CRCE: Coreference-Retention Concept Erasure in Text-to-Image Diffusion Models

Yuyang Xue
yuyang.xue@ed.ac.uk
Edward Moroshko
emoroshk@ed.ac.uk
Feng Chen
feng.chen@ed.ac.uk
Jingyu Sun
s2091784@ed.ac.uk
Steven McDonagh
smcdonag@ed.ac.uk
Sotirios A. Tsaftaris
s.tsaftaris@ed.ac.uk

School of Engineering
University of Edinburgh
Edinburgh, UK

Abstract

Text-to-Image diffusion models can produce undesirable content that necessitates concept erasure. However, existing methods struggle with under-erasure, leaving residual traces of targeted concepts, or over-erasure, mistakenly eliminating unrelated but visually similar concepts. To address these limitations, we introduce *CRCE*, a novel concept erasure framework that leverages Large Language Models to identify both semantically related concepts that should be erased alongside the target and distinct concepts that should be preserved. By explicitly modelling coreferential and retained concepts semantically, *CRCE* enables more precise concept removal, without unintended erasure. Additionally, we contribute *CorefConcept*, a comprehensive dataset encompassing objects, intellectual property, and personal identities, which we make publicly available to support future research in concept erasure. The code and dataset are available at here.

1 Introduction

Text-to-Image (T2I) diffusion models have demonstrated a remarkable ability to generate diverse, high-fidelity images from text prompts [30, 31, 33]. However, these powerful generative models may also produce undesirable concepts gleaned from extensive training corpora such as LAION-5B [35]. These concepts encompass *pornographic content*, *violent* or *hateful imagery* [43], *copyrighted art styles* or *specific private identities* [16, 37], *social biases* [24, 38], and other *sensitive* materials [28, 34]. Although the public release of these T2I models has made image generation technology widely accessible, the presence of such undesirable concepts raises serious legal, safety, and ethical concerns [14, 32, 36]. Consequently,

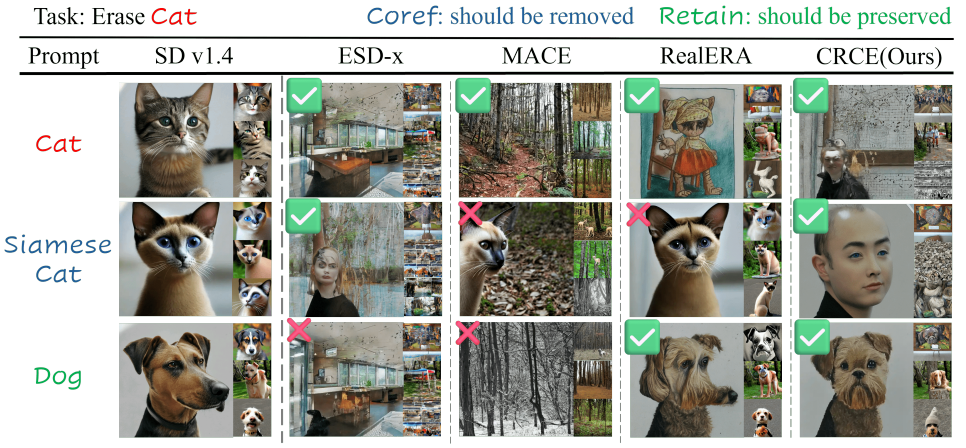

Figure 1: Consider the task of erasing the "Cat" concept. We define "Siamese cat" as coreference (coref), which should *also* be erased, and "Dog" as a retain concept, *i.e.* which should *not* be erased. We show examples from SD v1.4 before any erasure and proceed with results from concept erasure methods. Green checkmarks (✓) indicate successful erasures or retentions, while red crosses (×) highlight failures. Our approach effectively balances erasure and retention, reducing both under- and over-erasure issues compared to existing methods.

efforts have emerged to remove specific concepts so that models can no longer reproduce them – an approach collectively referred to as **concept erasure**.

Concept erasure modifies a pre-trained generative model to eliminate a target concept from its output while preserving the model's ability to generate the remaining concepts [18, 41]. Unlike post-intervention approaches such as filtering outputs and employing safety checkers at runtime [51], concept erasure pursues proactive solutions that alter the model's parameters, effectively removing the concept such that it cannot be easily evaded by cunning red-teaming prompts [39] or direct weight access [7].

However, despite progress in the field, we corroborate recent findings [4] identifying persistent issues, even with recent models. Fig. 1 shows this graphically. Erasing the concept "cat" can affect an unrelated concept "dog", while semantically related concepts such as "Siamese cat" may resist erasure. During concept erasure, some **coreferential** concepts (*corefs*) may remain *under-erased* (or have *concept residues* [22]), whereas **retained**, unrelated, concepts (*retains*) may be removed (*over-erased*). These issues occur not only with objects but also with identities. Although methods based on Euclidean distance, cosine similarity, and CLIP [29] have shown progress (further discussed in Sec. 2.2), these challenges remain largely unresolved.

In this paper, we propose an intuitive yet effective concept erasure framework named Coreference-Retention Concept Erasure (*CRCE*). Given a target concept, we employ a Large Language Model (LLM) to generate semantically accurate *coref* and *retain* concepts. These corefs and retains are jointly integrated in a new loss function, which effectively balances erasing the target concept (and its corefs) while preserving unrelated knowledge. *CRCE* can be easily generalised to diffusion-based models due to its flexibility and simplicity.

Our **main contributions** are summarised below:

1. To the best of our knowledge, *CRCE* is the first method to simultaneously address

both under- and over-erasure of concepts by leveraging LLMs to identify semantic relationships between concepts along their natural manifolds rather than relying solely on cosine similarity or Euclidean proximity.

2. We propose and validate a novel loss function that jointly integrates LLM-generated coref and retain concepts with weighted certainty levels, enabling targeting of the concept manifold while preserving unrelated knowledge. Comprehensive experiments show that our method outperforms existing concept erasure techniques.

3. We introduce *CorefConcept*, a comprehensive dataset that maps semantic relationships between target concepts and their corefs/retains with quantified certainty metrics, providing a valuable resource for future research on concept erasure and understanding of semantic relationships in embedding spaces.

2 Preliminaries

To set the scene for our approach, we start by introducing useful preliminary knowledge about the background and limitations of existing concept erasure methods, particularly regarding assumptions on the embedding space of CLIP.

2.1 Baseline: Erasing Stable Diffusion (ESD)

ESD [7] is a self-distillation method; a frozen copy of the original model guides a fine-tuned copy to eliminate the concept. Consider a pre-trained Stable Diffusion (SD) model with weights θ^*, a target concept embedding c that requires removal, and a copy of the model with weights θ that needs to be fine-tuned. During fine-tuning, ESD generates synthetic training noise by using the frozen model θ^* to guide the tuned model away from c. At each training step, a random noise latent x_t is generated at a specific diffusion timestep t, which is input into the tuned model θ, conditioned on a prompt containing concept c. Concurrently, the same noise is input into the frozen model θ^* with both the concept c and an unconditional empty prompt. These inputs are used to calculate an anchor noise prediction that represents the intended target concept. The tuned model θ is updated to align predictions more closely with this anchor. The process employs classifier-free guidance, described by:

$$\varepsilon_{\theta^*}^{\text{anchor}}(x_t, c, t) = \varepsilon_{\theta^*}(x_t, t) - \eta[\varepsilon_{\theta^*}(x_t, c, t) - \varepsilon_{\theta^*}(x_t, t)],$$

where η denotes the strength of the negative guidance. The fine-tuning objective seeks to make the predictions conform to this guided anchor, by minimizing[1]:

$$\mathcal{L}_{\text{ESD}} = \|\varepsilon_\theta(x_t, c, t) - \varepsilon_{\theta^*}^{\text{anchor}}(x_t, c, t)\|_2^2 . \tag{1}$$

Importantly, ESD's beauty lies in utilizing the frozen model to instruct the newly adjusted model using only inherent knowledge, effectively reversing the concept's influence. However, ESD has a key limitation: it focuses on removing the specific target concept without considering semantically related or unrelated concepts, as the example in Fig. 1 illustrates. Hence methods that better account for the semantic relationships between concepts are needed.

[1] We consistently omit $\mathbb{E}_{z_t,t,c,\varepsilon\sim\mathcal{N}(0,1)}$ for the sake of simplicity.

2.2 Related Work on Addressing Under-/Over-Erasure

Numerous methods aim to balance concept removal and utility in T2I models but do not address over-/under-erasure in totality. Techniques like CA [18] ablating the anchor to its super-class (*e.g.* "grumpy cat" to "cat") and attention-resteering FMN [41] suffer from under-erasure or unanalysed trade-offs. Closed-form editing methods such as UCE [8] and MACE [23] overlook synonym leakage or rely on non-automated semantic discovery. Others like SPM [25], RECE [9], Receler [13], and AdvUnlearn [42] use heuristics, shifting retains, red-teaming, gradient cues, or fixed banks, respectively, without fully resolving both erasure issues simultaneously. Retain set selection [2] can be problematic for abstract concepts and super-classes. While all above methods use anchors, regularisers, or adversarial retains to protect unrelated content, none jointly optimise a certainty-weighted set of coreferential and retain prompts, a gap our method addresses. A more comprehensive review of related work on latent diffusion models and concept erasure is provided in Appendix A.

