MapRoute: Semantic Routing for Precise Concept Erasure with Mapper

Supplementary Material

Appendix

(A).1. Surrogate Concept Sensitivity Test

To verify that the model is not sensitive to the choice of surrogate concept, we train the Mapper to learn the mapping from "A photo of Cameron Diaz" and "An image of Cameron Diaz" to four different surrogate concepts: "truck" and "truck", "A photo of Andrew Garfield" and "An image of Andrew Garfield", "A photo of Bruce Lee" and "An image of Bruce Lee", and "Akos Major style" and "Akos Major style". Each experiment is trained for 100 epochs. The results are shown in Table5.

Table 5. The results of the surrogate concept sensitivity test.

Surrogate Concept	Performance Metrics		
	$ACC_e \downarrow$	$ACC_s \uparrow$	$H_c \uparrow$
`Akos Major style`	**0.00**	90.80	**95.18**
`Andrew Garfield`	0.80	**91.28**	95.08
`Bruce Lee`	**0.00**	90.24	94.87
`truck`	**0.00**	90.44	94.98

Table 6. The results of 100 artistic style concepts erasure.

Method	Performance Metrics		
	$CLIP_e \downarrow$	$CLIP_s \uparrow$	$H_a \uparrow$
`UCE`	21.14	25.01	3.87
`MACE`	**17.45**	23.13	5.68
`Ours`	18.41	**25.71**	**7.30**

(A).2. Training Configurations

Object Erasure: To evaluate the erasure capability of MapRoute. For each object concept and synonym concept, we generate 200 images using the prompt "A photo of the concept" to compute the ACC score. The settings for object concepts and subclass concepts are provided in Table 11, while the synonym concepts follow the same configuration as in Mace[19]. Each experiment is trained for 500 epochs.

Celebrity Erasure: First, we randomly selected two celebrities, Adam Driver and Adriana Lima, to evaluate the erasure performance for individual celebrity concepts. Subsequently, following the setup in Mace, we conducted a collective erasure experiment involving 50 celebrities, with the erased and retained celebrity names listed in Table 12. We use five text prompts to generate images: "a portrait of {celebrity name}", "a sketch of {celebrity name}", "an oil painting of {celebrity name}", "{celebrity name} in an official photo", and "an image capturing {celebrity name} at a public event". In the single-celebrity erasure experiment, for the Erasure Group, we generate 50 images per celebrity per prompt, resulting in a total of 2,500 images. For the Retention Group, we generate 5 images per celebrity per prompt, also totaling 2,500 images. In the 50-celebrity erasure experiment, for the Erasure Group, we generate 5 images per celebrity per prompt, yielding 1,250 images in total. For the Retention Group, we generate 5 images per celebrity per prompt, resulting in 2,500 images. All other baselines follow the same experimental configuration. Portraits of these celebrities can be effectively generated using SD v1.4. The generated portraits are accurately recognizable by top-1 accuracy of the GIPHY Celebrity Detector (GCD). Each experiment is trained for 100 epochs.

Artistic Style Erasure: To evaluate the effectiveness of MapRoute in artistic style erasure, we first conducted single-artist erasure experiments using three artists: Brent Heighton, Brett Weston, and Brett Whiteley. For each, we trained the Mapper to learn the mapping from the style of artist name to the style of Van Gogh.

For the 100-artist erasure experiment, we followed the setup in Mace, selecting 100 artists as the erasure group and another 100 as the retention group, and trained the Mapper to map from each artist name style to the Van Gogh style.

Otherwise, to evaluate the accuracy of the results, for all artistic style erasure experiments, we use the CLIP-ViT-Large-Patch14 model to compute CLIP scores, rather than the CLIP-ViT-Base-Patch32 model used in Mace.

After training, we generated images for both groups using the same random seeds and the following five prompts: "Image in the style of artist name", "Art inspired by artist name", "Painting in the style of artist name", "A reproduction of art by artist name", and "A famous artwork by artist name". For the single-artist erasure experiment, we generated 50 images per prompt for the erasure group and 5 images per prompt for the retention group, resulting in a total of 2,750 images. For the 100-artist erasure experiment, both the erasure and retention groups generated 5 images per prompt, yielding 2,500 images per group. The complete lists of artists in the erasure and retention groups are provided in Table 13. Each experiment is trained for 100 epochs.

