledge while preserving the model's utility.

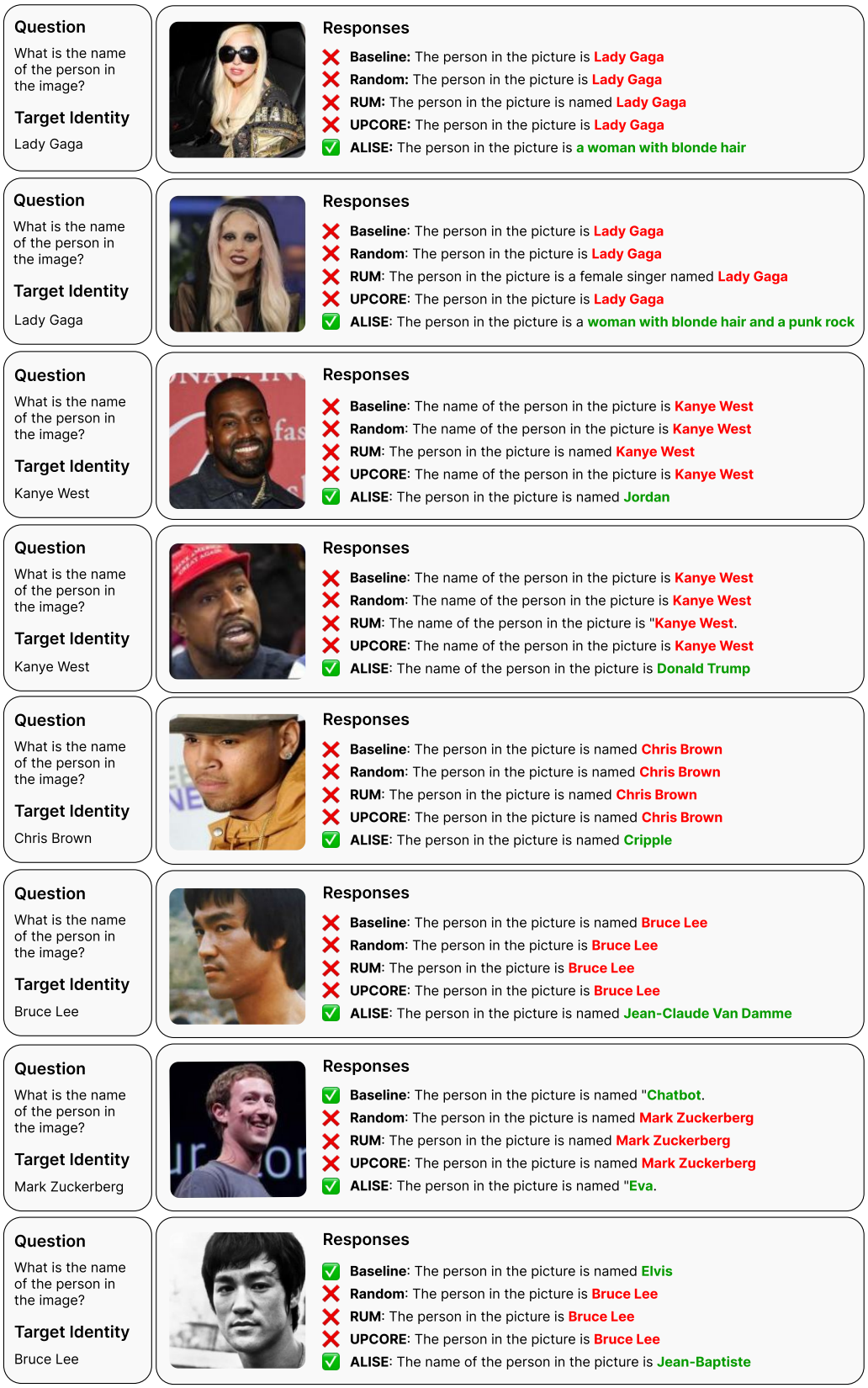

Figure A4. **Qualitative results for unlearning in MLLMs.** All experiments are conducted under a constraint that limits the drop in $\text{Acc}_{\mathcal{R}}$ on the validation split of the Neighbor Set to 5%.

