Appendix

A. Preliminaries

Latent diffusion model. Our method is implemented using Stable Diffusion (SD), also known as Latent Diffusion Models (LDM) [16]. This approach conducts the diffusion process within the latent space of an autoencoder. LDM is comprised of two principal components: a vector quantization autoencoder [2, 20] and a diffusion model [1, 7, 9, 16, 18, 19]. The autoencoder undergoes pretraining to transform images into spatial latent codes via an encoder $\mathbf{z} = \mathcal{E}(\mathbf{x})$, and it can reconstruct the images from these latent codes using a decoder $\mathbf{x} \approx \mathcal{D}(\mathcal{E}(\mathbf{x}))$. The diffusion model, on the other hand, is trained to generate latent codes that exist within the autoencoder's latent space. The training objective for the diffusion model is defined as follows [7, 16]:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathbf{z}\sim\mathcal{E}(\mathbf{x}),\mathbf{c},\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right], \tag{10}$$

where $\mathbf{z}_t$ is the noisy latent, t is the timestep, $\boldsymbol{\epsilon}$ is a standard Gaussian noise sample, $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ is the denoising network, and $\mathbf{c}$ is the conditioning embeddings, which can be encoded from text prompts, class labels, segmentation masks, among others [16]. During the inference process, Gaussian noise is sampled as a starting point $\mathbf{z}_T$ and successively denoised to produce a new latent code $\mathbf{z}_0$ through the well-trained denoising network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$. Ultimately, this latent code is transformed into an image via the pretrained decoder $\mathbf{x}_0 \approx \mathcal{D}(\mathbf{z}_0)$.

Cross-attention in text-to-image diffusion models. In text-to-image diffusion models, cross-attention mechanisms serve as the pivotal interface for the interplay between image and text modalities. Initially, a text prompt undergoes tokenization, converting it into a series of unique token embeddings. These embeddings are then processed by a text encoder (e.g., CLIP [14] or T5 [15]), resulting in a final set of embeddings, denoted as $\boldsymbol{\mathcal{P}} = [\mathbf{e}_1\ \mathbf{e}_2 \cdots \mathbf{e}_y]$, wherein each token's embedding $\mathbf{e}_i$ is enriched with information from the entire token sequence. These enhanced embeddings are subsequently introduced into the cross-attention modules, where they act as navigational beacons for the image synthesis process. At certain layer l and timestep t, the text embeddings are mapped using projection matrices, $\mathbf{W}_k$ and $\mathbf{W}_v$, to obtain the 'Keys' $\mathbf{k}_{t,l}$ and 'Values' $\mathbf{v}_{t,l}$, respectively. Concurrently, the image's features, $\mathbf{f}_{t,l}$, undergo the projection $\mathbf{W}_q$ to form the 'Queries' $\mathbf{q}_{t,l}$. The cross-attention mechanism computes the attention map as [16, 21]:

$$\mathbf{A}_{t,l} = \text{softmax}\left(\frac{\mathbf{q}_{t,l} \cdot \mathbf{k}_{t,l}^{\mathsf{T}}}{\sqrt{d}}\right), \tag{11}$$

where d is the scaling factor to normalize the dot product. The module then synthesizes image features by aggregating 'Values' with the attention weights, $\mathbf{o}_{t,l} = \mathbf{A}_{t,l} \cdot \mathbf{v}_{t,l}$. This process ensures that the generated images are intricately aligned with the input text, completing the text-to-image generation with high fidelity.

B. Closed-Form Solution Proof

In this section, we present a detailed derivation of the closed-form solution as written in Eq. (2). Our goal is to determine a refined matrix, denoted as $\mathbf{W}_k' \in \mathbb{R}^{d_1 \times d_2}$, which encourages the model to refrain from embedding residual information of the target phrase into other words, while preserving the prior knowledge. The loss function is defined in Eq. (1), which is:

$$\mathcal{L}\left(\mathbf{W}_k'\right) = \sum_{i=1}^{n}\left\|\mathbf{W}_k' \cdot \mathbf{e}_i^f - \mathbf{W}_k \cdot \mathbf{e}_i^g\right\|_2^2 + \lambda_1 \sum_{i=n+1}^{n+m}\|\mathbf{W}_k' \cdot \mathbf{e}_i^p - \mathbf{W}_k \cdot \mathbf{e}_i^p\|_2^2,$$

where $\lambda_1 \in \mathbb{R}^+$ is a hyperparameter, $\mathbf{e}_i^f \in \mathbb{R}^{d_2}$ is the embedding of a word co-existing with the target phrase, $\mathbf{e}_i^g \in \mathbb{R}^{d_2}$ is the embedding of that word when the target phrase is replaced with its super-category or a generic one, $\mathbf{e}_i^p \in \mathbb{R}^{d_2}$ is the embedding for preserving the prior, $\mathbf{W}_k \in \mathbb{R}^{d_1 \times d_2}$ is the pretrained weights, and n, m are the number of embeddings for mapping and preserving, respectively.

To seek the optimal $\mathbf{W}'_k$, we differentiate the loss function with respect to it and set the derivative equal to zero:

$$\frac{\partial \mathcal{L}\left(\mathbf{W}'_k\right)}{\partial \mathbf{W}'_k} = 2\sum_{i=1}^{n}\left(\mathbf{W}'_k \cdot \mathbf{e}_i^f - \mathbf{W}_k \cdot \mathbf{e}_i^g\right)\left(\mathbf{e}_i^f\right)^\mathsf{T} + 2\lambda_1 \sum_{i=n+1}^{n+m}\left(\mathbf{W}'_k \cdot \mathbf{e}_i^p - \mathbf{W}_k \cdot \mathbf{e}_i^p\right)\left(\mathbf{e}_i^p\right)^\mathsf{T} = 0 \tag{12}$$

$$\sum_{i=1}^{n}\mathbf{W}'_k \cdot \mathbf{e}_i^f \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} - \sum_{i=1}^{n}\mathbf{W}_k \cdot \mathbf{e}_i^g \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m}\mathbf{W}'_k \cdot \mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T} - \lambda_1 \sum_{i=n+1}^{n+m}\mathbf{W}_k \cdot \mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T} = 0 \tag{13}$$

$$\sum_{i=1}^{n}\mathbf{W}'_k \cdot \mathbf{e}_i^f \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m}\mathbf{W}'_k \cdot \mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T} = \sum_{i=1}^{n}\mathbf{W}_k \cdot \mathbf{e}_i^g \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m}\mathbf{W}_k \cdot \mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T} \tag{14}$$

$$\mathbf{W}'_k\left(\sum_{i=1}^{n}\mathbf{e}_i^f \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m}\mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T}\right) = \sum_{i=1}^{n}\mathbf{W}_k \cdot \mathbf{e}_i^g \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m}\mathbf{W}_k \cdot \mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T} \tag{15}$$

$$\mathbf{W}'_k = \left(\sum_{i=1}^{n}\mathbf{W}_k \cdot \mathbf{e}_i^g \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m}\mathbf{W}_k \cdot \mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T}\right) \cdot \left(\sum_{i=1}^{n}\mathbf{e}_i^f \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m}\mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T}\right)^{-1}. \tag{16}$$

To ensure the validity of the final step, it is crucial that the symmetric real matrix $\left(\sum_{i=1}^{n}\mathbf{e}_i^f \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m}\mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T}\right)$ is full rank. For any non-zero vector $\mathbf{x} \in \mathbb{R}^{d_2}$, we examine the following quadratic form:

$$\mathbf{x}^\mathsf{T} \cdot \left(\sum_{i=1}^{n}\mathbf{e}_i^f \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m}\mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T}\right) \cdot \mathbf{x} = \sum_{i=1}^{n}\left(\mathbf{x}^\mathsf{T}\mathbf{e}_i^f\right) \cdot \left(\mathbf{x}^\mathsf{T}\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m}\left(\mathbf{x}^\mathsf{T}\mathbf{e}_i^p\right) \cdot \left(\mathbf{x}^\mathsf{T}\mathbf{e}_i^p\right)^\mathsf{T} \tag{17}$$

$$= \sum_{i=1}^{n}\left\|\mathbf{x}^\mathsf{T}\mathbf{e}_i^f\right\|_2^2 + \lambda_1 \sum_{i=n+1}^{n+m}\left\|\mathbf{x}^\mathsf{T}\mathbf{e}_i^p\right\|_2^2 \geq 0. \tag{18}$$

The prior preserving embeddings $\mathbf{e}_i^p$ are computed by default using the MS-COCO dataset [12]. Due to the extensive number of terms involved in the summation, it is highly improbable for all terms $\left\|\mathbf{x}^\mathsf{T}\mathbf{e}_i^p\right\|_2^2$ in the sum to equal zero. Hence, in general cases, the matrix $\left(\sum_{i=1}^{n}\mathbf{e}_i^f \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m}\mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T}\right)$ is positive definite and thus invertible. The derivation is applicable to $\mathbf{W}'_v$ as well.

In addition to retaining general prior knowledge, akin to UCE [4], our framework extends support to allow users to highlight and preserve domain-specific concepts. This functionality is absent in most preceding frameworks. For instance, when two concepts share a strong correlation, removing one could potentially impair the generation quality of the other, which might be intended for preservation. Both general and domain-specific prior knowledge can be incorporated into the second term of Eq. (1). We set a weighting factor λ_3 to calibrate the significance attributed to each type of knowledge. Thus, Eq. (2) can be reformulated as follows:

$$\begin{aligned}\mathbf{W}'_k = &\left(\sum_{i=1}^{n}\mathbf{W}_k \cdot \mathbf{e}_i^g \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m'}\mathbf{W}_k \cdot \mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T} + \lambda_3 \sum_{i=n+m'}^{n+m}\mathbf{W}_k \cdot \mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T}\right) \\ &\cdot \left(\sum_{i=1}^{n}\mathbf{e}_i^f \cdot \left(\mathbf{e}_i^f\right)^\mathsf{T} + \lambda_1 \sum_{i=n+1}^{n+m'}\mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T} + \lambda_3 \sum_{i=n+m'}^{n+m}\mathbf{e}_i^p \cdot \left(\mathbf{e}_i^p\right)^\mathsf{T}\right)^{-1},\end{aligned} \tag{19}$$

where we have m' terms of general knowledge and $m - m'$ terms of domain-specific knowledge.

C. Implementation Details

All results from the original SD v1.4 and SD v2.1 are obtained without the application of negative prompts.

C.1. Experimental Setup Details

Object erasure. To assess the generality of erasure, we prepare three synonyms for each of the ten object classes in the CIFAR-10 dataset [10]. These synonyms are listed in Table 5. Since the object classes lack proper super-categories, we

Table 5. The synonyms and mapping concepts for the ten object classes in the CIFAR-10 dataset.

Object Classes	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Synonyms	Aircraft	Car	Avian	Feline	Hart	Canine	Amphibian	Equine	Vessel	Lorry
	Plane	Vehicle	Fowl	Kitty	Stag	Pooch	Anuran	Steed	Boat	Rig
	Jet	Motorcar	Winged Creature	Housecat	Doe	Hound	Tadpole	Mount	Watercraft	Hauler
Mapping Concepts (Randomly Sampled)	Ground	Sky	Street	Forest	Street	Sky	Forest	Forest	Ground	Sky

allocate generic concepts (e.g., sky or ground) to them, as also presented in Table 5. To evaluate the erasure capability of different methods, we use each finetuned model to generate 200, 600, and 1,800 images for three categories: the erased object (efficacy), its three synonyms (generality), and nine remaining objects (specificity). For the erased object, the prompt is *'a photo of the {erased class name}'*. For the three synonyms of the erased object, we use the prompt *'a photo of the {synonym of erased class name}'*. Lastly, for each of the nine remaining unaltered objects, the prompt is *'a photo of the {unaltered class name}'*.