2.3 CLIP Embedding Similarity vs. Semantic Similarity

Trained on a massive dataset of image-text pairs, CLIP [29] encodes images and text into a shared embedding space using a contrastive objective, making it a popular component in T2I models. CLIP similarity is a popular metric for exploiting and evaluating text/image embeddings by measuring the cosine similarity between their normalized vectors. Ideally, cosine similarity within CLIP's embedding space would reflect human notions of semantic similarity—conceptually similar images should be positioned close together, while dissimilar ones should be spread further apart. However, recent studies have demonstrated that human perception of semantic similarity does not always correspond with cosine distances within CLIP's representation space [6, 20]. In certain cases, CLIP similarity can appear counterintuitive from a human perspective, revealing a discrepancy between the model's learnt representations and human semantics. Fig. 2 illustrates the concept of dispersed prompt embeddings, in CLIP space, that are not clustered around the concept centre. When examining Euclidean distances, we find counterintuitive relationships: "dog"-"cat" (0.419) and "dog"-"pig" (0.466) are actually closer in the CLIP embedding space than "dog"-"service dog" (0.533) and "dog"-"guide dog" (0.539). Appendix D provides an extensive list of distances.

We argue that making inferences on the geometry of the CLIP embedding space (as RealEra [22] does) is not the most appropriate approach to fully determine how a model should erase concepts. Our approach instead bypasses this issue by explicitly finding sets of concepts that should be used as corefs and those that shall remain (retain) to directly guide the erasure process to reduce instances of under- and over-erasure.

3 Method: Coref-Retain Concept Erasure (CRCE)

To tackle both under- and over-erasure, our driving hypothesis is that these issues stem from how concepts are structured within the CLIP semantic space. At the core of our approach is the identification of coreferential (coref) concepts and retained (retain) concepts, which more accurately reflect semantic relationships. Rather than relying on naive spherical assumptions around target concepts, we leverage LLMs to guide the discovery of corefs that lie on the same semantic manifold as the target, and retain concepts that, while they may not be nearby in embedding space, should be preserved. This manifold-aware perspective enables more

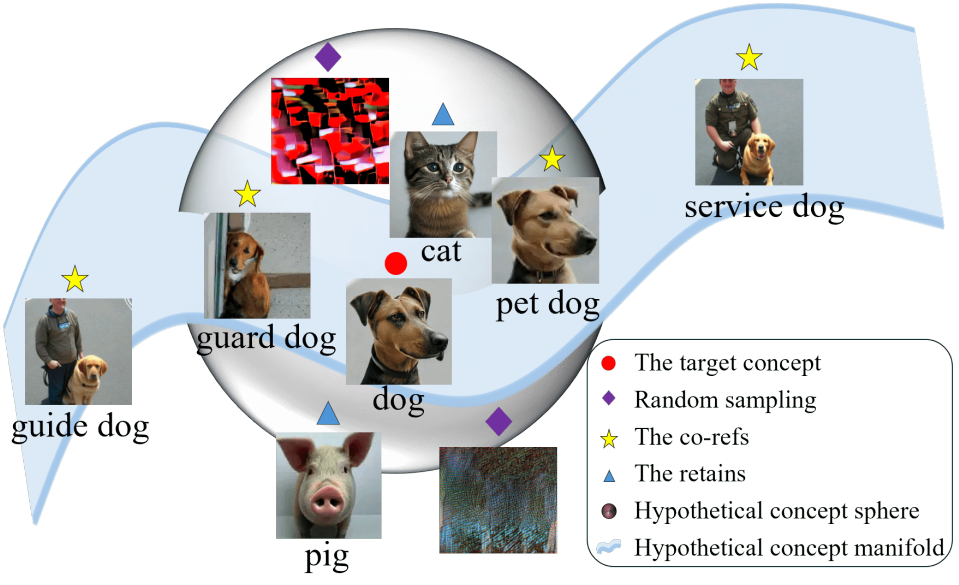

Figure 2: Illustration of how related and unrelated concepts to "dog" are arranged in CLIP's embedding space. The red dot marks the target concept "dog"; yellow stars (*e.g.* "guide dog", "service dog") represent coreferent concepts along the same semantic manifold, while blue triangles (*e.g.* "cat", "pig") denote unrelated concepts to be retained. Neglecting corefs and retains leads to under-/over-erasure. RealEra [22] samples random corefs in a spherical region, which poorly approximates the true semantic geometry—often capturing unrelated concepts (purple diamonds) - and ignores the non-Euclidean nature of concept relationships, where semantically distinct concepts may appear close in Euclidean space (blue triangles).

precise and controlled concept erasure. Building upon the ESD framework, we introduce two key innovations: (1) an LLM-guided mechanism for discovering coref and retain concepts, and (2) a synthetic dataset, *CorefConcept*, to support this process. Additionally, we propose a loss function that jointly removes target concepts and their corefs while preserving unrelated retains. Together, these contributions form our Coref-Retain Concept Erasing (CRCE) approach, a robust and comprehensive solution to mitigate both under-erasure and over-erasure. Fig. 3 illustrates the complete pipeline of our proposed CRCE method. The process begins with a target concept for erasure, which is passed to an LLM to generate appropriate corefs and retains with associated certainty levels. These generated concepts are then integrated into our loss function, which guides the model to effectively remove the target concept and its corefs while preserving retains.

3.1 LLM-based Generation of Corefs and Retains

We first formalise our definition of coreferences, as prior works may use the term with varying meanings. We define coreferences (corefs) as a set that includes the original concept itself, including its **synonyms and any possible expressions** that may lead to the same conceptual understanding. Although some concepts might lack direct synonyms, they can have coreferential expressions. For instance, corefs for "Mickey Mouse" could be "The first

mouse character by Walt Disney"; "The 45th President of the United States" could be an appropriate coref for "Donald Trump". For retained concepts, we take a more nuanced approach, beyond simply selecting arbitrary unrelated concepts. We define retains as concepts that share semantic proximity to the target concept without being coreferential. These concepts typically exist in the same categorical neighbourhood, but are fundamentally distinct entities. Importantly, we select retains that are close enough in the embedding space to potentially be affected by over-erasure, making them valuable indicators of erasure precision. For example, when erasing "dog", concepts such as "cat" and "pig" serve as ideal retains because they share categorical similarities (mammals) while being distinct in both taxonomy and human perception. This careful selection of semantically adjacent but distinct concepts allows us to effectively mitigate over-erasure. Note that using completely unrelated concepts (such as "airplane" when erasing "dog") would be trivially easy to retain and would fail to meaningfully evaluate the model's ability to avoid over-erasure, as such semantically distant concepts are unlikely to be affected even by imprecise erasure methods.

As illustrated in Fig. 2, randomly sampled embeddings often fail to accurately identify the correct corefs and retains, which are crucial for effective concept erasure. However, manually determining these corefs and retains for each concept is both time-consuming and labour-intensive. To address this challenge and systematically identify appropriate coreferential and retain concepts, we leverage the comprehensive semantic knowledge embedded in LLMs. These models excel at understanding nuanced relationships between concepts, making them ideal for generating semantically accurate concept associations that would be impractical to define manually at scale.

Our approach employs a structured prompting methodology that guides the LLM to generate distinct sets of concept relationships. For each target concept, the LLM produces a ranked list of coreferential concepts that maintain visual and semantic connections to the target. These corefs range from direct synonyms to more loosely associated terms, each assigned a certainty level ("Very High", "High", "Normal", "Low" or "Very Low") that quantifies the strength of their relationship to the target concept.

The prompt encourages corefs to prioritise synonyms and high-precision semantic equivalents, while retains focus on concepts that share categorical proximity without being identical. Our carefully crafted prompts explicitly discourage vague descriptions or unrelated concepts, promoting specificity and relevance in the generated associations. In cases where the LLM generates inappropriate concept relationships, the system supports iterative refinement through multi-round conversations, allowing domain experts to adjust and optimise the generated corefs and retains as needed. For a more detailed prompt template see Appendix C.

3.2 *CorefConcept*: A Dataset

We employ the data-processing method described above to create a dataset *CorefConcept* with three categories: objects, Intellectual Property (IP), and celebrities. We included CIFAR-10 class labels and some concepts that may be ambiguous (*e.g.* "bat (animal)" vs "bat (sports equipment)") to form the object category. IP and celebrity categories were generated using the ChatGPT-o1 reasoning assistant [15]. Specifically, for celebrities, we provided the prompt: "*Provide a list of well-known celebrities, including various gender, ethnicity, age, etc.*" For IP items, we prompted: "*Provide a list of well-known IP characters from different media, such as movies, books, cartoon, games, etc.*" The dataset contains 20 object concepts, 15 IP characters and 15 celebrities. Each concept was augmented with a list of 15 corefs and a list of 15 retains. We randomly selected 10 corefs per concept for the training set, with the

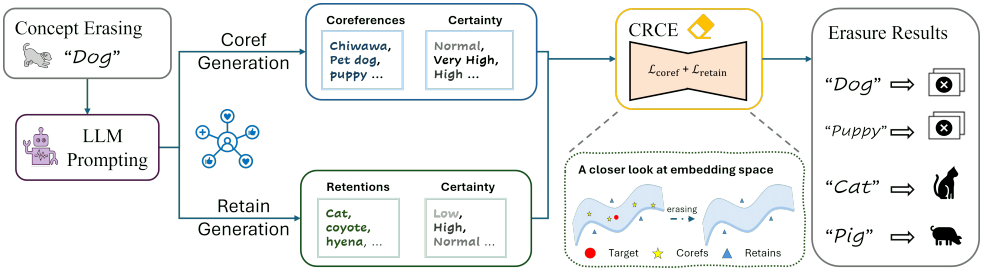

Figure 3: Overview of our proposed CRCE method. Our method erases a target concept (*e.g.* "dog") while preserving unrelated concepts. Using LLM prompting, we generate corefs (*e.g.* "Chihuahua", "puppy") and retains (*e.g.* "cat", "coyote"), each assigned a certainty score. The CRCE loss optimises both coref erasure and retain preservation, adjusting the embedding space to minimize unintended removals. The final erasure results show that the target ("dog") and its coreferential variants ("puppy") are erased, while unrelated concepts ("cat", "pig") remain intact, ensuring effective and controlled concept erasing.