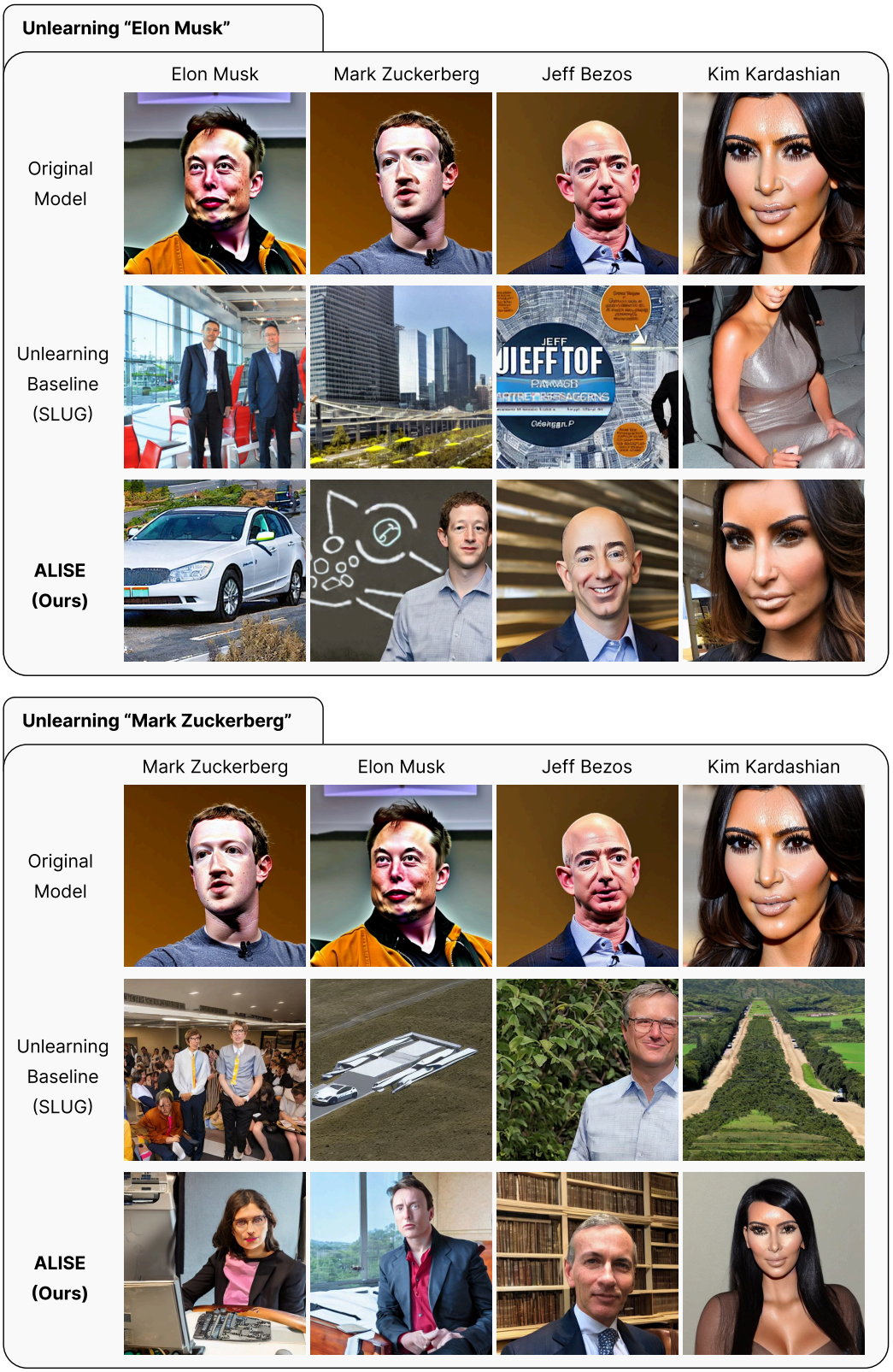

Figure A5. **Qualitative results for unlearning in T2I Generation.**