(A).3. Additional Ablation Studies

To further evaluate the model's optimal performance, we conduct ablation studies on the number of epochs in Stage 1

and the total training epochs for both celebrity erasure and artistic style erasure tasks. Based on the results, we ultimately select 10 epochs for Stage 1 and 100 epochs in total. Moreover, considering that the object erasure task removes entire class concepts, a more comprehensive set of subclass examples leads to better performance; under our experimental setup, training for 500 epochs in total yields the optimal results.

As shown in Table 7, when the number of training epochs in Stage 1 is set to 1 or 5, the model achieves very low $\mathrm{GCD_s}$ and $\mathrm{CLIP_s}$ values, indicating that it fails to learn the identity mapping in the concept embedding space. When the total number of training epochs is only 50, the $\mathrm{GCD_e}$ and $\mathrm{CLIP_e}$ values remain very high, suggesting that the model has not yet learned the one-to-one mapping from the target concept to the surrogate concept. However, when the number of epochs in Stage 1 is increased to 15, there is no significant improvement in performance. Therefore, we choose 10 epochs for Stage 1. Based on this setting, further increasing the total number of epochs to 200 or 300 does not yield noticeable performance gains. Hence, we select a total of 100 training epochs as the final configuration.

Table 7. For the erasure performance on individual celebrity and artistic style concepts, the GCD score is the average over the Adam Driver and Adriana Lima experiments, while the CLIP score is the average over the Brent Heighton, Brett Weston, and Brett Whiteley experiments. To evaluate the effectiveness of Stage 1 in learning the identity mapping, we compute only the $\mathrm{GCD_s}$ and $\mathrm{CLIP_s}$ values.

Epochs	Celebrity Erasure		Artistic style Erasure	
	$\mathrm{GCD_e}\downarrow$	$\mathrm{GCD_s}\uparrow$	$\mathrm{CLIP_e}\downarrow$	$\mathrm{CLIP_s}\uparrow$
1(stage1)	–	0.00	–	0.00
5(stage1)	–	23.12	–	15.02
10(stage1)	–	91.16	–	26.11
15(stage1)	–	91.16	–	26.12
50	89.31	91.32	25.18	26.09
100	0.00	90.92	16.85	26.13
200	0.00	91.16	16.80	26.14
300	0.00	91.14	16.82	26.13

Finally, we conducted an ablation study on the choice of Top-k using the COCO-30k dataset, while concurrently erasing the 100 art style listed in the 'Erasure Group' of Table 13. The results indicate that the model maintains robust generative performance when Top-k is below 30. This finding is consistent with the inherent capabilities of current generative models: once a prompt exceeds 30 distinct concepts, these models typically fail to generate the corresponding images accurately.The result is shown in Table 9.

Table 8. Ablation experiments on hyperparameter Top-k (COCO-30k).

Top-k	**1**	**5**	**10**	**30**	**50**
LPIPS	0.013	0.014	0.12	0.31	0.91

(A).4. Additional Experiments

The erasure performance on 100 artistic style concepts is presented in Table 6.

Following the methodology of UnlearnDiffAtk, we evaluated the erasure performance of the baseline models and MapRoute on the artistic style erasure task, specifically targeting Van Gogh style under white-box attacks. The results demonstrate that our model maintains superior generative performance even when subjected to such attacks.The result is shown in Table 10.

Table 9. Performance evaluation across different numbers of concepts.

# LPIPS	UCE	MapRoute
1	0.05	0.014
5	0.08	0.014
10	0.13	0.013
50	0.22	0.015
100	30.08	0.015

Furthermore, we evaluated the quality of the generated images. Following the experimental procedure of UCE, we configured MapRoute with a $Top\text{-}k$ value of 1. Utilizing the dataset of 1,000 artists provided therein, we analyzed the method's capability to erase multiple concepts by erasing n target artists while preserving the remaining $1,000 - n$ artists.The result is shown in Table 9.

Table 10. Attack performance of various methods against unlearned DMs by UnlearnDiffAtk[42] in Van Gogh style, measured by ASR averaged over perturbing 50 Van Goghrelated prompts.