Celebrity erasure. For assessing the task of multiple celebrity erasure, we establish a dataset consisting of 200 celebrities. Their portraits generated by SD v1.4 are recognizable with exceptional accuracy by the GIPHY Celebrity Detector (GCD) [5]. This dataset is divided into two groups: an erasure group with 100 celebrities and a retention group with 100 other celebrities. The full list for both groups is presented in Table 6. For evaluating the task of celebrity erasure, we use five text prompts to generate images. These prompts are as follows: *'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'*. The detailed breakdown of the number of images generated for each experiment is presented in Table 7.

Explicit content erasure. We adopt the same setting used in SA [6] to erase 'nudity', 'naked', 'erotic', and 'sextual' from SD v1.4. The mapping concept is set as 'a person wearing clothes'.

Artistic style erasure. We utilize the Image Synthesis Style Studies Database [8], which compiles a list of artists whose styles can be replicated by SD v1.4. From this database, we sample 200 artists and split them into two groups: an erasure group of 100 artists and a retention group with 100 other artists. The full list for both groups is presented in Table 8. To assess efficacy and specificity, we apply the same five prompts and seeds as in [4] for both the erased and retained artists group. These prompts include *'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 each of 100 artists, we use each prompt to generate five images, resulting in 25 images per artist. Thus, this yields 2500 images for each group.

C.2. Training Configurations

Implementation of previous works. In our series of four experiments, we focus on comparing our proposed method with existing methods, including ESD-u[1] [3], ESD-x [3], FMN[2] [22], SLD-M[3] [17], UCE[4] [4], AC[5] [11], and SA[6] [6]. Notably, SA [6] demands extensive resources, requiring 4 RTX A6000s and over 12 hours of training for concept erasure. Consequently, we have not replicated their findings due to these extensive requirements. Instead, we align our explicit content erasure task with SA's settings [6], and we employ their reported experimental results for our comparative analysis. Beyond SA [6], we implement each existing method following their recommended configurations for various erasure types (such as objects, style, or nudity). It is important to note that several methods (e.g., FMN [22] and AC [11]) are not tailored for erasing multiple concepts. In our preliminary tests, we observe that without altering the algorithm or further tuning the suggested parameters, training for multiple concepts—either sequentially or in parallel—yielded comparably mediocre results, marked by either inadequate specificity or generality. Consequently, we opt for a parallel training manner for them when erasing multiple concepts to save resources.

[1] https://github.com/rohitgandikota/erasing
[2] https://github.com/SHI-Labs/Forget-Me-Not
[3] https://github.com/ml-research/safe-latent-diffusion
[4] https://github.com/rohitgandikota/unified-concept-editing
[5] https://github.com/nupurkmr9/concept-ablation
[6] https://github.com/clear-nus/selective-amnesia

Table 6. **The Evaluation Setup for Celebrity Erasure:** Our celebrity dataset contains an erasure group with 100 celebrities and a retention group with another 100 celebrities. Portraits of these celebrities can be effectively generated using SD v1.4. The generated portraits are accurately recognizable by the GIPHY Celebrity Detector (GCD) with an accuracy exceeding 99%. To perform erasure experiments involving 1, 5, 10, and 100 celebrities, a corresponding number of celebrities are selected from the erasure group for each experiment. In all cases, the entire retention group is utilized.

Group	# of Celebrities to Be Erased	Mapping Concept	Celebrity
Erasure Group	1	'a woman'	*'Melania Trump'*
	5	'a person'	*'Adam Driver', 'Adriana Lima', 'Amber Heard', 'Amy Adams', 'Andrew Garfield'*
	10	'a person'	*'Adam Driver', 'Adriana Lima', 'Amber Heard', 'Amy Adams', 'Andrew Garfield', 'Angelina Jolie', 'Anjelica Huston', 'Anna Faris', 'Anna Kendrick', 'Anne Hathaway'*
	100	'a person'	*'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', 'Jennifer Lopez', 'Jeremy Renner', 'Jessica Biel', 'Jessica Chastain', 'John Oliver', 'John Wayne', 'Johnny Depp', 'Julianne Hough', 'Justin Timberlake', 'Kate Bosworth', 'Kate Winslet', 'Leonardo Dicaprio', 'Margot Robbie', 'Mariah Carey', 'Melania Trump', 'Meryl Streep', 'Mick Jagger', 'Mila Kunis', 'Milla Jovovich', 'Morgan Freeman', 'Nick Jonas', 'Nicolas Cage', 'Nicole Kidman', 'Octavia Spencer', 'Olivia Wilde', 'Oprah Winfrey', 'Paul Mccartney', 'Paul Walker', 'Peter Dinklage', 'Philip Seymour Hoffman', 'Reese Witherspoon', 'Richard Gere', 'Ricky Gervais', 'Rihanna', 'Robin Williams', 'Ronald Reagan', 'Ryan Gosling', 'Ryan Reynolds', 'Shia Labeouf', 'Shirley Temple', 'Spike Lee', 'Stan Lee', 'Theresa May', 'Tom Cruise', 'Tom Hanks', 'Tom Hardy', 'Tom Hiddleston', 'Whoopi Goldberg', 'Zac Efron', 'Zayn Malik'*
Retention Group	1, 5, 10, and 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 7. The detailed breakdown of the number (#) of images generated for each celebrity erasure experiment.

# of Celebrities to Be Erased	Celebrity Group	# of Images Generated for Each Celebrity	Total # of Generated Images
1	Erasure Group	250	250
	Retention Group	25	2500
5	Erasure Group	50	250
	Retention Group	25	2500
10	Erasure Group	25	250
	Retention Group	25	2500
100	Erasure Group	25	2500
	Retention Group	25	2500

Table 8. **The Evaluation Setup for Artistic Style Erasure:** We sample 200 artists from the Image Synthesis Style Studies Database [8]. They are split into two groups: an erasure group with 100 artists and a retention group with another 100 artists. The artworks of these artists can be successfully replicated by SD v1.4.

Group	# of Artistic Styles to Be Erased	Mapping Concept	Artist
Erasure Group	100	'art'	*'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 Rivière', '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 Griffiths', 'Casey Weldon', 'Caspar David Friedrich', 'Cassius Marcellus Coolidge', 'Catrin Welz-Stein', 'Cedric Peyravernay', 'Chad Knight', 'Chantal Joffe', 'Charles Addams', 'Charles Angrand', 'Charles Blackman', 'Charles Camoin', 'Charles Dana Gibson', 'Charles E. Burchfield', 'Charles Gwathmey', 'Charles Le Brun', 'Charles Liu', 'Charles Schridde', 'Charles Schulz', 'Charles Spencelayh', 'Charles Vess', 'Charles-Francois Daubigny', 'Charlie Bowater', 'Charline Von Heyl', 'Chaïm Soutine', 'Chen Zhen', 'Chesley Bonestell', 'Chiharu Shiota', 'Ching Yeh', 'Chip Zdarsky', 'Chris Claremont', 'Chris Cunningham', 'Chris Foss', 'Chris Leib', 'Chris Moore', 'Chris Ofili', '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 Efimovich Arkhipov', 'Adam Elsheimer', 'Adam Hughes', 'Adam Martinakis', 'Adam Paquette', 'Adi Granov', 'Adolf Hirémy-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', 'Affandi', '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 Durer', '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'*

Implementation of MACE. This section details the implementation of MACE, focusing on the hyperparameters applied across various experimental scenarios, as outlined in Table 9. For the erasure of explicit content, we leverage general prior knowledge estimated from the MSCOCO dataset, without incorporating any domain-specific knowledge. For the erasure of celebrity likenesses and artistic styles, MACE again utilizes the general prior knowledge from the MSCOCO dataset. Additionally, domain-specific knowledge is employed, which is calculated based on the corresponding retention groups that users wish to preserve. Object erasure presents a special case where the prior knowledge from the MSCOCO dataset includes the concepts we aim to erase (e.g., cat, dog, or airplane). Thus, a direct application of the standard approach is not feasible. To address this, we modify the loss function. Instead of using the second term in Eq. (1), we use $\|\mathbf{W}'_k - \mathbf{W}_k\|_2^2$ to preserve the original knowledge.

D. Additional Evaluation Results of Erasing the CIFAR-10 Classes

Table 10 presents the results of erasing the final six object classes of the CIFAR-10 dataset [10]. Our approach achieves the highest harmonic mean across the erasure of these six object classes. This highlights the exceptional erasure capabilities of our approach, effectively balancing specificity and generality.

Table 9. **Hyperparameters Utilized in MACE Across Different Experimental Sets.**

Erasure Type	Segment	LoRA Training Step	Learning Rate	$\lambda_1 = \lambda_2$	λ_3	Rank r
Object	Airplane	50	1.0×10^{-5}	1000.0	-	1
	Automobile	50	1.0×10^{-5}	100.0	-	1
	Bird	50	1.0×10^{-5}	10.0	-	1
	Cat	50	1.0×10^{-5}	1000.0	-	1
	Deer	50	1.0×10^{-5}	10.0	-	1
	Dog	50	1.0×10^{-5}	10.0	-	1
	Frog	50	1.0×10^{-5}	0.4	-	1
	Horse	50	1.0×10^{-5}	1.0	-	1
	Ship	50	1.0×10^{-5}	1000.0	-	1
	Truck	50	1.0×10^{-5}	0.1	-	1
Celebrity	1 Celebrity	50	1.0×10^{-4}	1.0×10^{-4}	0.8	1
	5 Celebrities	50	1.0×10^{-4}	1.0×10^{-4}	5.0	1
	10 Celebrities	50	1.0×10^{-4}	1.0×10^{-4}	8.0	1
	100 Celebrities	50	1.0×10^{-4}	1.0×10^{-4}	20.0	1
Artistic Style	100 Artistic Styles	50	1.0×10^{-4}	1.0×10^{-4}	8.0	1
Explicit Content	'Nudity', 'Naked', 'Erotic', 'Sexual'.	120	1.0×10^{-5}	7.0×10^{-7}	-	1

Table 10. **Evaluation of Erasing the CIFAR-10 Classes:** Results for the final six individual classes are presented. CLIP classification accuracies are reported for each erased class in three sets: the erased class itself (Acc_e, efficacy), the nine remaining unaffected classes (Acc_s, specificity), and three synonyms of the erased class (Acc_g, generality). The harmonic means H_o reflect the comprehensive erasure capability. All presented values are denoted in percentage (%). The classification accuracies of images generated by the original SD v1.4 are presented for reference.