remaining 5 used for the test set. Full details, including prompts used and dataset samples, are available in Appendices C.1 and E, respectively. This dataset provides a valuable resource for studying coreference and disambiguation in vision-language models, supporting research on reasoning and representation learning across diverse and ambiguous concepts. [2]

3.3 Coreference and Retention-Aware Loss

We integrate generated coref and retain samples into the loss function to improve concept erasure performance. We add terms to the ESD loss to explicitly account for coref and retain concepts with their associated certainty levels. The loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{ESD}} + \frac{1}{M} \sum_{c' \in \text{Coref}} \mathcal{C}_{c'} \cdot \left[\| \varepsilon_\theta(c') - \varepsilon_{\theta^*}^{\text{anchor}}(c) \|_2^2 \right] + \frac{1}{N} \sum_{r \in \text{Retain}} \mathcal{C}_r \cdot \left[\| \varepsilon_\theta(r) - \varepsilon_{\theta^*}(r) \|_2^2 \right],$$

where $\mathcal{L}_{\text{ESD}}$ represents the ESD loss from Eq. 1. We introduce two additional terms that incorporate coref concepts c' and retain concepts r with each weighted by their certainty levels $\mathcal{C}_{c'}$ and $\mathcal{C}_r$ (with the certainty of each concept determined by the LLM). For each iteration, we randomly sample M corefs and N retains from the training corefs and the training retains. The discrete certainty levels are $\{1.0, 0.8, 0.6, 0.4, 0.2\}$ corresponding from "Very High" to "Very Low". Hence, concepts with higher certainty contribute more significantly to the loss, thereby prioritising alignment with target corefs and retains.

4 Experiments

4.1 Experimental Setup

Approaches for comparison: We treat ESD-x [7] as a baseline method in the fine-tuning concept erasure paradigm. We also compare our method to other state-of-the-art methods:

[2] https://github.com/vios-s/CRCE-Coreference-Retention-Concept-Erasure-in-Text-

Table 1: Quantitative comparison of concept erasure methods across Object, Intellectual Property (IP), and Celebrity categories. Top-3 accuracy are marked in **bold**, underlined, and *italics*, respectively. Our method (CRCE) consistently achieves the most balanced accuracy, effectively removing the target concepts and their coref (Acc_U, Acc_C^{test}) while preserving unrelated retained concepts (Acc_R^{test}). Empirically, we set the number of corefs and retains to 5 and 3, according to Table 7 in Appendix F.2.

	Object					Intellectual Property (IP)					Celebrity				
	$Acc_U \downarrow$	$Acc_C^{train} \downarrow$	$Acc_C^{test} \downarrow$	$Acc_R^{train} \uparrow$	$Acc_R^{test} \uparrow$	$Acc_U \downarrow$	$Acc_C^{train} \downarrow$	$Acc_C^{test} \downarrow$	$Acc_R^{train} \uparrow$	$Acc_R^{test} \uparrow$	$Acc_U \downarrow$	$Acc_C^{train} \downarrow$	$Acc_C^{test} \downarrow$	$Acc_R^{train} \uparrow$	$Acc_R^{test} \uparrow$
SD [🖼]	83.33	87.22	91.60	81.17	80.63	88.33	37.46	30.13	86.80	86.75	90.90	17.12	17.94	92.64	93.24
ESD-x [🖼]	3.33	37.72	39.92	73.82	69.89	3.33	10.67	7.73	62.80	71.16	*1.81*	11.09	10.76	72.9	68.37
FMN [🖼]	4.00	50.78	50.45	30.06	25.92	4.93	4.45	*3.89*	2.88	2.51	9.2	4.38	4.00	7.92	5.23
UCE [🖼]	1.27	8.52	8.40	17.91	14.58	*1.17*	1.38	6.57	1.01	1.02	1.47	3.14	3.03	5.83	5.00
SPM [🖼]	36.00	31.04	33.29	14.57	15.45	16.86	9.52	10.13	5.87	8.05	16.85	15.78	19.41	33.94	34.35
Receler [🖼]	2.40	**6.53**	**7.09**	11.68	8.71	0.27	**0.77**	**1.74**	0.87	0.91	0.53	*2.87*	1.76	4.11	4.96
MACE [🖼]	*2.10*	*12.10*	*9.05*	62.73	68.84	1.33	*3.33*	3.73	44.53	3.73	**0.00**	3.33	2.40	29.73	30.40
RealEra [🖼]	48.88	70.49	72.60	**88.92**	*77.15*	3.33	16.67	14.66	77.73	**81.81**	12.72	11.78	11.53	82.19	85.40
CRCE-Sphere	26.67	58.11	62.40	86.57	**84.84**	20.00	13.73	12.00	*75.20*	*79.47*	7.27	11.78	11.53	82.19	*82.70*
CRCE-Fixed	8.75	19.27	32.21	78.23	72.39	7.66	4.80	4.26	67.94	69.47	5.57	**1.36**	*2.59*	*81.54*	79.45
CRCE	**1.25**	12.81	26.31	*85.20*	79.56	**0.00**	6.27	7.46	**78.80**	80.25	*1.81*	2.04	**1.29**	**87.35**	**87.02**

FMN [41], UCE [8], SPM [25], Receler [13], MACE [23], and RealEra [22]. See Appendix B for more information about implementation details.

Evaluation metrics: Conventional concept erasure evaluations use classifiers (*e.g.* ResNet [10]) or CLIP Score [11] to assess erasure and preservation accuracy. However, classifiers face out-of-distribution (OOD) issues, and CLIP similarities may not reliably reflect semantic alignment with concepts [20] (see Sec. 2.3). To address this, we propose an alternative evaluation protocol leveraging VLMs, which are well-suited for recognising complex visual patterns across diverse domains. Specifically, we select the open-source Qwen2 VL [40] to instantiate performance evaluation. We prompt the model using the prompt: "*Observe and describe the image and check whether it has the same concept as* <X>", where the variable is instantiated by an instance of an object, character, or celebrity identity. We restrict the model to provide binary output (*yes* or *no*), as a straightforward classifier.

We assess concept erasure performance with metrics that evaluate both erasure effectiveness and concept preservation: $Acc_U \downarrow$ measures how effectively the primary target concept is removed; $Acc_C^{train} \downarrow$ captures the erasure of coref concepts seen during fine-tuning, which are semantically related and should also be erased; To assess generalisation, $Acc_C^{test} \downarrow$ evaluates erasure accuracy on unseen coref concepts; On the preservation side, $Acc_R^{train} \uparrow$ reflects how well unrelated but semantically close concepts are retained during training, while $Acc_R^{test} \uparrow$ measures the model's ability to generalise this preservation to unseen retain concepts. This train/test split approach assesses both performance on concepts used during optimisation and generalisation to unseen concepts. An effective method will have low values for erasure metrics and high values for preservation metrics, suggesting better discrimination between related and unrelated concepts.

4.2 Quantitative Results

Table 1 shows a clear pattern: existing erasure methods excel on one side of the coref-retain trade-off but stumble on the other, whereas **CRCE balances both**. From Acc_U, Acc_C^{test} and Acc_R^{test} for Object, we can observe that RealEra, which retains concepts by sampling from a random CLIP-sphere, preserves unrelated concepts best, *e.g.* it keeps retain accuracy at 77.15%, yet leaves the target itself recognisable in 48.88% of cases and coreferential variants in 72.60%, underscoring severe under-erasure. Receler achieves the opposite extreme: by steering cross-attention to a single anchor it drives target accuracy to 2.40% and

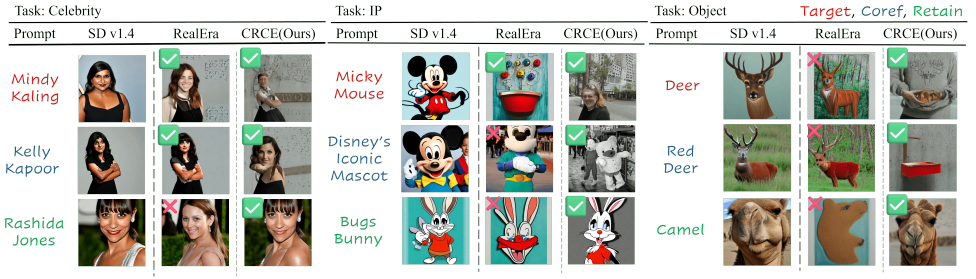

Figure 4: Comparison of concept erasure effectiveness between RealEra [22] and our method. "Mindy Kaling" (celebrity), "Mickey Mouse" (IP), and "Deer" (object) are targeted for removal along with their corefs. CRCE successfully erases corefs while accurately retaining related yet distinct entities, demonstrating superior precision compared to RealEra.

corefs to 7.09%, but the collateral damage is strong, with retain accuracy collapsing to 8.71% on the test split, showing serious over-erasure. Adapter-based SPM (latent anchoring) and attention-reweighting FMN attempt to strike middle ground—where retain accuracy rises to 15.45% to 25.92%, respectively, yet more than one-third of corefs are not erased, while robustness-oriented UCE wipes out almost everything, leaving retain accuracy below 15%. MACE does a relatively good job at removing both the target and its corefs, while preserving two thirds of retains. In contrast, CRCE drives the accuracy of under-erasure to single digits and keeps retain accuracy above 79% in every domain: on Objects it lowers Acc_U to 1.25% and coref accuracy to 12.8% while still scoring 79.6% on retains; on Intellectual-Property (IP) the corresponding numbers are 0%, 7.46%, and 80.25%; on Celebrities 1.81%, 1.29%, and 87.02%. These quantitative margins match visual inspection (see Appendix H for more details): rivals that protect image quality often leave clear visual traces of the target (*e.g.* object "horse" cannot be erased using FMN or RealEra), whereas those that purge aggressively distort or erase desirable details (*e.g.* SPM and MACE often produce over-smoothed textures, Receler's prompts yield pure colours). CRCE alone removes every instance together with their corefs while leaving retains untouched, empirically demonstrating that the LLM-derived coref/retain manifolds and the certainty-weighted objective deliver the sought-after balance between erasure precision and content preservation. Fig. 4 presents the selected visual comparisons between our method and RealEra across object, celebrity, and IP categories. Comprehensive results comparing all methods are presented in Appendix H.