Artistic Style:	**Van Gogh**									
Unlearned DMs:	ESD		FMN		AC		UCE		MapRoute	
	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3
No Attack	2.00%	16.00%	10.00%	32.00%	12.00%	52.00%	62.00%	78.00%	**2.00%**	**2.00%**
UnlearnDiffAtk	32.00%	76.00%	56.00%	90.00%	77.00%	92.00%	94.00%	100.00%	**10.00%**	**12.00%**

Table 11. We train the Mapper to learn the mapping from object concepts and their subclass concepts to the surrogate concept.

Object concept	subclass concept	surrogate concept
'airplane'	*'plane' 'jet' 'aircraft' 'private jet' 'flying' 'spacecraft' 'helicopter' 'drone' 'business jet'*	'vehicle'
'automobile'	*'car' 'automotive' 'vehicle' 'motorcar' 'sedan' 'SUV' 'Electric car' 'drone' 'Sports car'*	'Portrait of Van Gogh'
'bird'	*'birdie' 'chicken' 'avian' 'fowl' 'winged creature' 'pigeon' 'parrot' 'swan' 'flamingo'*	'cat'
'cat'	*'cats' 'kitty' 'gato' 'tabby cat' 'kitten' 'ragdoll' 'bengal' 'russian blue' 'siamese'*	'dog'
'deer'	*'muntjac' 'moose' 'barasingha' 'mule deer' 'reindeer' 'stag' 'doe' 'hart' 'red deer'*	'horse'
'dog'	*'poodle' 'chihuahua' 'hound' 'bulldog' 'doggy' 'puppy' 'dogs' 'canine' 'pooch'*	'bird'
'frog'	*'anuran' 'amphibian' 'tadpole' 'tree frog' 'marsh frog' 'toad' 'bullfrog' 'poison dart frog' 'horned frog'*	'cat'
'horse'	*'arabian horse' 'andalusian' 'shire horse' 'arabian horse' 'icelandic horse' 'equine' 'steed' 'mare' 'pony'*	'car'
'ship'	*'sailing' 'boat' 'pirate ship' 'sail boat' 'vessel' 'watercraft' 'Aircraft carrier' 'Tugboat' 'Hovercraft'*	'vehicle'
'truck'	*'Semi-trailer truck' 'Dump truck' 'Tanker truck' 'vehicle' 'Fire truck' 'lorry' 'rig' 'hauler' 'truck'*	'cat'

Table 12. We train the Mapper to learn the mapping between the textual embeddings of "A photo of {erased celebrity}" and "An image of {erased celebrity}" and the corresponding embeddings of "A photo of Bruce Lee" and "An image of Bruce Lee." Subsequently, we generate images following Mace[19] generation configuration and evaluate the results using quantitative metrics, as illustrated in the figure.