Method	Deer Erased				Dog Erased				Frog Erased			
	$\text{Acc}_\text{e} \downarrow$	$\text{Acc}_\text{s} \uparrow$	$\text{Acc}_\text{g} \downarrow$	$H_\text{o} \uparrow$	$\text{Acc}_\text{e} \downarrow$	$\text{Acc}_\text{s} \uparrow$	$\text{Acc}_\text{g} \downarrow$	$H_\text{o} \uparrow$	$\text{Acc}_\text{e} \downarrow$	$\text{Acc}_\text{s} \uparrow$	$\text{Acc}_\text{g} \downarrow$	$H_\text{o} \uparrow$
FMN [22]	98.95	94.13	60.24	3.04	97.64	98.12	96.95	3.94	91.60	94.59	63.61	19.10
AC [11]	99.45	98.47	64.78	1.62	98.50	98.57	95.76	3.29	99.92	98.62	92.44	0.24
UCE [4]	11.88	98.39	8.94	92.34	13.22	98.69	14.63	89.90	20.86	98.32	18.50	85.53
SLD-M [17]	57.62	98.45	39.91	59.53	94.27	98.53	82.84	12.35	81.92	98.19	59.78	33.20
ESD-x [3]	19.01	96.98	10.19	88.77	28.54	96.38	44.49	70.78	11.56	97.37	13.73	90.45
ESD-u [3]	18.14	73.81	6.93	82.17	27.03	89.75	28.52	77.24	12.32	88.05	7.62	89.32
Ours	13.47	97.71	6.08	**92.48**	11.07	96.77	10.86	**91.47**	11.45	97.75	13.08	**90.83**
SD v1.4 [16]	99.87	98.49	70.02	-	98.74	98.62	98.25	-	99.93	98.49	92.04	-

Method	Horse Erased				Ship Erased				Truck Erased			
	$\text{Acc}_\text{e} \downarrow$	$\text{Acc}_\text{s} \uparrow$	$\text{Acc}_\text{g} \downarrow$	$H_\text{o} \uparrow$	$\text{Acc}_\text{e} \downarrow$	$\text{Acc}_\text{s} \uparrow$	$\text{Acc}_\text{g} \downarrow$	$H_\text{o} \uparrow$	$\text{Acc}_\text{e} \downarrow$	$\text{Acc}_\text{s} \uparrow$	$\text{Acc}_\text{g} \downarrow$	$H_\text{o} \uparrow$
FMN [22]	99.63	93.14	46.61	1.10	97.97	98.21	96.75	3.70	97.64	97.86	95.37	4.62
AC [11]	99.74	98.63	45.29	0.77	98.18	98.50	77.47	4.97	98.50	98.61	95.12	3.40
UCE [4]	4.66	98.32	12.70	93.42	6.13	98.41	21.44	89.44	20.58	98.16	50.00	70.13
SLD-M [17]	81.76	98.44	36.71	37.14	89.24	98.56	41.02	24.99	91.06	98.72	80.62	17.29
ESD-x [3]	16.86	97.02	15.05	87.96	33.35	97.93	34.78	73.99	36.06	97.24	44.29	68.38
ESD-u [3]	17.69	82.23	9.89	84.73	18.38	94.32	15.93	86.33	26.11	85.35	21.47	78.98
Ours	4.89	97.48	7.85	**94.86**	8.58	98.56	14.40	**91.56**	7.29	98.38	9.38	**93.79**
SD v1.4 [16]	99.78	98.50	45.74	-	98.64	98.63	64.16	-	98.89	98.60	95.00	-

E. Concept-Focal Importance Sampling

Figure 7 presents a graph plotting the probability density function $\xi(t)$ defined in Eq. (5), which is:

$$\xi(t) = \frac{1}{Z}\left(\sigma\left(\gamma(t - t_1)\right) - \sigma\left(\gamma(t - t_2)\right)\right),$$

where Z is a normalizer, $\sigma(x)$ is the sigmoid function $1/(1 + e^{-x})$, with t_1 and t_2 as the bounds of a high probability sampling interval ($t_1 < t_2$), and γ as a temperature hyperparameter. We set $t_1 = 200, t_2 = 400$, and $\gamma = 0.05$ throughout our experiments.

This strategy excels particularly in eliminating mass concepts that share overlapping terms. However, it is important to note that when erasing a smaller number of concepts or mass concepts that do not share overlapping terms, the improvements tend to be incremental. Beyond enhancing specificity, this design significantly boosts training effectiveness by fostering a more concentrated and efficient learning process.

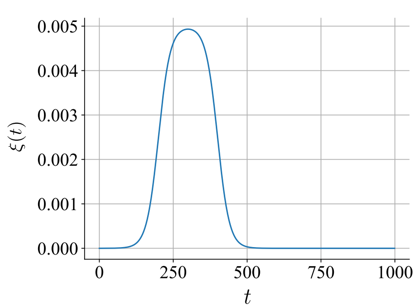


Figure 7. The graph of probability density function of timestep t for reference.

F. Additional Ablation Studies and Applications

Additional ablation studies. We also carry out ablation studies focused on independently adjusting either the 'Key' or 'Value' projection matrices, as detailed in Table 11. Intriguingly, exclusively finetuning the'Value' projection matrices for an identical number of steps can result in the deterioration of unrelated concepts, as indicated by a lower Acc_s. While fine-tuning only the 'Value' projection matrices might seem efficient for obtaining satisfactory outcomes with minimal adjustments, its peak performance is notably inferior.

Table 11. **Ablation Study on the Impact of LoRA Finetuned Projection Matrices in Erasing 100 Celebrities.** All presented values are denoted in percentage (%).

Config	Variation	Tuning Step	Metrics		
			$\text{Acc}_\text{e}\downarrow$	$\text{Acc}_\text{s}\uparrow$	$H_\text{c}\uparrow$
5	Tune Key Only	50 steps	12.72	80.54	83.77
6	Tune Value Only	50 steps	0.77	38.65	55.63
		25 steps	3.28	62.74	76.11
		10 steps	10.47	77.81	83.26
		5 steps	12.72	80.45	83.73
		3 steps	14.46	81.70	83.58
		1 steps	15.39	82.37	83.48
Ours	Tune Key & Value	50 steps	4.31	84.56	**89.78**

Additional applications. MACE possesses the capability to simultaneously erase different types of concepts, such as both a celebrity likeness and artistic style, as shown in Figure 8 (a). Furthermore, MACE is compatible with distilled fast diffusion models (e.g., Latent Consistency Model [13]), with an example presented in Figure 8 (b).

(a) Erase different types of concepts simultaneously **(b) MACE on LCM (4-step generation)**

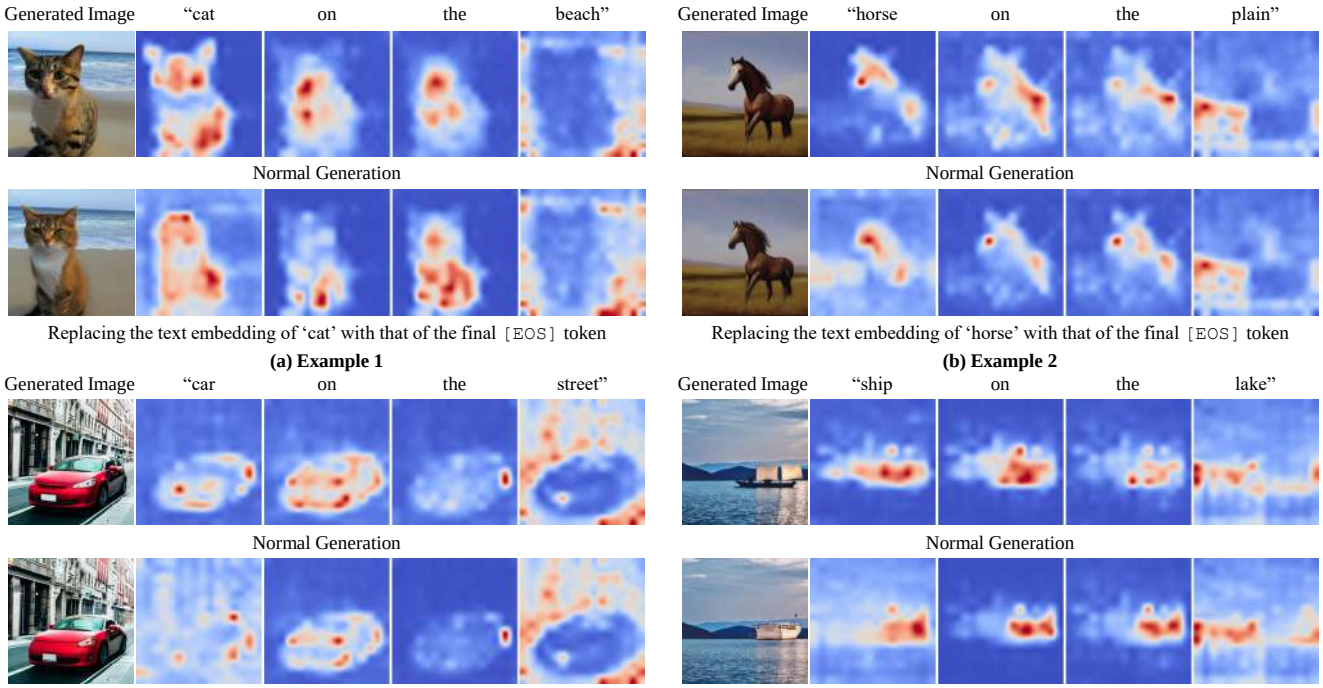
Figure 8. **(a) MACE can simultaneously erase different types of concepts.** The left image is generated from the original SD v1.4. The right one is generated by the MACE finetuned SD v1.4 which erases the concepts of a celebrity (Bill Clinton) and a artistic style (Brent Heighton). Both images are generated using the prompt *'Bill Clinton walking, Brent Heighton style'*. **(b) MACE is compatible with distilled diffusion models**. The left image is generated from the original LCM Dreamshaper v7. The right one is generated by the MACE finetuned LCM which erases the concept of 'Trump'. Both images are generated using the prompt *'a photo of Trump playing guitar'*.

G. Additional Qualitative Results

Figure 9 provides further instances of concept generation utilizing residual information. Despite substituting the text embedding of the core concept with that of the final `[EOS]` token, the attention maps corresponding to the remaining words clearly delineate the contours of the targeted concept. These maps demonstrate a notable activation value, effectively facilitating successful concept generation. Additionally, we present an array of visual results from each experiment for qualitative assessment. The corresponding figure indices are listed in Table 12. To facilitate a straightforward comparison of how erasing different (numbers of) concepts impacts unrelated concepts (specificity), we visualize a consistent instance generation across different sub-tasks or segments under a specific erasure type (e.g., car for object erasure or Bill Murray for celebrity erasure).

Table 12. Summary of tasks with their figure indices.

Erasure Type	Segment	Figure Index
Object Erasure	Airplane	Figure 10
	Automobile	Figure 11
	Bird	Figure 12
	Cat	Figure 13
	Deer	Figure 14
	Dog	Figure 15
	Frog	Figure 16
	Horse	Figure 17
	Ship	Figure 18
	Truck	Figure 19
Celebrity Erasure	1 Celebrity	Figure 20
	5 Celebrities	Figure 21
	10 Celebrities	Figure 22
	100 Celebrities	Figure 23 Figure 24
Artistic Style Erasure	100 Artistic Styles	Figure 25
Explicit Content Erasure	-	Figure 26



(a) Example 1 **(b) Example 2** **(c) Example 3** **(d) Example 4**

Figure 9. **Additional Examples of Generating Concepts Using Residual Information:** In every example presented, the first row illustrates images normally generated by SD v1.4, while the second row displays images generated after replacing the text embedding of the key concept with that of the final [EOS] token. Despite this replacement of the key concept's text embedding, the attention maps of the remaining words distinctly highlight the contours of the intended concept, exhibiting a high activation value.