4.3 Ablation Study: Variations of CRCE

To better understand the impact of different sampling strategies on concept erasure performance, We ablate the sampling strategy via two variants: **CRCE-Sphere** employs RealEra's CLIP-sphere sampling strategy but uses our loss. Results show that it reduces Acc_U on Objects from 48.88% to 26.67% and lifts retain accuracy from 72.60% to 84.84%, confirming the effectiveness of our loss even under Euclidean assumptions. **CRCE-Fixed** trains on one static coref/retain list, giving consistent but narrow supervision. **CRCE** (our full method) randomly samples new coref/retain prompts at each step, covering the manifold more completely and therefore beats CRCE-Sphere/Fixed across all metrics as shown in Table 1. The gaps between **Sphere** to **CRCE-Fixed** to **CRCE** underscore that (i) CLIP space

is not isotropic and (ii) dynamic sampling prevents over-fitting to a local concept region.

The results of other two ablation studies on *the coref and retain certainty* and *their number needed* are available in Appendix F. These ablations show that structured certainty scores, especially for corefs, are crucial for effective and robust concept erasure, as random or uniform certainty leads to significant performance degradation. Additionally, using a moderate number of corefs and retains (*e.g.* $M = 5$, $N = 3$) achieves the best balance across metrics, while overly large values introduce semantic interference and reduce generalisation performance.

5 Conclusions

We introduce CRCE, a novel concept erasure approach for T2I diffusion models that tackles under- and over-erasure by leveraging LLMs to identify semantically related (coref) and unrelated (retain) concepts, enabling more precise erasure compared to distance-based sampling methods. Our *CorefConcept* dataset and tailored loss enable fine-grained erasure and preservation. CRCE outperforms prior methods across diverse tasks and is not restricted to the ESD backbone but can seamlessly integrate with any diffusion-based concept erasure framework. Limitations such as difficult cases of concept entanglement are addressed in Appendix G. Future work may extend CRCE to stylistic or explicit concepts, potentially incorporating human feedback to refine semantic relationships.

Acknowledgements

Y. Xue thanks additional financial support from the School of Engineering, the University of Edinburgh. S.A. Tsaftaris acknowledges support from the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme (grant RCSRF1819/8/25), and the UK's Engineering and Physical Sciences Research Council (EPSRC) support via grant EP/X017680/1, and the UKRI AI programme and EPSRC, for CHAI - EPSRC AI Hub for Causality in Healthcare AI with Real Data [grant number EP/Y028856/1].

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[2] Anh Bui, Trang Vu, Long Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them. *Int. Conf. Learn. Represent.*, 2025.

[3] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *Int. Conf. Mach. Learn.*, 2023.

[4] Somnath Basu Roy Chowdhury, Kumar Avinava Dubey, Ahmad Beirami, Rahul Kidambi, Nicholas Monath, Amr Ahmed, and Snigdha Chaturvedi. Fundamental limits of perfect concept erasure. In *International Conference on Artificial Intelligence and Statistics*, pages 901–909. PMLR, 2025.

[5] 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. In *Int. Conf. Learn. Represent.*, 2023.

[6] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Adv. Neural Inform. Process. Syst.*, 2023.

[7] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Int. Conf. Comput. Vis.*, pages 2426–2436, 2023.

[8] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Winter Conf. Appl. Comput. Vis.*, pages 5111–5120, 2024.

[9] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. pages 73–88. Springer, 2024.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *Empirical Methods in Natural Language Processing*, 2021.

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Int. Conf. Learn. Represent.*, 1(2):3, 2022.

[13] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. pages 360–376. Springer, 2024.

[14] Tatum Hunter. Ai porn is easy to make now. for women, that's a nightmare. *The Washington Post*, 2023.

[15] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[16] Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. Ai art and its impact on artists. In *AAAI*, pages 363–374, 2023.

[17] Changhoon Kim and Yanjun Qi. A comprehensive survey on concept erasure in text-to-image diffusion models. *arXiv preprint arXiv:2502.14896*, 2025.

[18] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22691–22702, 2023.

[19] Black Forest Lab. Announcing flux1.1 [pro] and the bfl api, Oct 2024. URL https://blackforestlabs.ai/announcing-flux-1-1-pro-and-the-bfl-api/.

[20] Siting Li, Pang Wei Koh, and Simon Shaolei Du. On erroneous agreements of clip image embeddings. *arXiv preprint arXiv:2411.05195*, 2024.

[21] Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. In *Eur. Conf. Comput. Vis.*, pages 93–109. Springer, 2024.

[22] Yufan Liu, Jinyang An, Wanqian Zhang, Ming Li, Dayan Wu, Jingzi Gu, Zheng Lin, and Weiping Wang. Realera: Semantic-level concept erasure via neighbor-concept mining. *arXiv preprint arXiv:2410.09140*, 2024.

[23] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6430–6440, 2024.

[24] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Adv. Neural Inform. Process. Syst.*, 36:56338–56351, 2023.

[25] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. pages 7559–7568, 2024.

[26] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas.

[27] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Removing nsfw concepts from vision-and-language models for text-to-image retrieval and generation. *Eur. Conf. Comput. Vis.*, 2024.

[28] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pages 3403–3417, 2023.

[29] 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 *Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021.

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022.

[32] Kevin Roose. An ai-generated picture won an art prize. artists aren't happy. *The New York Times*, 2022.

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inform. Process. Syst.*, 35:36479–36494, 2022.

[34] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22522–22531, 2023.

[35] 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. *Adv. Neural Inform. Process. Syst.*, 35:25278–25294, 2022.

[36] Riddhi Setty. Ai art generators hit with copyright suit over artists' images. *Bloomberg Law*, 2023.

[37] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6048–6058, 2023.

[38] Lukas Struppek, Dom Hintersdorf, Felix Friedrich, Patrick Schramowski, Kristian Kersting, et al. Exploiting cultural biases via homoglyphs in text-to-image synthesis. *Journal of Artificial Intelligence Research*, 78:1017–1068, 2023.

[39] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *Int. Conf. Learn. Represent.*, 2024.

[40] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[41] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 1755–1764, 2024.

[42] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. 2024.

[43] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *Eur. Conf. Comput. Vis.*, pages 385–403. Springer, 2024.

A Additional Related Work

A.1 Text-to-Image (T2I) Diffusion Models

Text-to-Image (T2I) generation is a branch of generative modelling that focuses on producing images from textual descriptions, where diffusion-based solutions currently dominate the field. Stable Diffusion (SD) [31] constitutes a highly popular open-source latent diffusion model (LDM) that reduces computational costs by operating within a compressed latent space. eDiff-I [1] uses an ensemble of specialised diffusion experts to handle highly compositional prompts, while Muse [3] explores masked generative transformers for parallelised image synthesis. Similarly, Flux [19] proposes an architecture that combines diffusion and autoregressive components to balance the generation speed with high-quality outputs. Emerging research on concept erasure addresses pressing concerns around privacy and content moderation, highlighting the ethical and legal dimensions of large-scale T2I models. We next review latent diffusion models (LDMs) in more detail. In our study, we build upon SD v1.4 [31] as the foundational model for implementing concept erasure algorithms, given its recent popularity and quasi benchmark status.

LDM [31], typically have three main components: 1) a pre-trained vision autoencoder to compress high-dimensional image data into a low-dimensional latent representation. The encoding network $\mathcal{E}(\cdot)$ maps an image x to a latent variable z, and the decoding network $\mathcal{D}(\cdot)$ reconstructs the image from the latent space such that $\mathcal{D}(z) = \hat{x} \approx x$. 2) The text encoder transforms text prompts into conditioning vectors, allowing a condition over the generation process. The textual prompt p is embedded as $c = \mathcal{T}(p)$, where $\mathcal{T}$ is a text encoder, typically CLIP [29]. 3) A latent diffusion model for iterative denoising that accepts text embedding c through the cross-attention layers in the core generative process, governed by a U-Net-based [26] LDM that progressively refines noisy latent representations towards high-fidelity outputs along the diffusion trajectory.

A.2 Concept Erasure

The primary objective of concept erasure is to condition the model such that it inherently removes the concept in response to undesired prompts. This can be accomplished through the following paradigms.