Group	# of Celebrities to Be Erased	Surrogate Concept	Celebrity
Erasure Group	50	'Bruce Lee'	*'Adam Driver', 'Adriana Lima', 'Amber Heard', 'Amy Adams', 'Andrew Garfield', 'Angelina Jolie', 'Anjelica Huston', 'Anna Faris', 'Anna Kendrick', 'Anne Hathaway', 'Arnold Schwarzenegger', 'Barack Obama', 'Beth Behrs', 'Bill Clinton', 'Bob Dylan', 'Bob Marley', 'Bradley Cooper', 'Bruce Willis', 'Bryan Cranston', 'Cameron Diaz', 'Channing Tatum', 'Charlie Sheen', 'Charlize Theron', 'Chris Evans', 'Chris Hemsworth', 'Chris Pine', 'Chuck Norris', 'Courteney Cox', 'Demi Lovato', 'Drake', 'Drew Barrymore', 'Dwayne Johnson', 'Ed Sheeran', 'Elon Musk', 'Elvis Presley', 'Emma Stone', 'Frida Kahlo', 'George Clooney', 'Glenn Close', 'Gwyneth Paltrow', 'Harrison Ford', 'Hillary Clinton', 'Hugh Jackman', 'Idris Elba', 'Jake Gyllenhaal', 'James Franco', 'Jared Leto', 'Jason Momoa', 'Jennifer Aniston', 'Jennifer Lawrence'*
Retention Group	100	–	*'Aaron Paul', 'Alec Baldwin', 'Amanda Seyfried', 'Amy Poehler', 'Amy Schumer', 'Amy Winehouse', 'Andy Samberg', 'Aretha Franklin', 'Avril Lavigne', 'Aziz Ansari', 'Barry Manilow', 'Ben Affleck', 'Ben Stiller', 'Benicio Del Toro', 'Bette Midler', 'Betty White', 'Bill Murray', 'Bill Nye', 'Britney Spears', 'Brittany Snow', 'Bruce Lee', 'Burt Reynolds', 'Charles Manson', 'Christie Brinkley', 'Christina Hendricks', 'Clint Eastwood', 'Countess Vaughn', 'Dakota Johnson', 'Dane Dehaan', 'David Bowie', 'David Tennant', 'Denise Richards', 'Doris Day', 'Dr Dre', 'Elizabeth Taylor', 'Emma Roberts', 'Fred Rogers', 'Gal Gadot', 'George Bush', 'George Takei', 'Gillian Anderson', 'Gordon Ramsey', 'Halle Berry', 'Harry Dean Stanton', 'Harry Styles', 'Hayley Atwell', 'Heath Ledger', 'Henry Cavill', 'Jackie Chan', 'Jada Pinkett Smith', 'James Garner', 'Jason Statham', 'Jeff Bridges', 'Jennifer Connelly', 'Jensen Ackles', 'Jim Morrison', 'Jimmy Carter', 'Joan Rivers', 'John Lennon', 'Johnny Cash', 'Jon Hamm', 'Judy Garland', 'Julianne Moore', 'Justin Bieber', 'Kaley Cuoco', 'Kate Upton', 'Keanu Reeves', 'Kim Jong Un', 'Kirsten Dunst', 'Kristen Stewart', 'Krysten Ritter', 'Lana Del Rey', 'Leslie Jones', 'Lily Collins', 'Lindsay Lohan', 'Liv Tyler', 'Lizzy Caplan', 'Maggie Gyllenhaal', 'Matt Damon', 'Matt Smith', 'Matthew Mcconaughey', 'Maya Angelou', 'Megan Fox', 'Mel Gibson', 'Melanie Griffith', 'Michael Cera', 'Michael Ealy', 'Natalie Portman', 'Neil Degrasse Tyson', 'Niall Horan', 'Patrick Stewart', 'Paul Rudd', 'Paul Wesley', 'Pierce Brosnan', 'Prince', 'Queen Elizabeth', 'Rachel Dratch', 'Rachel Mcadams', 'Reba Mcentire', 'Robert De Niro'*

Table 13. We train the Mapper to learn the mapping between the textual embeddings of "{erased artistic} style" and the corresponding embeddings of "{Van Gogh} style". Subsequently, we generate images following [19] generation configuration and evaluate the results using quantitative metrics, as illustrated in the figure.