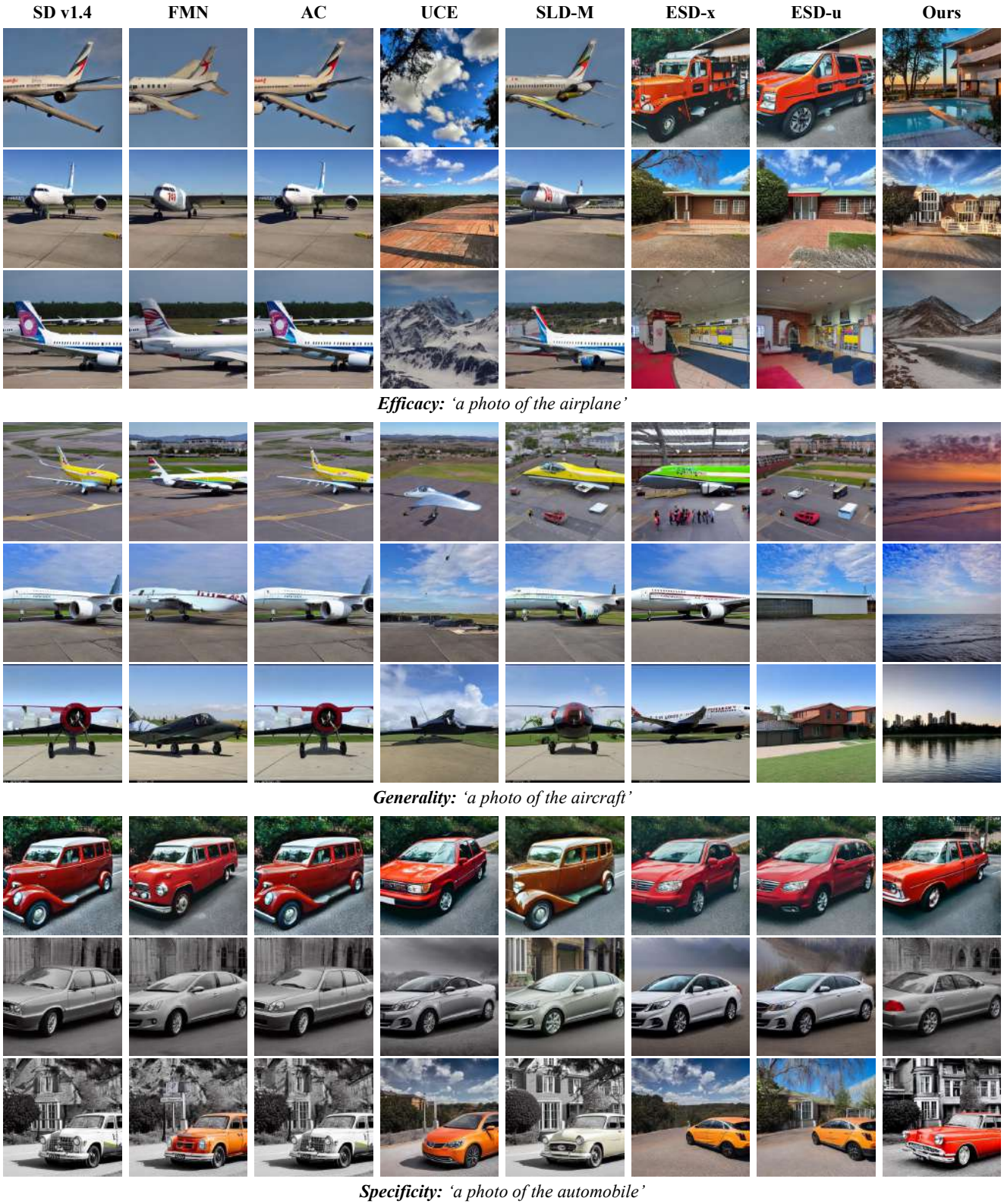


Figure 10. Qualitative comparison on **airplane erasure**. The images on the same row are generated using the same random seed.

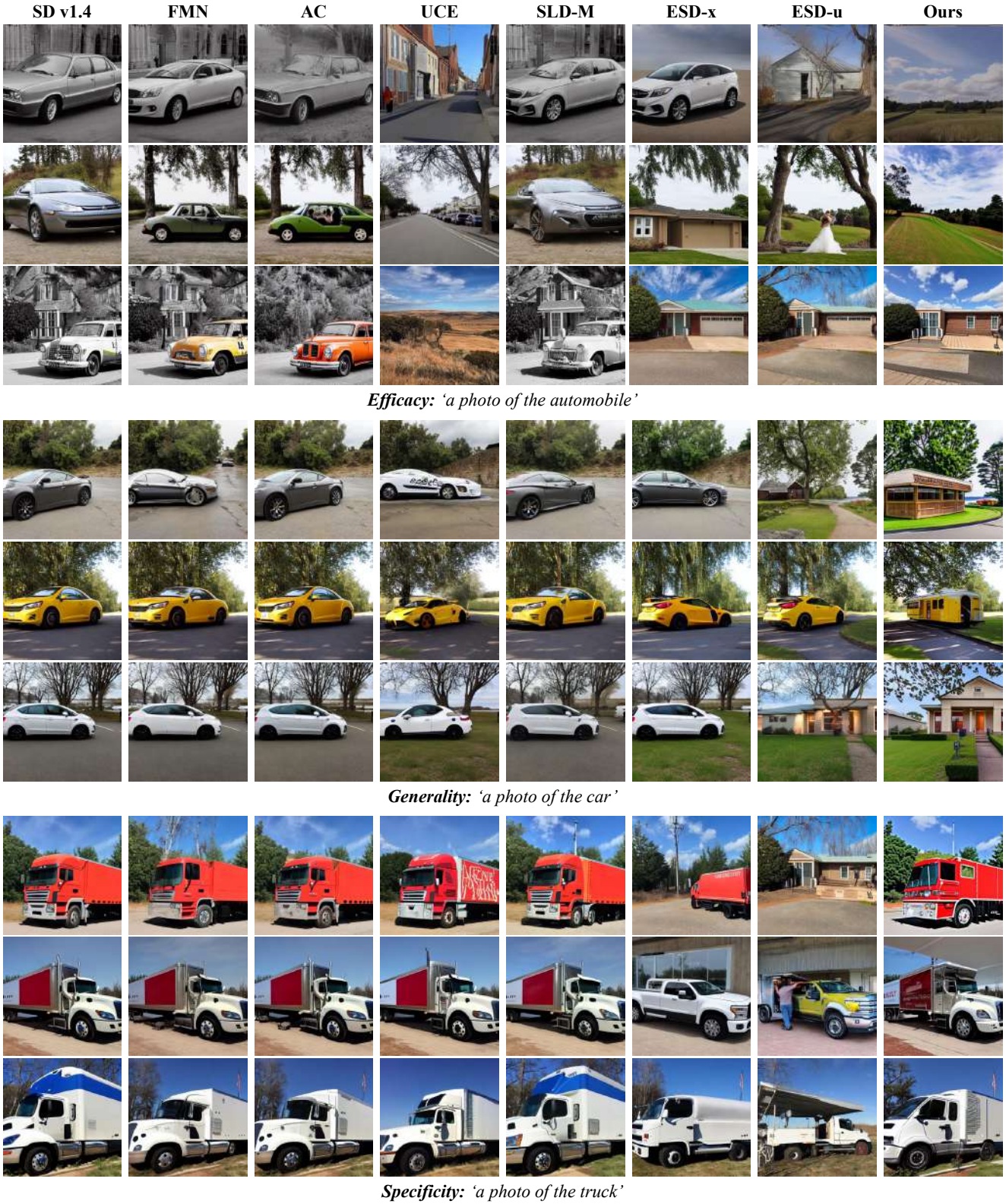


Figure 11. Qualitative comparison on **automobile erasure**. The images on the same row are generated using the same random seed.

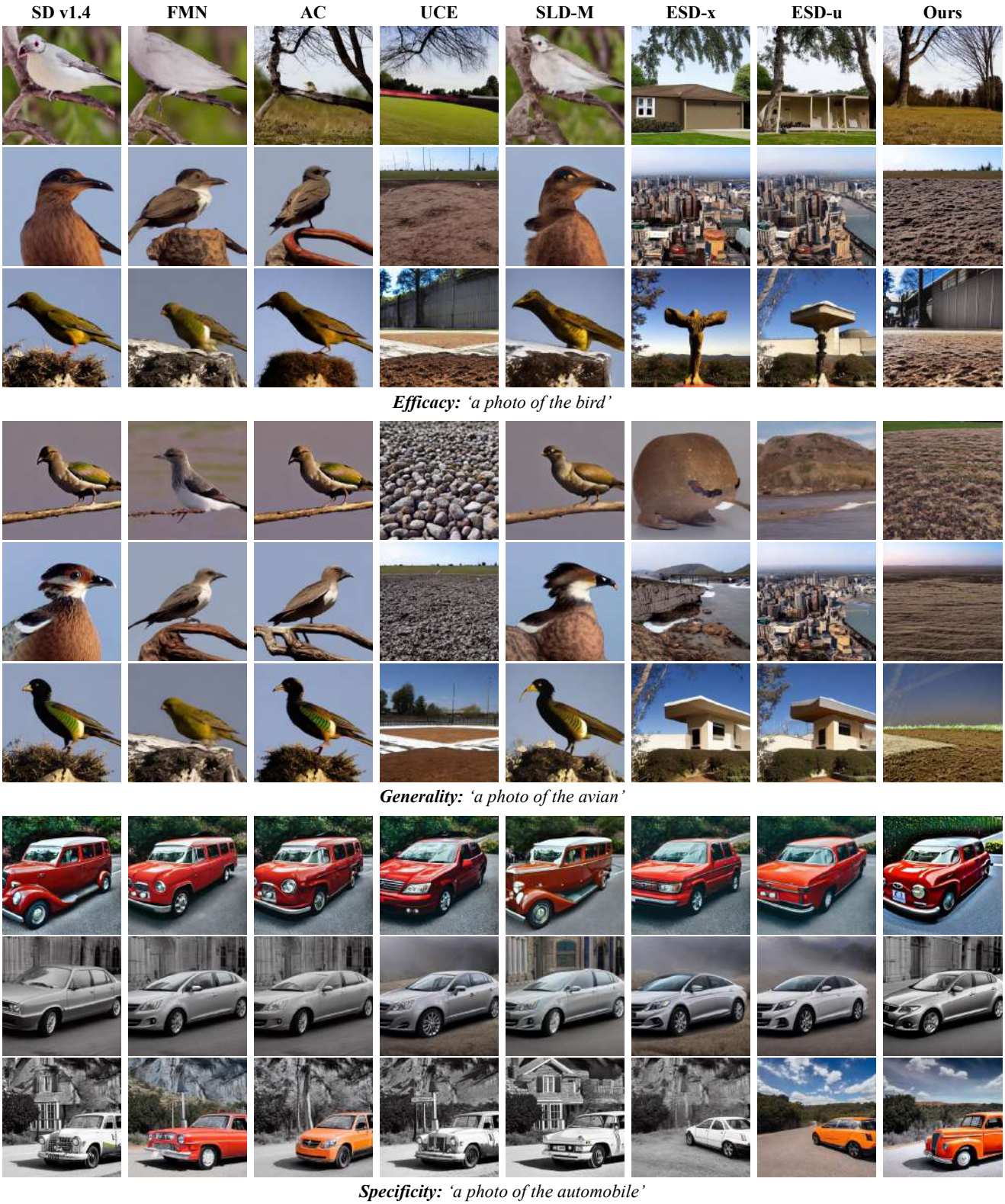


Figure 12. Qualitative comparison on **bird erasure**. The images on the same row are generated using the same random seed.