Closed-form Model Editing offers a direct mathematical update to model parameters without requiring iterative training. A general framework for closed-form model editing utilises least squares-based optimisation, primarily focusing on the key and value projection matrices within the LDM cross-attention module. UCE [8] employs the closed-form solution for concept erasure, allowing scalable moderation and debiasing. MACE [23] uses adapters for large-scale concept erasure incorporating LoRA [12]. RealEra [22] inherited the MACE strategy and adds random sampling of neighbouring concept embeddings lying in a sphere in CLIP embedding space to the erasure target. As demonstrated in Sec. 2.3, relying on geometric assumptions about the CLIP embedding space fundamentally misrepresents the true semantic relationships between concepts. This class of approaches is appealing due to their non-iterative nature; however, we note that they often struggle with fine-grained control over concept erasure, potentially leading to unintended distortions in related but distinct concepts or fail to fully disentangle deeply entangled concepts, making them less effective for complex or nuanced modifications compared to iterative approaches.

Fine-tuning is one of the most intuitive strategies to remove unwanted concepts from

T2I models, which we also adopt in our work. FMN [41] minimises attention activation to redirect attention mechanisms, effectively eliminating certain concepts. ESD [7] modifies noise prediction to remove concepts and conducts a detailed study on the most suitable module within the LDM for fine-tuning. We have reviewed relevant details of ESD in Sec. 2.1. SalUn [5] proposes saliency-based weights to achieve concept removal. Instead of adjusting the U-Net in LDM directly, Safe-CLIP [27] and Latent Guard [21] employ safe and unsafe pairs data to fine-tune the CLIP embedding. However, fine-tuning still poses the risks of over-erasure and unintended degradation of the model's capabilities. For further details, refer to a recent comprehensive survey paper [17].

B Implementation details

We conducted a comprehensive evaluation of various concept erasure methods based on their source code: ESD [7][3], MACE [23][4], FMN [41][5], UCE [8][6], SPM [25][7] and Receler [13][8]. For RealEra [22], due to the absence of publicly available source code, we re-implemented the method based on the descriptions provided in their publication.

In the ESD framework, although ESD-u (full U-Net fine-tuning) is recommended for object-based erasure, preliminary experiments indicated significant over-erasure, adversely affecting retention accuracy. Consequently, we adopted ESD-x (cross-attention strict), which confines fine-tuning to the key (K) and value (V) weights within cross-attention modules. Other hyperparameters, including the negative guidance parameter η, were set to 1, consistent with the original paper.

For MACE, we followed the recommended configuration. Given that MACE employs anchor concepts to guide the erasure process, and our object categories often lack clearly defined super-categories (or are themselves super-categories), we utilized generic anchor concepts such as "sky" or "ground," as suggested in MACE's appendix.

In the case of FMN, we increased the number of optimisation steps to 100 to achieve a cleaner erasure result, as the default 35 steps were insufficient for most concepts. We did not employ textual inversion, considering the concepts were already inherent in the text encoder. Additionally, we prepared five images for each concept to facilitate the attention steering configuration. The resulting outputs appear somewhat blurry.

For UCE, we applied the default hyperparameter settings for each concept and designated the "concept type" as "object".

Regarding SPM, we utilized the default hyperparameters, set the "surrogate concept" as an empty string, and configured the training mode to "erase-with-la". The total number of iterations was set to 500, with "lr-warmup-steps" set to 100.

In the case of Receler, we employed the default settings.

As the official implementation of RealEra was unavailable and explicit reproduction instructions were limited, we reconstructed RealEra's approach atop ESD-x, integrating their random sampling strategy of neighbouring concept embeddings as outlined in their methodology.

[3] https://github.com/rohitgandikota/erasing
[4] https://github.com/Shilin-LU/MACE/tree/main
[5] https://github.com/SHI-Labs/Forget-Me-Not/
[6] https://github.com/rohitgandikota/unified-concept-editing
[7] https://github.com/Con6924/SPM/
[8] https://github.com/jasper0314-huang/Receler

Unless otherwise specified, all experiments, including our CRCE method, were fine-tuned for 500 iterations using a fixed learning rate of 1×10^{-5}. Other configurations adhered to the details provided in the original ESD implementation. All models were trained on Stable Diffusion 1.4 [31] from CompVis, utilising a single NVIDIA A100 80GB GPU.

C The Prompt Template for Corefs and Retains Generation

The prompt is carefully engineered to prioritize strong visual and conceptual associations with the target concept. Additionally, we incorporate hierarchical certainty levels to distinguish between highly relevant and loosely related terms. The structured approach helps mitigate under- and over-erasure issues, ensuring that concept erasure remains effective while preserving visually distinct but related elements that should not be unintentionally removed. All generations are based on ChatGPT-o1 [15].

C.1 Task Instructions

The Task Instruction.

This research program focuses on the concept erasing behaviour of Text-2-Image models, especially Stable Diffusion v1.4. The goal is to explore the under-/over-erasure behaviour of the current concept erasure algorithms. As part of this academic study, you will generate coreferential (coref) lists and retention (retains) lists for specified concepts.

1. You will be given a concept that can belong to one of the following categories: Object, Intellectual Property (IP), and Celebrity.
2. Your objective is to provide a list of 15 corefs concepts that correspond to the specified target concept.
3. The corefs should be visually related to the target concept.
4. That means these prompts can be used to generate an related image from the corefs for generative models such as Stable Diffusion.
5. Do not use very vague and general description.
6. Better not to include other irrelevant concept in the prompt.
   (a) For example, when we find corefs for the celebrity “Samuel L. Jackson”, the bad prompt such as “Frequent co-star with Bruce Willis” will let the T2I generative model generate “Bruce Willis” but not “Samuel L. Jackson” himself.
   (b) The input concept could have multiple meanings, if you find the word to be ambiguous, you can use each of its meaning to form a JSON list. For example, apple may refer to “the fruit apple” or “the tech company apple”, so you need to generate two sets of answers.

C.2 Coreferential and Retention Concept Certainty

Concept Certainty Criteria.

1. Order the concepts by their relevance or confidence:
   (a) The first item should be a synonym or the most accurate descriptor of the concept.
   (b) The last item should be the most vague or loosely related concept.
   (c) If it is possible, give more high certainty coreferential as many as possible
2. Assign a level of certainty to each item. Use the following scale: from "Very High" to "High", "Normal", "Low", and "Very Low".

The level of certainty should be based on these **Certainty Criteria**:

1. Visual and Semantic Relevance:
   (a) Coref entries are chosen because they are visually or conceptually similar to the target concept, ensuring they can prompt related images in generative models.
   (b) Retain entries are selected to represent similar but distinct concepts that should remain intact if the target concept is erased.
2. Contextual Specificity:
   (a) For celebrity and IP concepts, details such as roles, iconic traits, and narrative associations are incorporated to ensure the corefs accurately represent the subject.
3. Avoiding Vague Descriptions:
   (a) The generated terms aim to be specific enough to avoid misinterpretations by generative models.
   (b) Unrelated or overly generic descriptors are avoided to maintain high relevance.
4. Balancing Similarity and Distinctiveness:
   (a) Coref lists focus on descriptors that are tightly aligned with the target concept.
   (b) Retain lists include similar entities that are visually related but not identical, ensuring that the concept erasing process does not inadvertently remove associated, yet distinct, concepts.

D CLIP Embedding Distances

We show distances in CLIP embedding space using the target "Dog" in Table. 2. In the CLIP embedding space, the term "cat" surprisingly exhibits a higher cosine similarity (0.9128) to "dog" compared to some direct corefs like "service dog" (0.8580) or "show dog" (0.8827), highlighting potential pitfalls when relying solely on Euclidean proximity or cosine similarity for concept erasure tasks.

Corefs intended for simultaneous removal, generally exhibit high cosine similarity to "Dog", with "pet dog" (0.9192) and "puppy" (0.9166) among the closest. However, certain retained concepts (*e.g.* "wolf" at 0.8734 and "cat" at 0.9126) also have significant similarity scores, illustrating the embedding-space overlap between semantically distinct yet visually related concepts. Conversely, some retains have lower similarity scores, such as "jackal" (0.7756), potentially reflecting clear conceptual distinctions. These observations emphasise the importance of carefully structured, LLM-generated certainty criteria and semantic mapping, as implemented in our CRCE method, to accurately distinguish coref concepts from visually or semantically similar retains in the CLIP embedding space.

E *CorefConcept*: A Dataset with Corefs and Retains

We present the *CorefConcept* dataset, for evaluating concept erasure tasks in T2I models. The dataset includes three distinct categories: Object, IP, and Celebrity. Each category is annotated with precise coref and retain concepts, accompanied by clearly defined certainty levels ranging from "Very High" to "Very Low". Table. 3, Table. 4, and Table. 5 show one example for each category.

F Ablation Study

F.1 Ablation: Is the certainty necessary?

Table. 6 presents a detailed ablation study investigating the robustness of our method to noise in certainty estimates for coreference and retain concepts. Each row corresponds to a different certainty configuration: **CRCE-nocert** assigns uniform certainty values (all set to 1); CRCE-ours uses structured certainty scores generated by a Large Language Model (LLM); **CRCE-coref-*** perturbs only coreference certainty while keeping retain certainty flat (1); **CRCE-retain-*** does the opposite, and **CRCE-both-*** adds random noise to both. The ***-0** rows serve as baselines without perturbation.