Group	# of artistic styles to Be Erased	Surrogate Concept	artistic style
Erasure Group	100	Van Gogh	*'Brent Heighton', 'Brett Weston', 'Brett Whiteley', 'Brian Bolland', 'Brian Despain', 'Brian Froud', 'Brian K. Vaughan', 'Brian Kesinger', 'Brian Mashburn', 'Brian Oldham', 'Brian Stelfreeze', 'Brian Sum', 'Briana Mora', 'Brice Marden', 'Bridget Bate Tichenor', 'Briton Rivie're', 'Brooke Didonato', 'Brooke Shaden', 'Brothers Grimm', 'Brothers Hildebrandt', 'Bruce Munro', 'Bruce Nauman', 'Bruce Pennington', 'Bruce Timm', 'Bruno Catalano', 'Bruno Munari', 'Bruno Walpoth', 'Bryan Hitch', 'Butcher Billy', 'C. R. W. Nevinson', 'Cagnaccio Di San Pietro', 'Camille Corot', 'Camille Pissarro', 'Camille Walala', 'Canaletto', 'Candido Portinari', 'Carel Willink', 'Carl Barks', 'Carl Gustav Carus', 'Carl Holsoe', 'Carl Larsson', 'Carl Spitzweg', 'Carlo Crivelli', 'Carlos Schwabe', 'Carmen Saldana', 'Carne Griffths', 'Casey Weldon', 'Caspar David Friedrich', 'Cassius Marcellus Coolidge', 'Catrin Welz-Stein', 'Cedric Peyravernay', 'Chad Knight', 'Chantal Jofe', 'Charles Addams', 'Charles Angrand', 'Charles Blackman', 'Charles Camoin', 'Charles Dana Gibson', 'Charles E. Burchffeld', 'Charles Gwathmey', 'Charles Le Brun', 'Charles Liu', 'Charles Schridde', 'Charles Schulz', 'Charles Spencelayh', 'Charles Vess', 'Charles-Francois Daubigny', 'Charlie Bowater', 'Charline Von Heyl', 'Chaim Soutine', 'Chen Zhen', 'Chesley Bonestell', 'Chiharu Shiota', 'Ching Yeh', 'Chip Zdarsky', 'Chris Claremont', 'Chris Cunningham', 'Chris Foss', 'Chris Leib', 'Chris Moore', 'Chris Offli', 'Chris Saunders', 'Chris Turnham', 'Chris Uminga', 'Chris Van Allsburg', 'Chris Ware', 'Christian Dimitrov', 'Christian Grajewski', 'Christophe Vacher', 'Christopher Balaskas', 'Christopher Jin Baron', 'Chuck Close', 'Cicely Mary Barker', 'Cindy Sherman', 'Clara Miller Burd', 'Clara Peeters', 'Clarence Holbrook Carter', 'Claude Cahun', 'Claude Monet', 'Clemens Ascher'*
Retention Group	100	–	*'A.J. Casson', 'Aaron Douglas', 'Aaron Horkey', 'Aaron Jasinski', 'Aaron Siskind', 'Abbott Fuller Graves', 'Abbott Handerson Thayer', 'Abdel Hadi Al Gazzar', 'Abed Abdi', 'Abigail Larson', 'Abraham Mintchine', 'Abraham Pether', 'Abram Effmovich Arkhipov', 'Adam Elsheimer', 'Adam Hughes', 'Adam Martinakis', 'Adam Paquette', 'Adi Granov', 'Adolf Hire 'my-Hirschl', 'Adolph Gottlieb', 'Adolph Menzel', 'Adonna Khare', 'Adriaen van Ostade', 'Adriaen van Outrecht', 'Adrian Donoghue', 'Adrian Ghenie', 'Adrian Paul Allinson', 'Adrian Smith', 'Adrian Tomine', 'Adrianus Eversen', 'Afarin Sajedi', 'Af andi', 'Aggi Erguna', 'Agnes Cecile', 'Agnes Lawrence Pelton', 'Agnes Martin', 'Agostino Arrivabene', 'Agostino Tassi', 'Ai Weiwei', 'Ai Yazawa', 'Akihiko Yoshida', 'Akira Toriyama', 'Akos Major', 'Akseli Gallen-Kallela', 'Al Capp', 'Al Feldstein', 'Al Williamson', 'Alain Laboile', 'Alan Bean', 'Alan Davis', 'Alan Kenny', 'Alan Lee', 'Alan Moore', 'Alan Parry', 'Alan Schaller', 'Alasdair McLellan', 'Alastair Magnaldo', 'Alayna Lemmer', 'Albert Benois', 'Albert Bierstadt', 'Albert Bloch', 'Albert Dubois-Pillet', 'Albert Eckhout', 'Albert Edelfelt', 'Albert Gleizes', 'Albert Goodwin', 'Albert Joseph Moore', 'Albert Koetsier', 'Albert Kotin', 'Albert Lynch', 'Albert Marquet', 'Albert Pinkham Ryder', 'Albert Robida', 'Albert Servaes', 'Albert Tucker', 'Albert Watson', 'Alberto Biasi', 'Alberto Burri', 'Alberto Giacometti', 'Alberto Magnelli', 'Alberto Seveso', 'Alberto Sughi', 'Alberto Vargas', 'Albrecht Anker', 'Albrecht Dürer', 'Alec Soth', 'Alejandro Burdisio', 'Alejandro Jodorowsky', 'Aleksey Savrasov', 'Aleksi Briclot', 'Alena Aenami', 'Alessandro Allori', 'Alessandro Barbucci', 'Alessandro Gottardo', 'Alessio Albi', 'Alex Alemany', 'Alex Andreev', 'Alex Colville', 'Alex Figini', 'Alex Garant'*