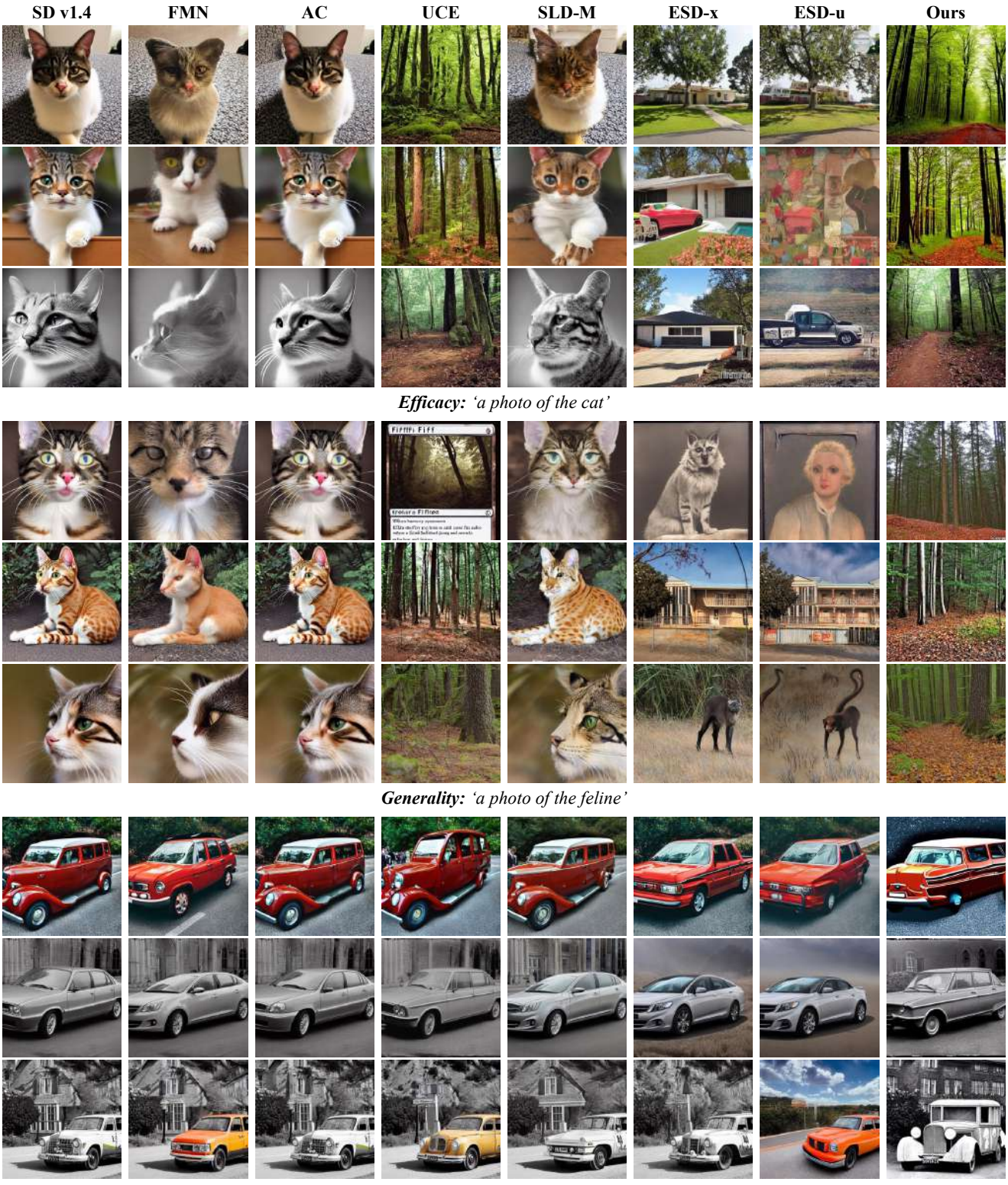


Figure 13. Qualitative comparison on **cat erasure**. The images on the same row are generated using the same random seed.

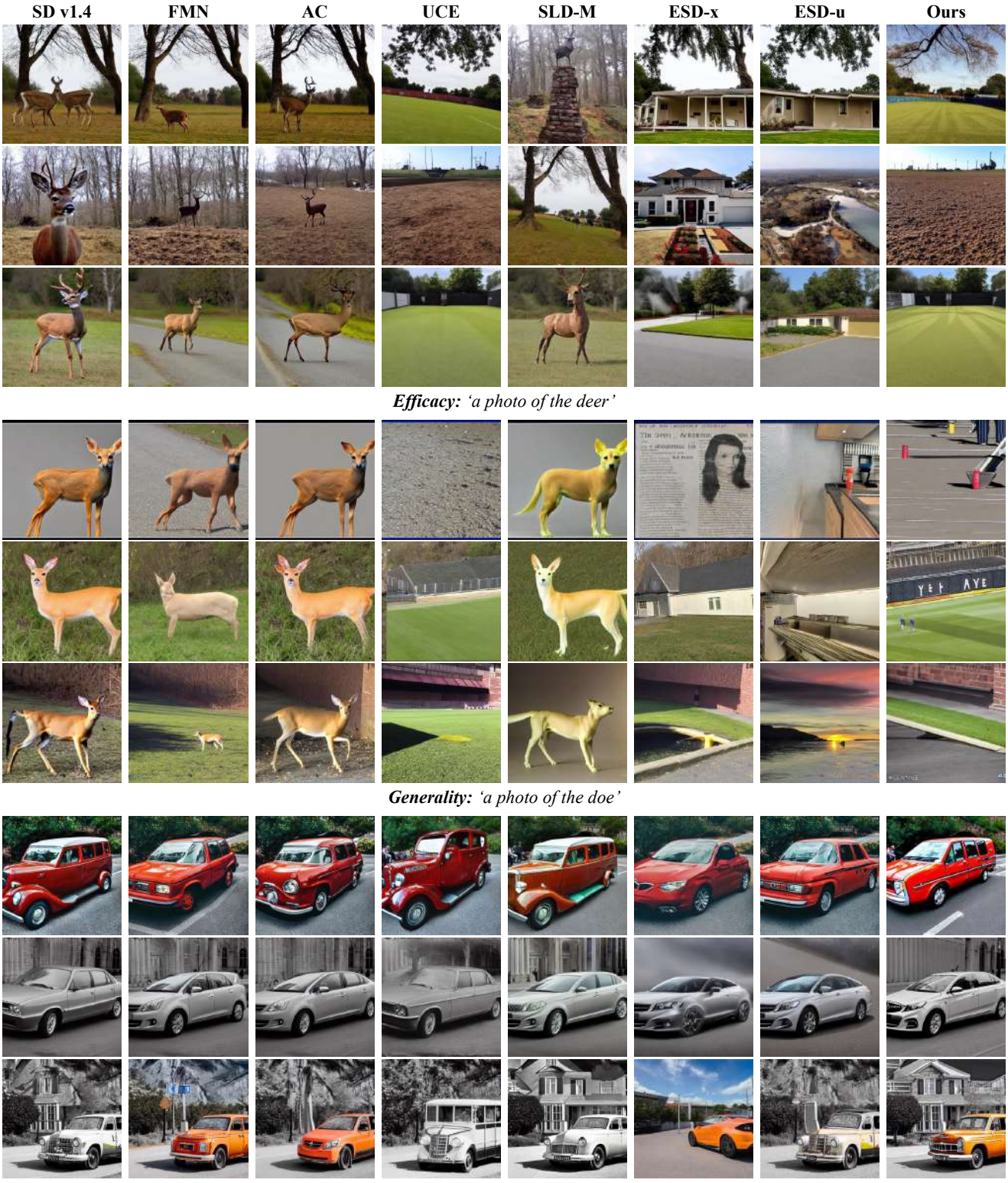


Figure 14. Qualitative comparison on **deer erasure**. The images on the same row are generated using the same random seed.

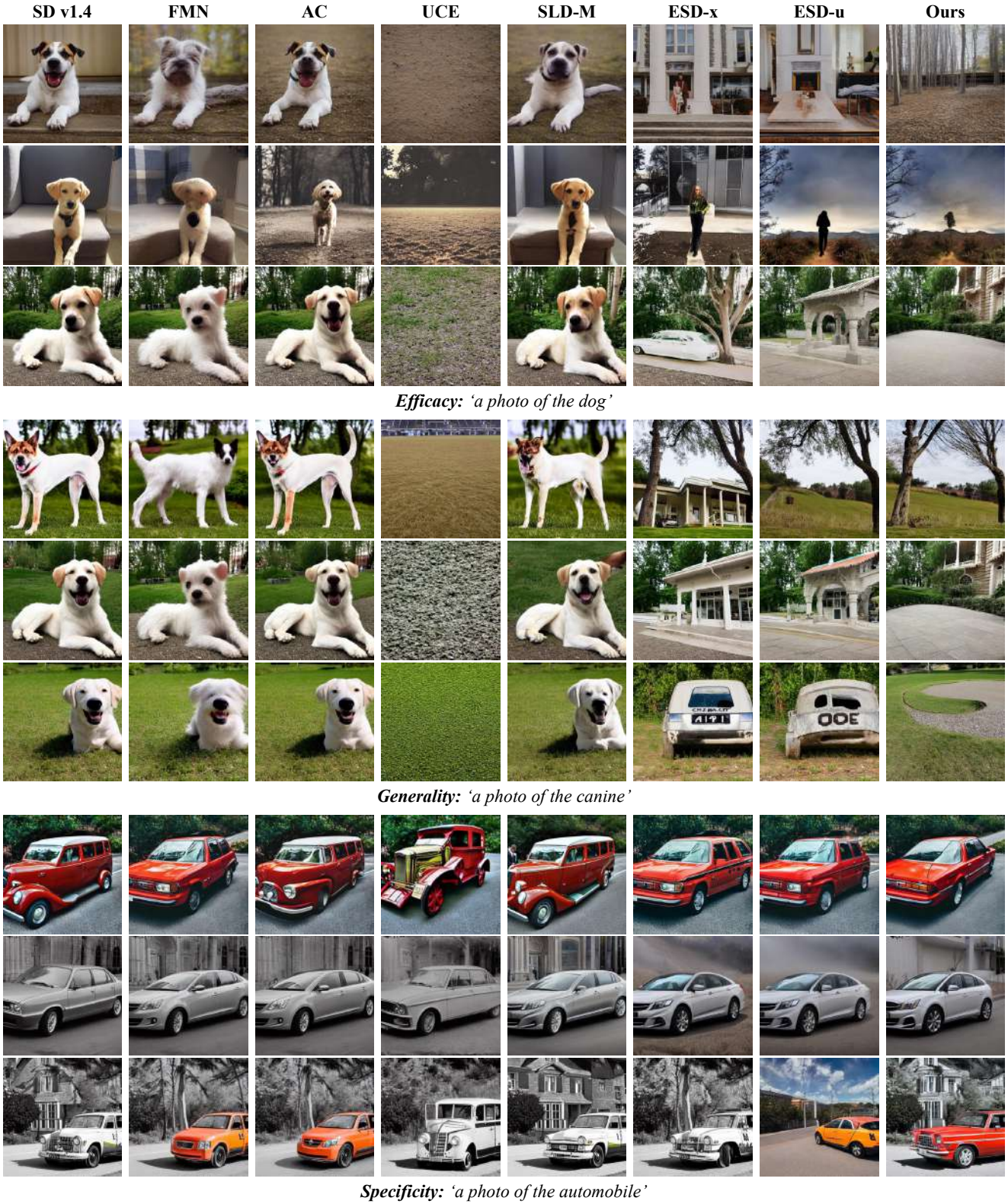


Figure 15. Qualitative comparison on **dog erasure**. The images on the same row are generated using the same random seed.