The results clearly demonstrate that our full model (**CRCE-ours**) achieves the best overall performance, significantly outperforming all baselines in target erasure (Acc_{U}) and coreference generalization ($Acc_{\mathrm{C}}^{\mathrm{test}}$). Perturbing coref certainty (**CRCE-coref-***) leads to the most severe degradation in coref erasure, while perturbing retain certainty (**CRCE-retain-***) degrades retention but still preserves erasure accuracy. The **CRCE-both-*** variants exhibit moderate robustness to noise but never outperform the structured certainty setup. These findings highlight that accurate certainty — especially for coreference — is essential for effective and robust concept erasure, confirming the superiority of our LLM-based certainty design.

Table 2: An example of "Dog"'s cosine similarity and Euclidean distance compared to its corefs and retains.

Group	Words	Cosine Similarity↑	Euclidean Distance↓
coref	domestic dog	0.8927	0.4633
coref	house dog	0.9112	0.4213
coref	pet dog	0.9199	0.4002
coref	pooch	0.9166	0.4083
coref	puppy	0.9122	0.4190
coref	family dog	0.9213	0.3967
coref	canine companion	0.8843	0.4810
coref	dog breed	0.8853	0.4791
coref	working dog	0.8853	0.4790
coref	guard dog	0.8695	0.5109
coref	guide dog	0.8549	0.5388
coref	service dog	0.8580	0.5328
coref	show dog	0.8827	0.4845
coref	mongrel	0.8559	0.5369
coref	hound	0.8377	0.5697
retain	wolf	0.8734	0.5031
retain	coyote	0.8105	0.6156
retain	jackal	0.7756	0.6699
retain	fox	0.8724	0.5051
retain	dingo	0.8564	0.5359
retain	dhole	0.7176	0.7515
retain	raccoon dog	0.7182	0.7508
retain	hyena	0.7200	0.7484
retain	domestic cat	0.8212	0.5980
retain	pig	0.8913	0.4662
retain	ferret	0.7746	0.6715
retain	monkey	0.8288	0.5851
retain	goat	0.8431	0.5602
retain	sheep	0.8247	0.5921
retain	cat	0.9122	0.4190

Table 3: An example of IP category from *CorefConcept* Dataset.

category	Example	type	Train Samples	Train Certainty	Testing Samples	Test Certainty
IP	Katniss Everdeen	coref	"Mockingjay symbol", "District 12 tribute", "bow-and-arrow warrior", "arena survivor", "rebel leader in Panem", "Capitols outspoken challenger", "Suzanne Collins protagonist", "winner of the 74th Hunger Games", "Peeta Mellarks ally", "Girl on Fire archer"	"Very High", "High", "High", "High", "High", "Normal", "Normal", "Normal", "Normal", "Low"	"the Hunger Games heroine", "revolutionary youth icon", "symbol of defiance", "sibling protector to Prim", "victor turned rebel figure"	"Very High", "Low", "Low", "Low", "Very Low"
		retain	"Bella Swan (Twilight)", "Hermione Granger", "Clary Fray (Mortal Instruments)", "Rose Hathaway (Vampire Academy)", "Lena Duchannes (Beautiful Creatures)", "Feyre (A Court of Thorns and Roses)", "Scarlet Benoit (Lunar Chronicles)", "Celaena Sardothien (Throne of Glass)", "Alina Starkov (Shadow and Bone)", "Kelsea Glynn (Queen of the Tearling)"	"High", "High", "High", "High", "High", "Normal", "Normal", "Normal", "Normal", "Low"	"Tris Prior (Divergent)", "Laia (An Ember in the Ashes)", "Karou (Daughter of Smoke and Bone)", "Tessa Gray (Infernal Devices)", "Arya Stark (Game of Thrones)"	"High", "Low", "Low", "Low", "Very Low"

Table 4: An example of Object category from *CorefConcept* Dataset.

Category	Example	Type	Train Samples	Train Certainty	Testing Samples	Test Certainty
Object	Horse	coref	"mare", "stallion", "colt", "filly", "equine", "pony", "racehorse", "draft horse", "steed", "war horse"	"Very High", "Very High", "High", "High", "High", "Normal", "Normal", "Normal", "Low", "Low"	"domestic horse", "thoroughbred", "Arabian horse", "light riding horse", "wild horse"	"Very High", "Low", "Low", "Low", "Very Low"
		retain	"mule", "zebra", "onager", "moose", "bull", "yak", "water buffalo", "goat", "sheep", "reindeer"	"High", "High", "Normal", "Normal", "Normal", "Low", "Low", "Low", "Very Low", "Very Low"	"donkey", "llama", "alpaca", "hippopotamus", "giraffe"	"Very High", "Very Low", "Very Low", "Very Low", "Very Low"
	Bat (animal)	coref	"chiropteran", "nocturnal bat", "fruit bat", "insectivorous bat", "vampire bat", "megabat", "microbat", "bat colony creature", "cave-dwelling bat", "winged nocturnal mammal"	"Very High", "Very High", "High", "High", "High", "Normal", "Normal", "Normal", "Low", "Low"	"flying mammal", "guano-producing bat", "flying fox", "bat-like mammal", "archaic chiroptera"	"Very High", "Low", "Low", "Low", "Very Low"
		retain	"sugar glider", "owl", "bird", "flying fish", "moth", "butterfly", "colugo", "pterosaur", "dragon", "fairy"	"High", "High", "Normal", "Normal", "Normal", "Normal", "Low", "Low", "Low", "Low"	"flying squirrel", "angel", "raccoon", "rat", "fox"	"Very High", "Low", "Very Low", "Very Low", "Very Low"

Table 5: An example of Celebrity category from CorefConcept Dataset.

category	Example	type	Train Samples	Train Certainty	Testing Samples	Test Certainty
Celebrity	Tom Cruise	coref	"Mission Impossible lead", "Ethan Hunt", "Maverick the pilot", "Minority Report Lead", "Lestat Interview with the Vampire", "Jerry Maguire lead", "Scientology adherent", "Rain Man's Brother", "box-office megastar", "Nicole Kidman's Former Spouse"	"Very High", "High", "High", "High", "High", "Normal", "Normal", "Normal", "Normal", "Low"	"Top Gun pilot actor", "charismatic screen presence", "producer and actor duo", "international film sensation", "dramatic roles veteran"	"Very High", "Low", "Low", "Low", "Very Low"
		retain	"Keanu Reeves", "Leonardo DiCaprio", "Johnny Depp", "Matt Damon", "Ben Affleck", "George Clooney", "Ryan Gosling", "Christian Bale", "Hugh Jackman", "Orlando Bloom"	"High", "High", "High", "Normal", "Normal", "Normal", "Normal", "Normal", "Low", "Low"	"Brad Pitt", "Robert Pattinson", "Jake Gyllenhaal", "Chris Pratt", "Mark Wahlberg"	"High", "Low", "Low", "Low", "Very Low"

Table 6: Ablation study comparing different certainty perturbation strategies. Reds are the best and blues are the worst. The top 3 is marked in **bold**, underlined, and *italics*, respectively.

Method	$Acc_{U} \downarrow$	$Acc_{C}^{train} \downarrow$	$Acc_{C}^{test} \downarrow$	$Acc_{R}^{train} \uparrow$	$Acc_{R}^{test} \uparrow$
CRCE-nocert	2.20	19.50	36.00	86.27	79.78
CRCE-both-0.2	3.89	18.84	31.76	85.92	78.88
CRCE-both-0.4	4.17	*20.12*	32.58	86.35	79.76
CRCE-coref-0	4.34	19.43	34.36	*88.32*	**82.29**
CRCE-coref-0.2	4.31	**21.60**	**36.60**	88.60	*81.87*
CRCE-coref-0.4	*3.75*	*20.87*	*35.49*	**89.19**	82.00
CRCE-retain-0	**5.06**	13.72	*27.53*	**83.46**	76.59
CRCE-retain-0.2	4.79	15.67	29.12	*84.27*	*76.93*
CRCE-retain-0.4	*4.65*	*14.49*	26.63	83.95	**76.15**
CRCE-ours	**1.25**	**12.81**	**26.31**	85.20	79.56

F.2 Ablation: How many corefs and retains are optimal?

To investigate optimal hyperparameter values, we conduct experiments to perform a thorough sweep. We select values in range $\{1,3,5,10\}$ for the number of corefs (M) and retains (N). Table. 7 shows the results.

Contrary to what intuition might suggest, the largest values for M and N do not yield the optimal performance. While $M = 10, N = 1$ achieves the lowest Acc_{U} score (1.11%), indicating excellent target concept erasure, this configuration fails to maintain balanced performance across other metrics. Instead, configurations with mid-range values; $M = 5, N = 3$ demonstrate superior performance on coref erasure generalisation ($Acc_{C}^{test} = 26.31\%$) while maintaining strong target concept removal.

As M and N increase, the model encounters more potentially conflicting signals between what should be erased and what should be preserved. The table shows that configurations with large values ($M = 10, N = 10$) often suffer from degraded coref erasure performance ($Acc_{C}^{test} = 54.20\%$), suggesting the model struggles to establish clear conceptual boundaries when presented with too many corefs/retains. This counterintuitive finding can be explained by the concept of semantic interference. Furthermore, the retention metrics (Acc_{R}^{train} and Acc_{R}^{test}) reach their peak at $M = 5, N = 5$ (90.19% and 87.36% respectively), slightly outperforming larger configurations like $M = 10, N = 10$ (88.92% and 86.73%). We conjecture that beyond a certain threshold, additional examples create diminishing returns and can even become counterproductive by introducing instability into the learning process. The optimal configuration appears to provide sufficient corefs/retains to define the concept manifold without overwhelming the model with redundant or potentially contradictory information.