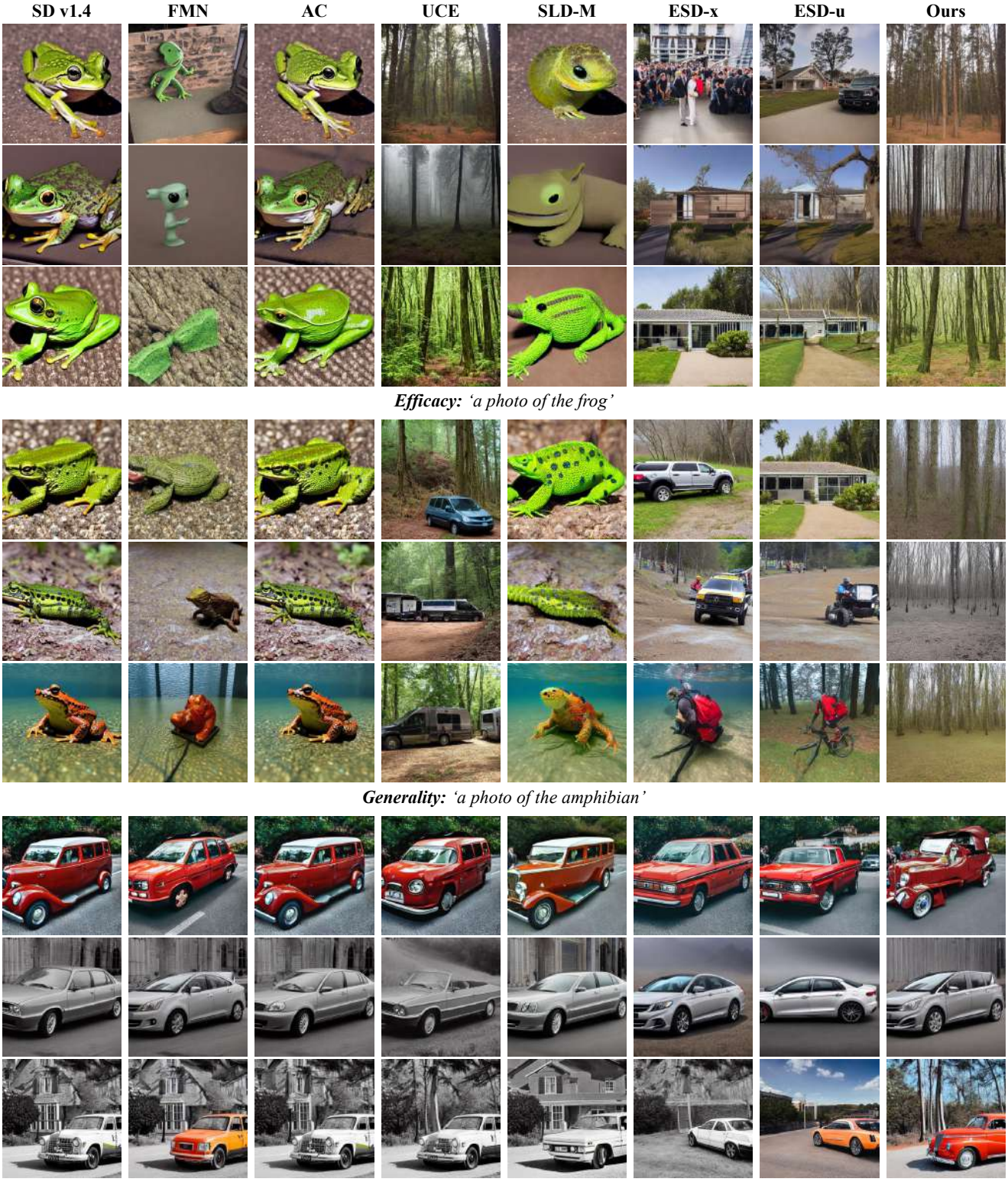


Figure 16. Qualitative comparison on **frog erasure**. The images on the same row are generated using the same random seed.

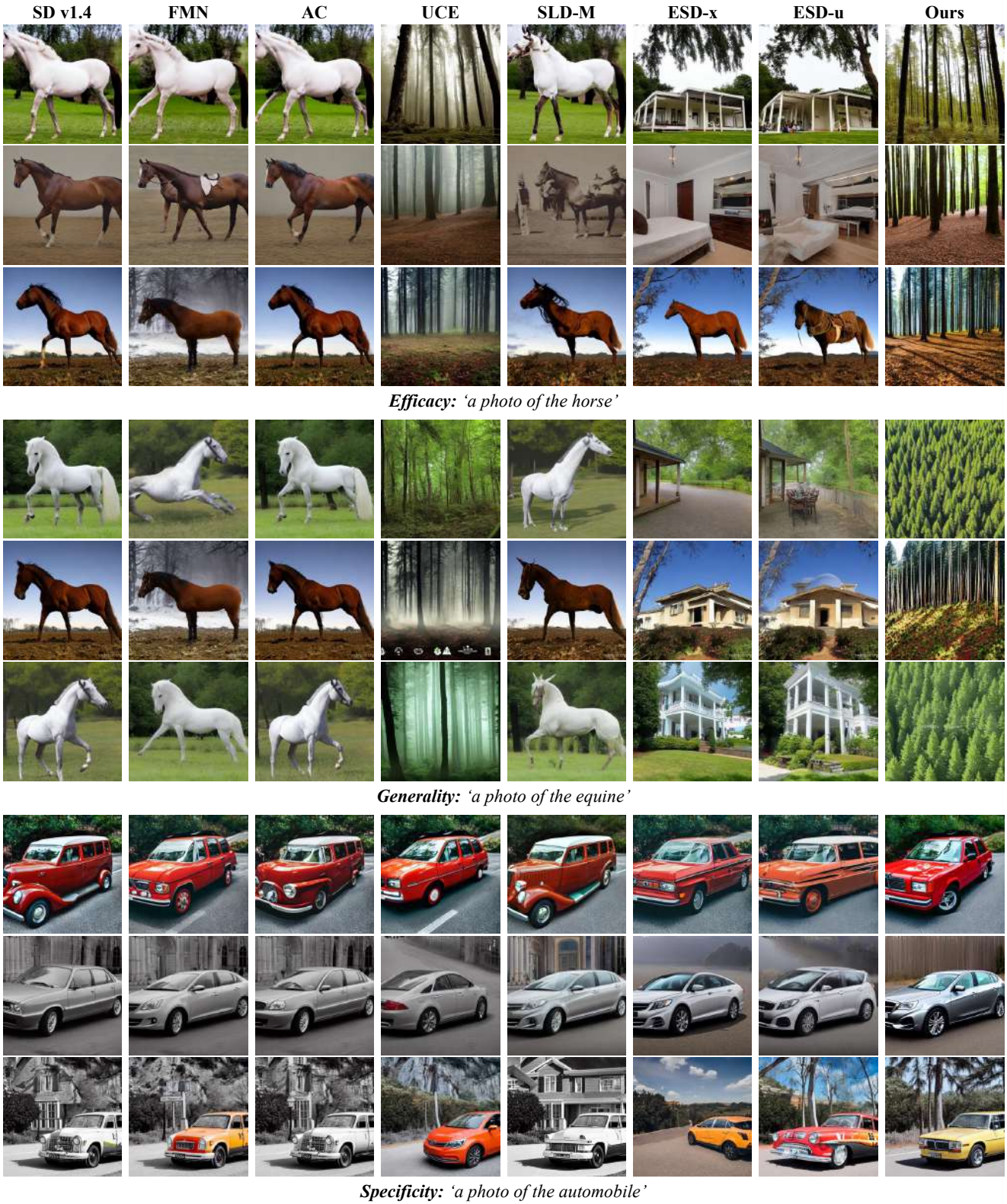


Figure 17. Qualitative comparison on **horse erasure**. The images on the same row are generated using the same random seed.

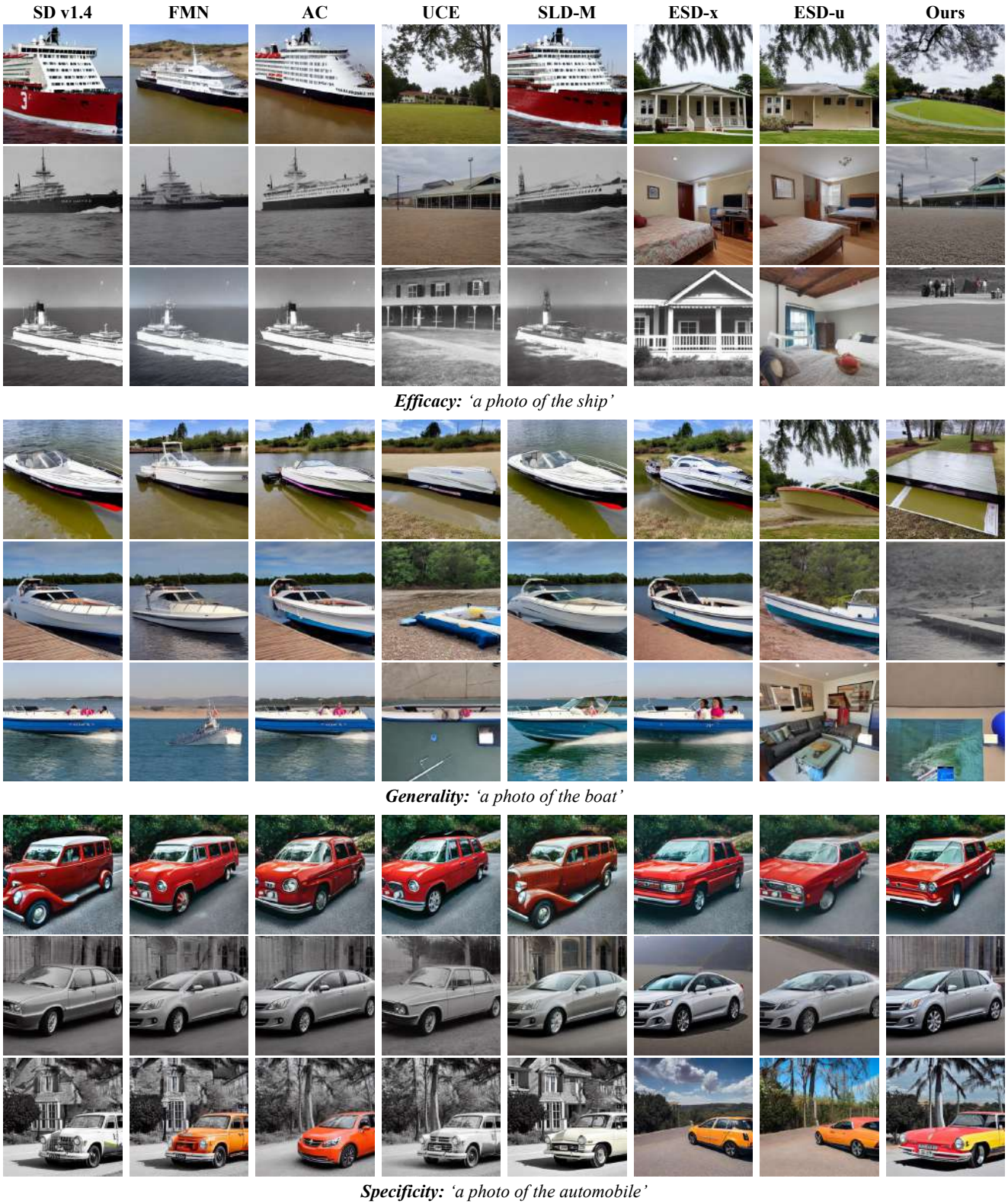


Figure 18. Qualitative comparison on **ship erasure**. The images on the same row are generated using the same random seed.

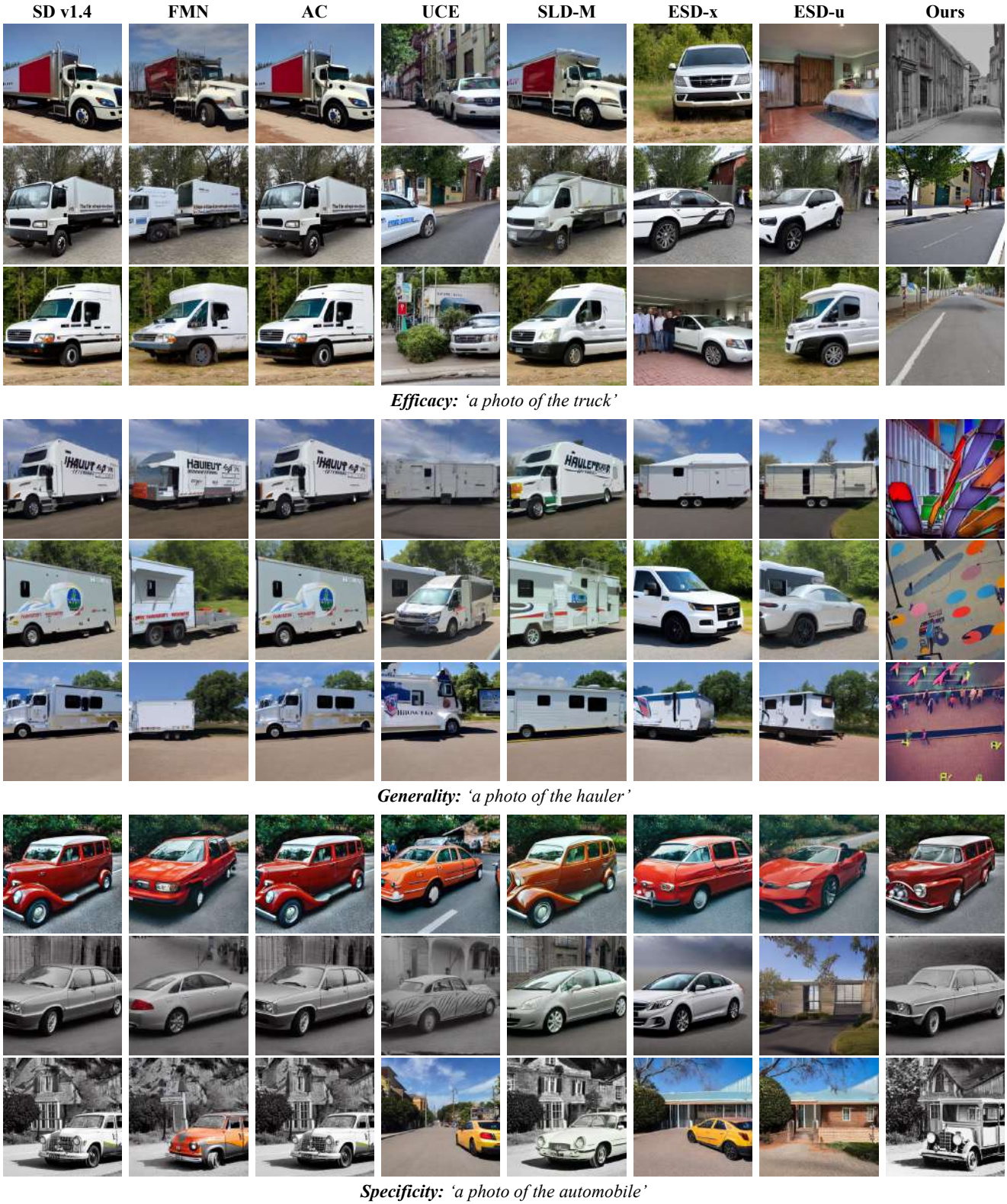


Figure 19. Qualitative comparison on **truck erasure**. The images on the same row are generated using the same random seed.

Figure 20. Qualitative comparison on **1-celebrity erasure**. The images on the same row are generated using the same random seed. Melania Trump is in the erasure group, while Bill Murray and Amanda Seyfried are in the retention group (See Table 6).

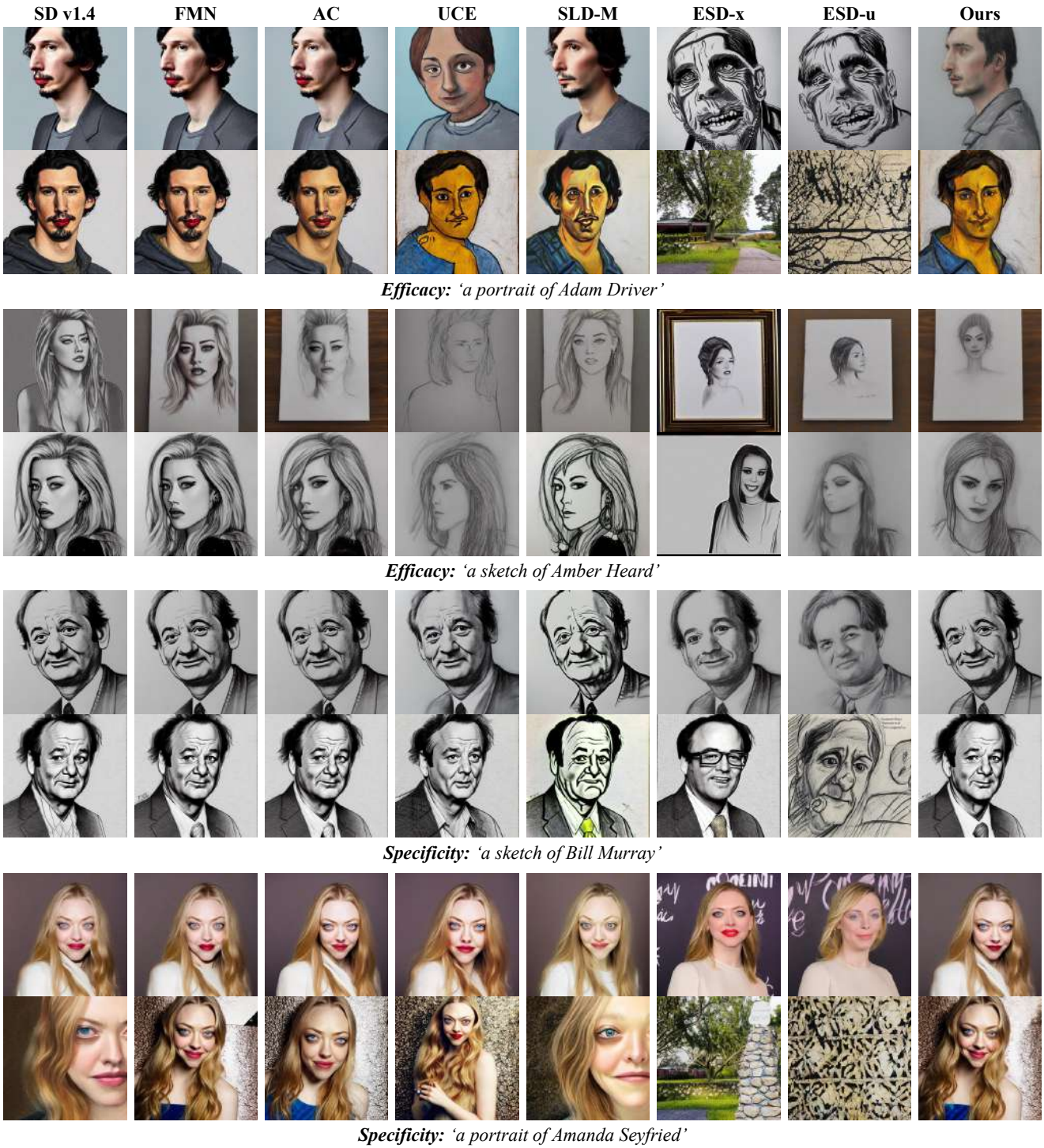


Figure 21. Qualitative comparison on **5-celebrity erasure**. The images on the same row are generated using the same random seed. Adam Driver and Amber Heard are in the erasure group, while Bill Murray and Amanda Seyfried are in the retention group (See Table 6).

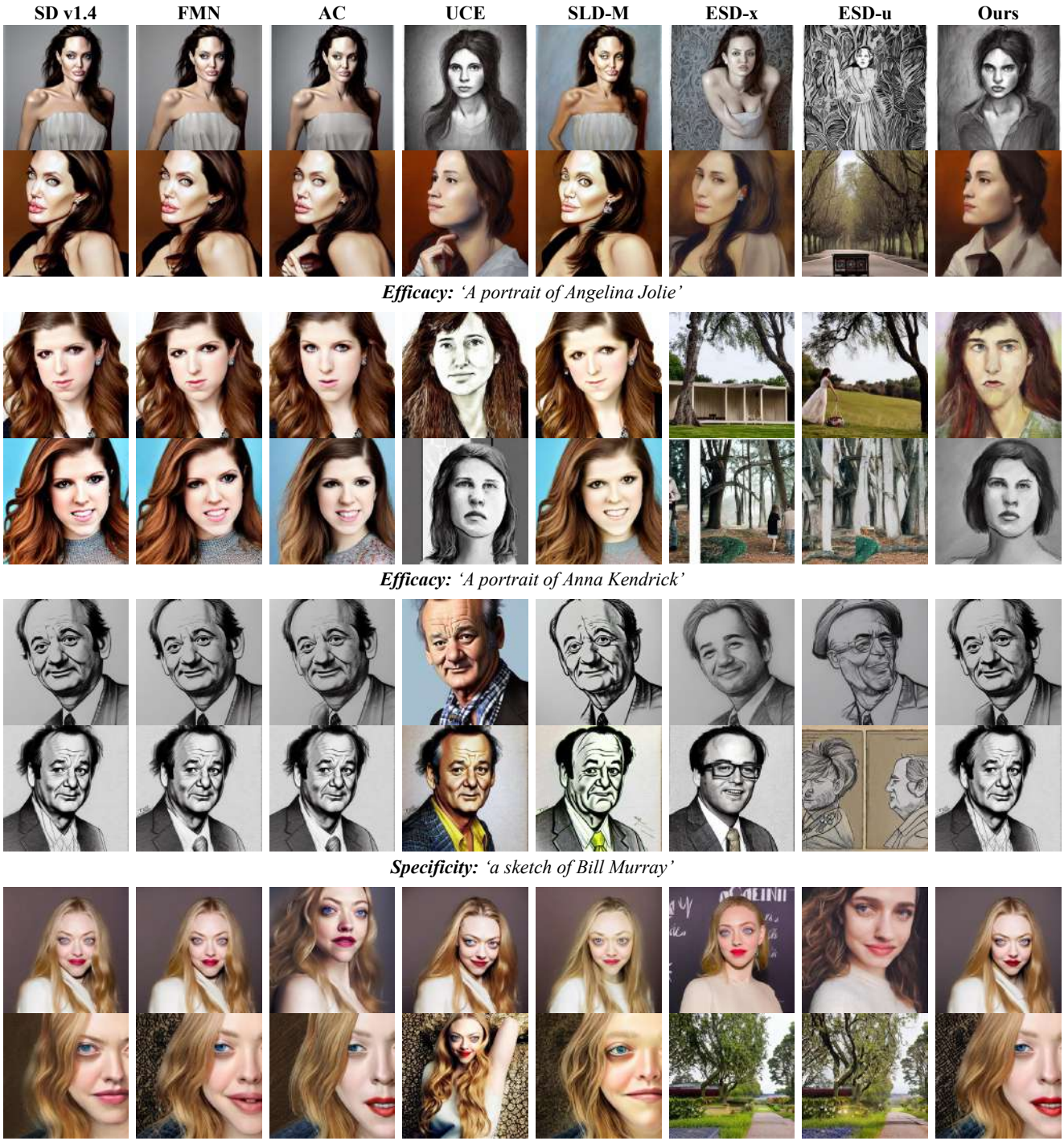


Figure 22. Qualitative comparison on **10-celebrity erasure**. The images on the same row are generated using the same random seed. Angelina Jolie and Anna Kendrick are in the erasure group, while Bill Murray and Amanda Seyfried are in the retention group (See Table 6).