G Limitations

Despite its overall effectiveness, Fig. 5 indicates that CRCE occasionally over-erases concepts. For example, when attempting to erase the concept "joker," the post-erasure model fails to generate "Batman." This issue likely stems from intrinsic biases within the target model (SD): when erasing a coref such as "Gotham City Antagonist," the strong conceptual binding between "Gotham City" and "Batman" causes unintended erasure of "Batman." Our

Table 7: Evaluation metrics for different M and N values for the number of corefs and retains, respectively. Best scores are highlighted in bold.

	$M \backslash N$	1	3	5	10
$Acc_{U} \downarrow$	1	3.33	2.22	2.22	4.44
	3	2.40	3.33	3.20	3.20
	5	4.80	1.25	4.44	2.22
	10	**1.11**	4.44	1.17	3.33
$Acc_{C}^{train} \downarrow$	1	49.30	19.4	20.79	21.38
	3	20.85	44.72	22.55	22.36
	5	21.04	**12.81**	49.80	19.40
	10	19.60	19.90	19.47	49.30
$Acc_{C}^{test} \downarrow$	1	55.80	34.20	34.40	35.00
	3	37.20	48.3	37.86	38.13
	5	37.86	**26.31**	55.20	31.40
	10	34.40	34.40	33.26	54.20
$Acc_{R}^{train} \uparrow$	1	79.01	86.76	87.54	87.05
	3	87.89	82.61	88.15	87.76
	5	88.75	85.20	**90.19**	87.05
	10	87.35	86.47	86.01	88.92
$Acc_{R}^{test} \uparrow$	1	55.8	79.15	79.57	81.47
	3	81.66	73.62	81.52	80.97
	5	81.38	79.56	**87.36**	80.42
	10	78.94	79.78	78.68	86.73

CRCE method, even with a reasonably generated set of coref and retained terms, currently cannot fully overcome these inherent model biases.

Figure 5: This figure demonstrates how SD v1.4 incorrectly overfits the concept of "Gotham City" with "The Batman". While "Gotham City Antagonist" is a valid coreference for "The Joker", erasing "The Joker" also distorts "The Batman", revealing implicit model biases from the internal T2I model.

H Additional Qualitative Results

We provide additional visual comparisons demonstrating our method (CRCE) on object, IP, and celebrity erasure tasks. Fig. 6 illustrates object erasure, showing CRCE effectively removes "Horse" along with corefs "Pony" and "War Horse", while retaining semantically distinct concepts "Mule" and "Sheep". Fig. 7 presents the erasure of the "Apple (fruit)"

object, highlighting our method's ability to erase closely related corefs "Golden Delicious" and "Granny Smith" without impacting retains "Pear" and "Banana". Fig. 8 and Fig. 9 extend evaluations to IP and celebrities, respectively, showing that CRCE removes targeted concepts "Batman" and "Beyoncé" and their corefs without affecting visually or conceptually related retains.

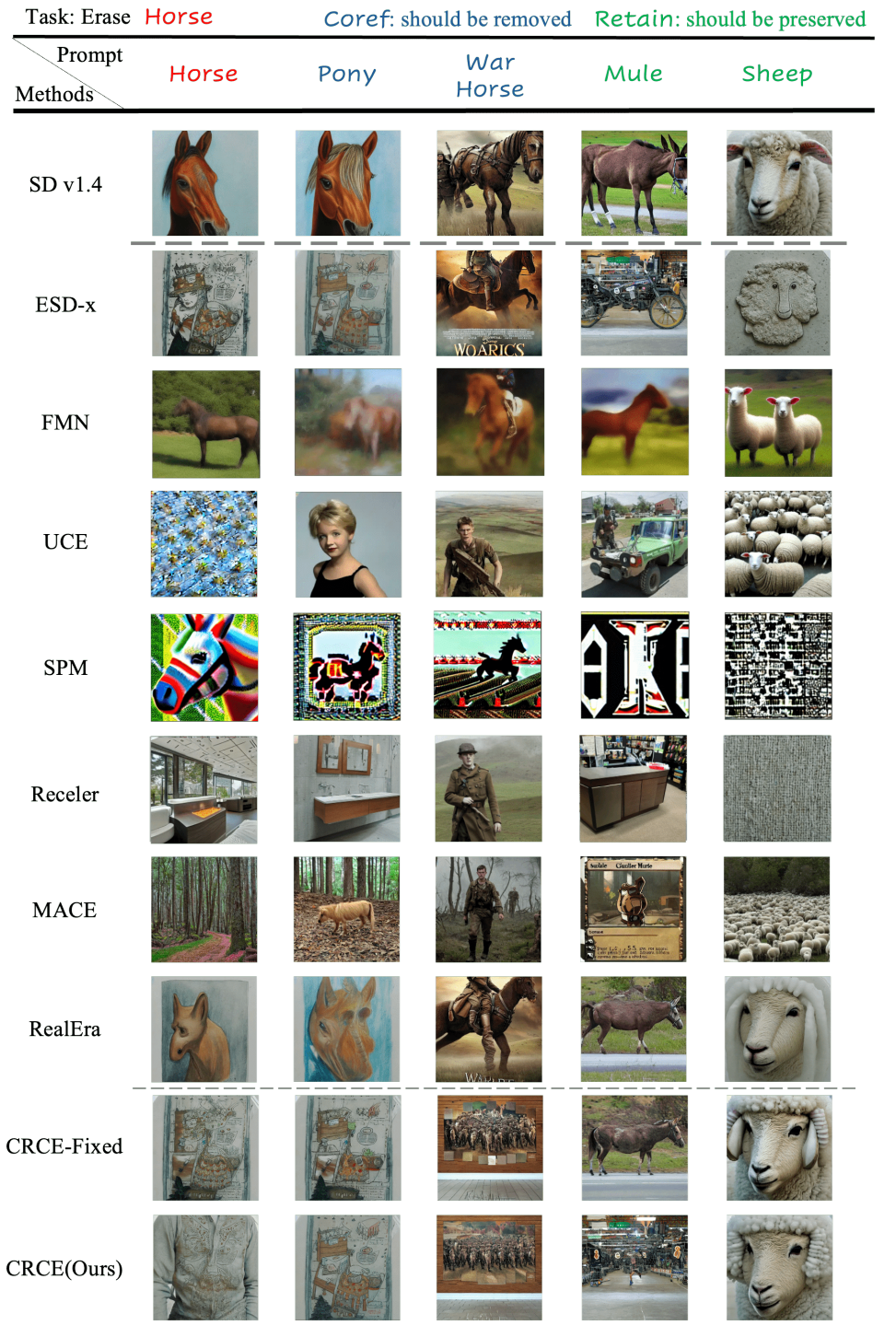

Figure 6: Qualitative Comparison on object: **Horse** erasure, together with its corefs *pony, war horse* and retains *mule, sheep*.

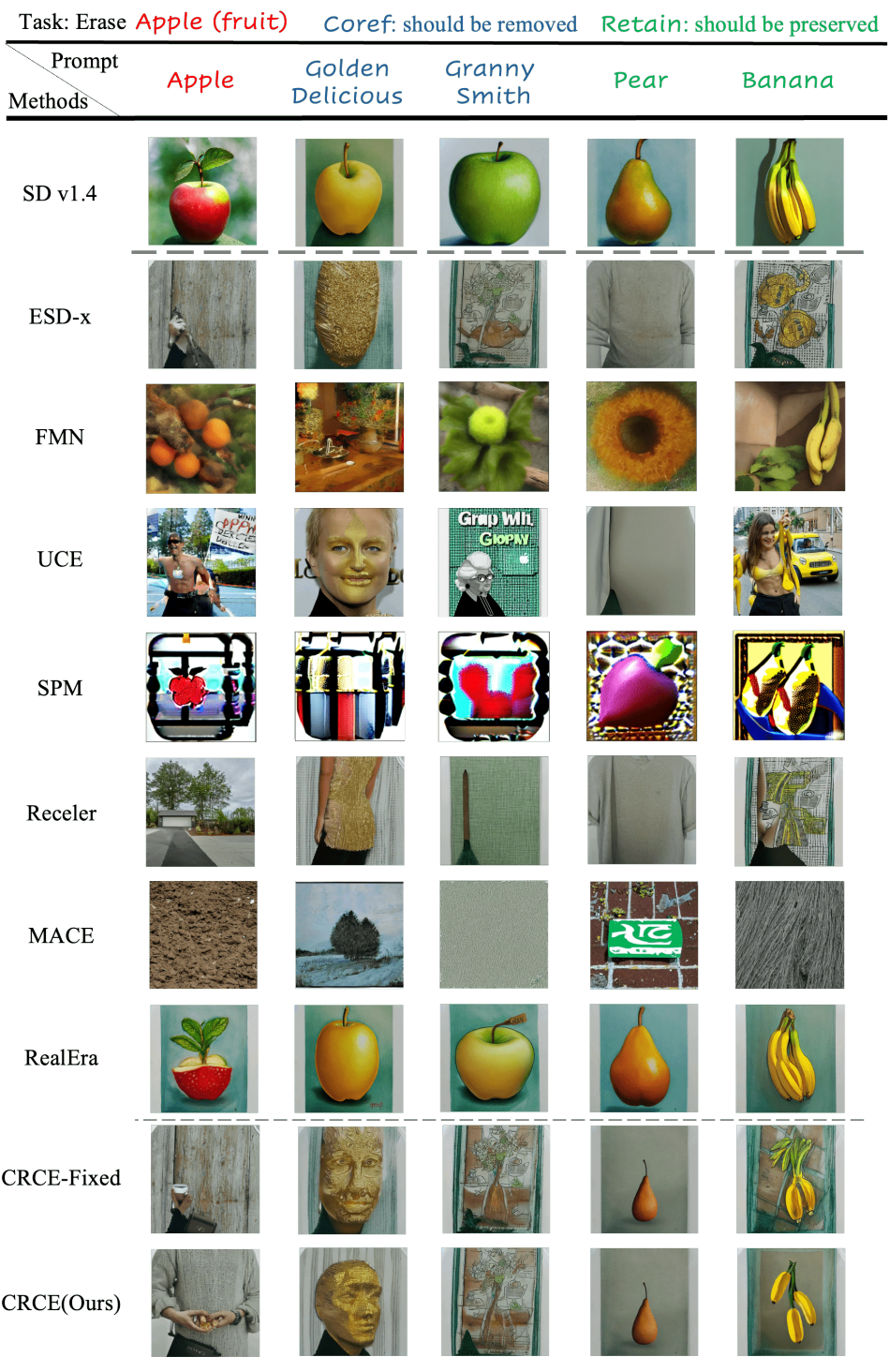

Figure 7: Qualitative Comparison on object: **Apple** erasure, together with its corefs *Golden Delicious, Granny Smith* and retains *Pear, Banana*.