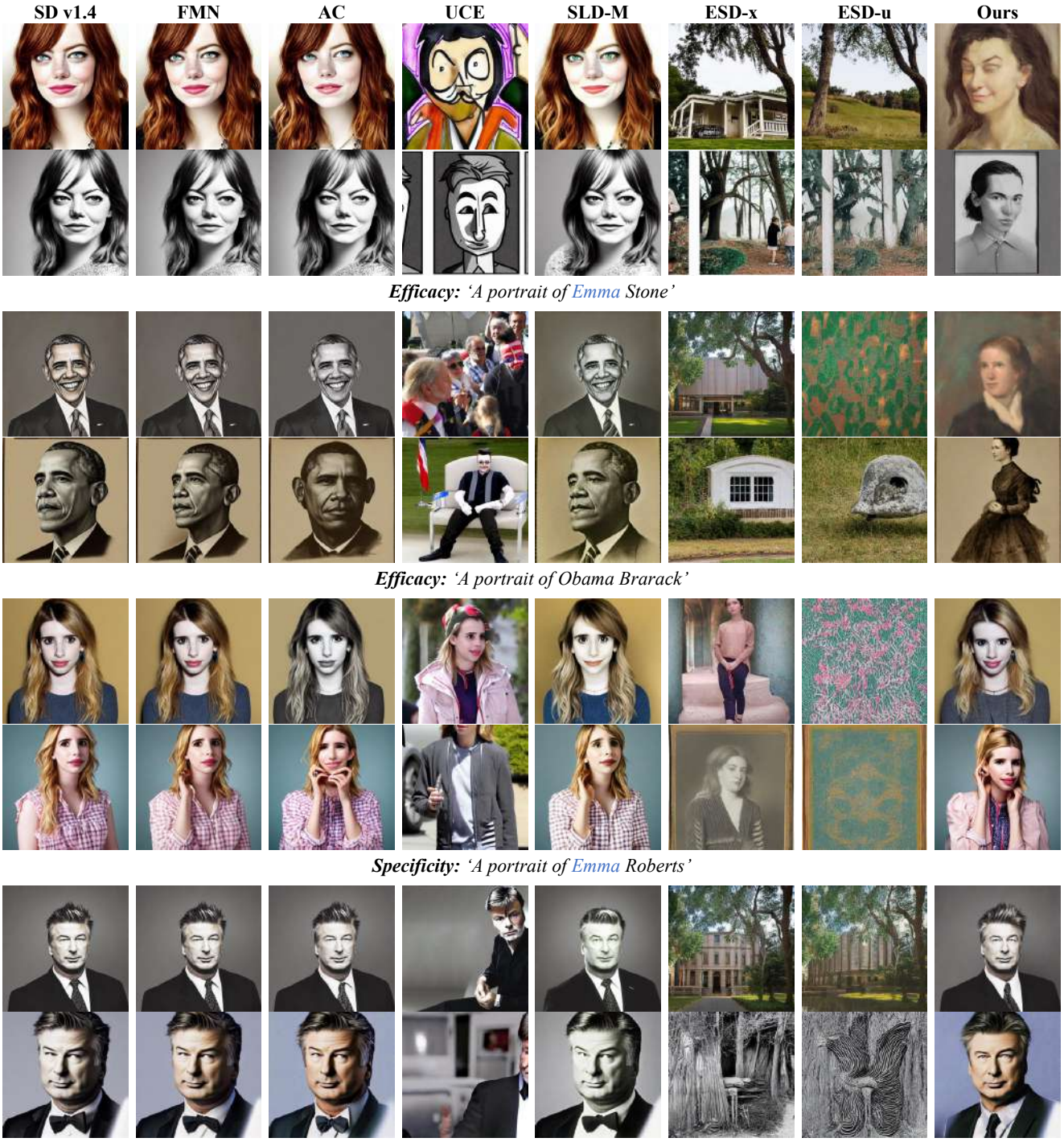


Figure 23. Qualitative comparison on **100-celebrity erasure**. The images on the same row are generated using the same random seed. Emma Stone and Obama Brarack are in the erasure group, while Emma Roberts and Alec Baldwin are in the retention group (See Table 6).

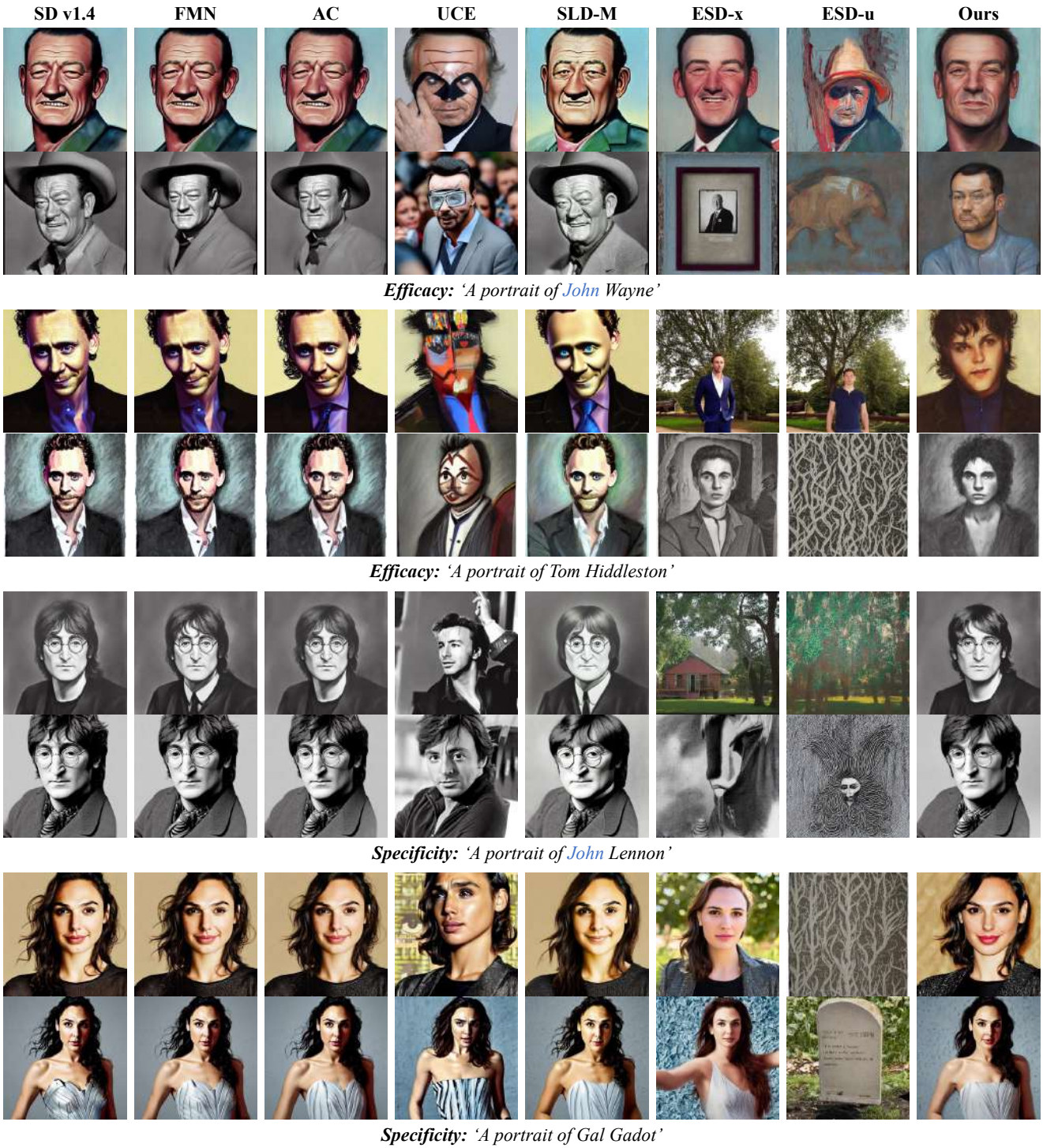


Figure 24. Qualitative comparison on **100-celebrity erasure**. The images on the same row are generated using the same random seed. John Wayne and Tom Hiddleston are in the erasure group, while John Lennon and Gal Gadot are in the retention group (See Table 6).

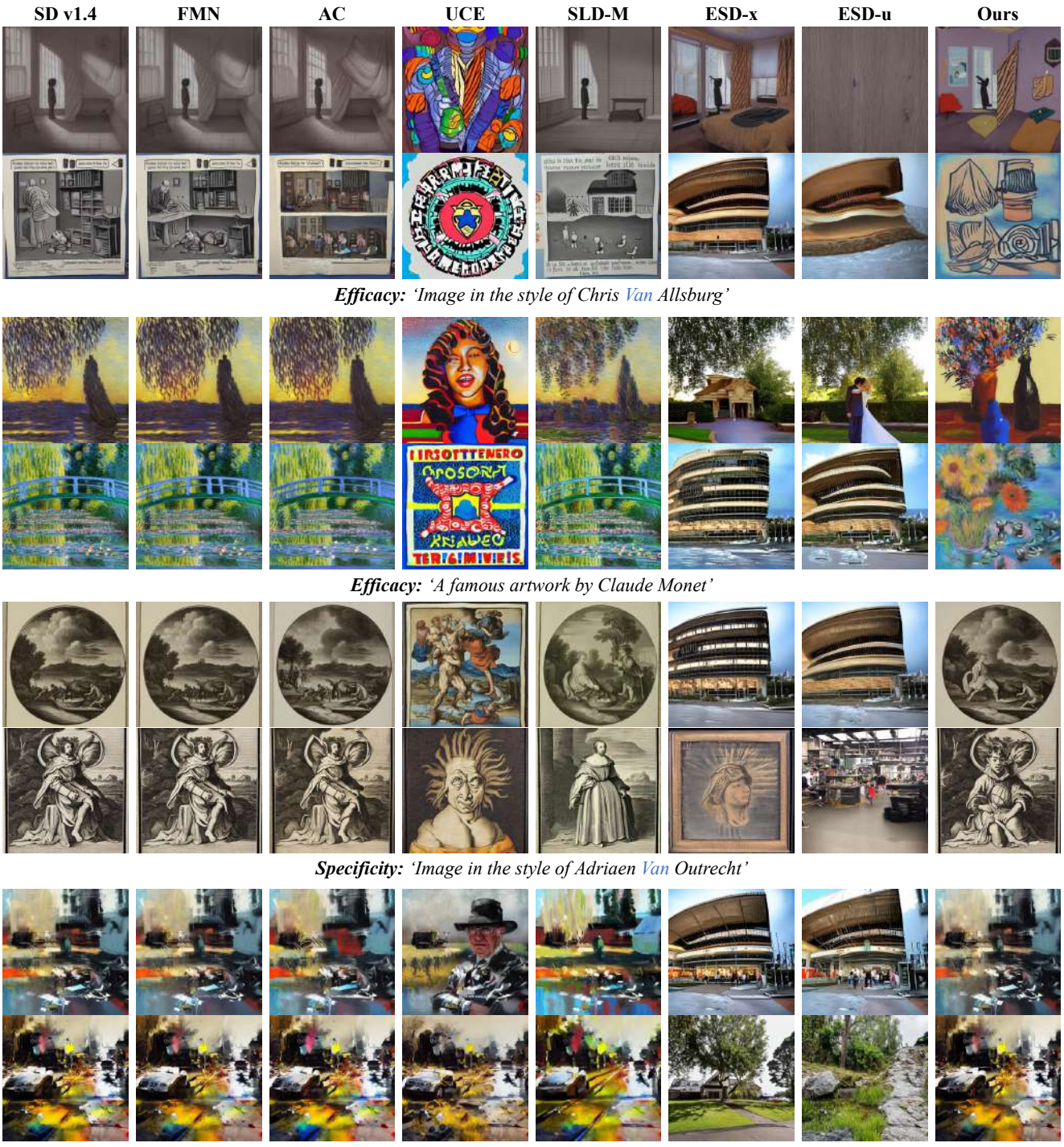


Figure 25. Qualitative comparison on **erasing 100 artistic styles**. The images on the same row are generated using the same random seed. Chris Van Allsburg and Claude Monet are in the erasure group, while Adriaen Van Outrecht and Adrian Ghenie are in the retention group (See Table 8).

Figure 26. Qualitative comparison on **explicit content erasure**. The images on the same row are generated using the same random seed. The sensitive parts are masked by authors.

References

[1] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1

[2] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1

[3] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. *arXiv preprint arXiv:2303.07345*, 2023. 3, 6

[4] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. *arXiv preprint arXiv:2308.14761*, 2023. 2, 3, 6

[5] Nick Hasty, Ihor Kroosh, Dmitry Voitekh, and Dmytro Korduban. Giphy celebrity detector. https://github.com/Giphy/celeb-detection-oss. 3

[6] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *arXiv preprint arXiv:2305.10120*, 2023. 3

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[8] Surea I, Proxima Centauri B, Erratica, and Stephen Young. Image synthesis style studies. https://www.aiartapps.com/ai-art-apps/image-synthesis-style-studies. 3, 5

[9] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 1

[10] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 5

[11] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 3, 6

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[13] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 7

[14] 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*, pages 8748–8763. PMLR, 2021. 1

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551, 2020. 1

[16] 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*, pages 10684–10695, 2022. 1, 6

[17] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 3, 6

[18] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1

[19] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[20] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[22] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023. 3, 6