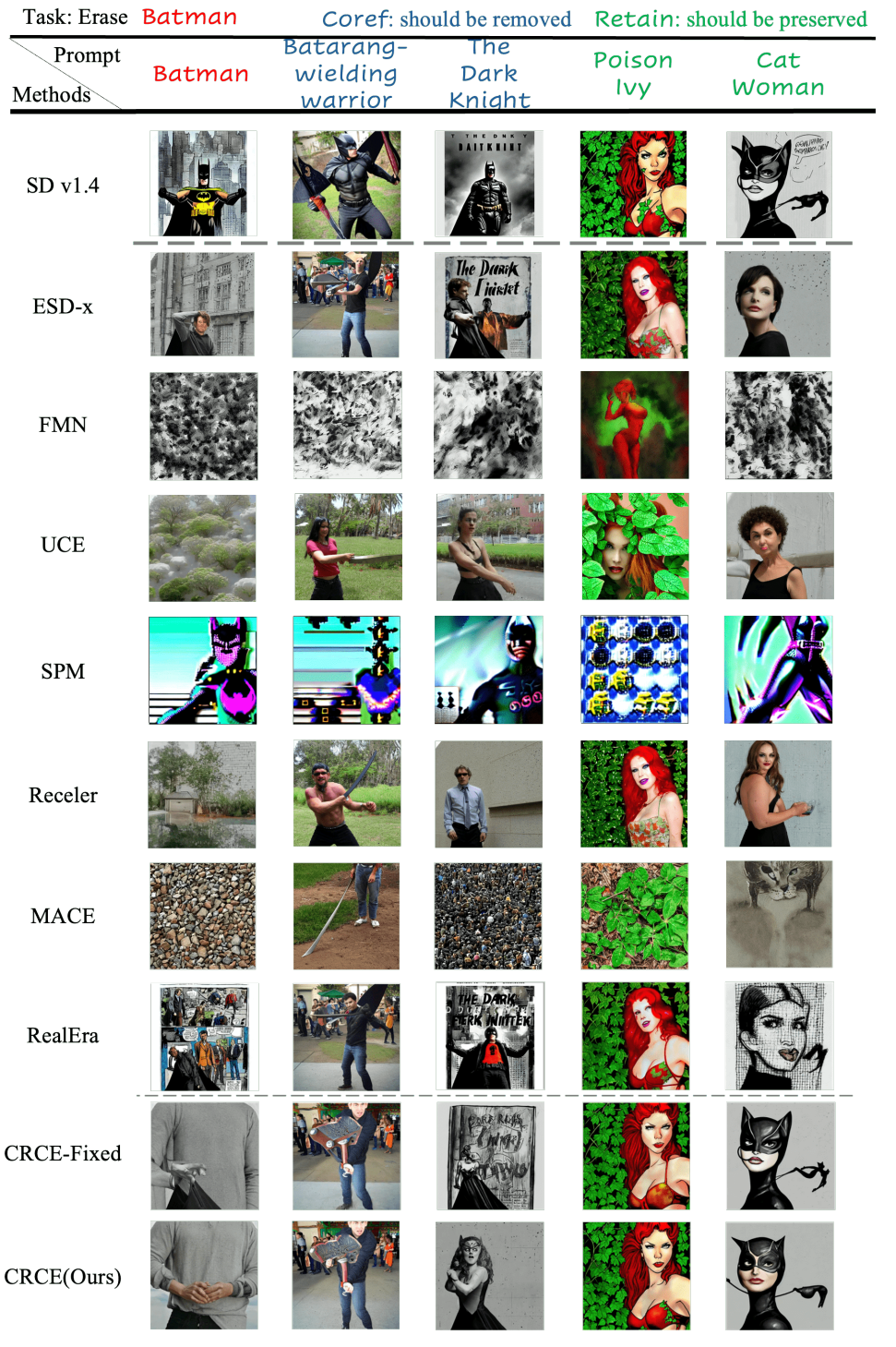

Figure 8: Qualitative Comparison on IP: **Batman** erasure, together with its corefs *Batarang-wielding warrior, The Dark Knight* and retains *Poison Ivy, Catwoman*.

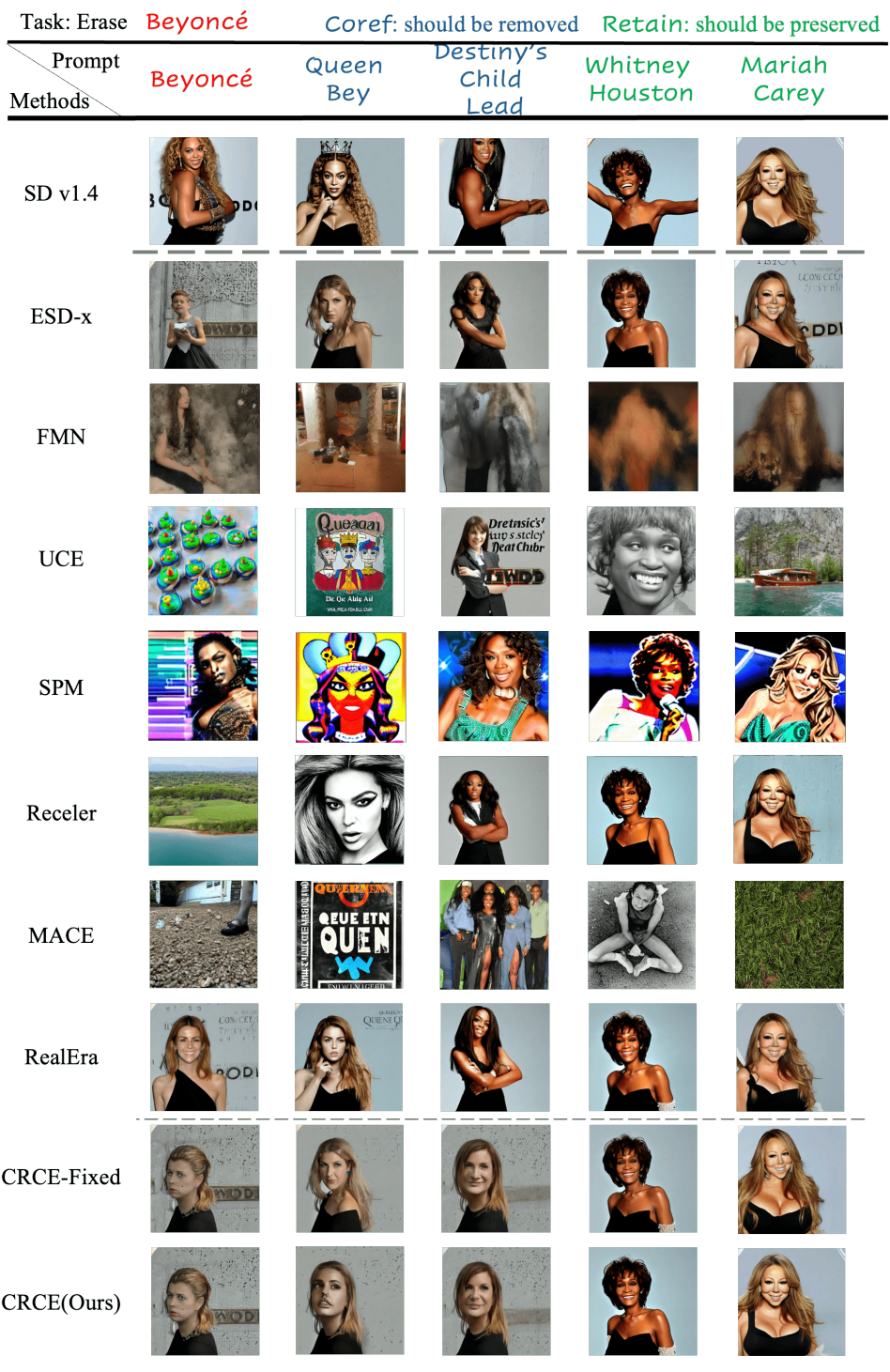

Figure 9: Qualitative Comparison on Celebrity: **Beyoncé** erasure, together with its corefs *Queen Bey, Destiny's Child Lead* and retains *Whitney Houston, Mariah Carey*.