The Illusion of Unlearning: The Unstable Nature of Machine Unlearning in Text-to-Image Diffusion Models

Supplementary Material

A. Additional Fine-tuning Experiments Details

A.1. Concepts Used for Fine-tuning

In our study, we sequentially fine-tuned the unlearned models on ten distinct concepts to assess the revival potential of the unlearned information. For the object unlearning task, we focused on the "golf ball" concept and related objects. In the style unlearning task, we concentrated on the "Van Gogh" artistic style along with other related artistic movements and styles. We also did celebrity unlearning and NSFW content unlearning also.

A.2. Process of Obtaining Related Concepts

To systematically gather related concepts, we employed the OpenAI API with the GPT-4o model [25] to generate hierarchical lists of concepts associated with our target concepts. This approach ensured a comprehensive exploration of both closely and distantly related concepts, which is crucial for evaluating the robustness of machine unlearning methods.

A.2.1 Hierarchical Concept Generation

For both object and style unlearning tasks, we crafted specific prompts to guide the GPT-4o model in generating relevant concepts.

- **Object Unlearning Prompt:**

```
Generate semantically similar concepts to
    the provided input concept, organized
    into five levels of decreasing
    similarity.

Instructions:
- Begin by identifying concepts that are
    most closely related to the input
    concept.
- Gradually expand to broader or more
    distantly related concepts,
    organizing them into levels of
    decreasing similarity.
- Ensure that each level has concepts
    that are less related than those in
    the preceding level.
- The last level should be the least
    similar concept related to input
    concept similar to random/general
    concept.

Steps
1. Input Analysis: Begin by understanding
    the input concept, identifying its
    key characteristics and essential
    features.
2. Concept Search: Look for related
    concepts using those key
    characteristics as a basis. Start
    with those sharing direct
    similarities.
3. Level Organization: Organize the
    discovered concepts into five levels.
    Level 1 should contain the most
    similar concepts, with each
    subsequent level containing concepts
    of decreasing similarity.
4. Validation: Ensure each concept on a
    level is less similar than those on
    the previous level.

At every level, find 2 concepts. So,
    there are a total of 10 concepts.

Output Format:
- The output should be a numbered list
    from 1 to 10, without any
    explanations or additional commentary.
- Use the following structure: 1. 2. 3.
    4. 5. 6. 7. 8. 9. 10.
```

- **Style Unlearning Prompt:**

```
Generate a list of 10 artistic styles
    that are semantically related to the
    provided input style, organized into
    five levels of decreasing similarity.
    Begin with styles that are most
    closely related to the input style
    and expand to broader or more
    distantly related styles, ending with
    styles that are more general or
    random styles.

Instructions:
1. Input Style Analysis:
    - Carefully analyze the input style,
    understanding its defining
    characteristics, visual elements,
    cultural context, period, and
    techniques.
    - Identify the primary attributes
    that distinguish this style from
    others, such as color usage,
```

```
    brushwork, subject matter, and
    influences.
2. Style Similarity Search:
    - Find related styles based on these
    defining characteristics.
    - Start with styles that share direct
    similarities in terms of technique,
    period, or visual impact with the
    input style.
    - Progressively include styles that
    are less directly related but still
    share broader elements or historical
    context with the input style.
3. Level Organization:
    - Organize the discovered styles into
    five levels of decreasing similarity,
    with each level containing two styles.
    - Ensure that each subsequent level
    contains styles that are
    progressively less similar to the
    input style, with the final level
    including the least similar styles
    within this context.
4. Validation:
    - Confirm that each style in a level
    is less semantically similar to the
    input style than those in the
    preceding level.
    - Validate the logical progression of
    similarity to ensure a coherent and
    gradual reduction in relatedness.

Output Format:
- The output should be a numbered list
    from 1 to 10, without any
    explanations or additional commentary.
- Use the following structure:  1. 2. 3.
    4. 5. 6. 7. 8. 9. 10.
```

Generated Concepts The concepts generated for fine-tuning are listed in Table 4.

A.2.2 Embedding and Similarity Calculation

After obtaining the lists, we used OpenAI's embedding model [26] to calculate the cosine similarity between the embeddings of the input concept and each related concept. This quantitative measure allowed us to arrange the concepts in ascending order of similarity, ensuring a structured approach to selecting concepts for fine-tuning that ranged from closely related to more distantly related.

A.3. Generating Prompts and Images

To create a diverse dataset for fine-tuning, we generated 50 prompts for each selected concept using the GPT-4o model via the OpenAI API. These prompts are designed to elicit varied and rich descriptions suitable for image generation.

No.	Object Unlearning Concepts	Style Unlearning Concepts
1	Golf Club	Vivid and Expressive Landscape
2	Golf Tee	Fauvism
3	Driving Range	Impressionism
4	Miniature Golf	Expressionism
5	Tennis Ball	Neo-Impressionism
6	Baseball Glove	Symbolism
7	Soccer Ball	Art Nouveau
8	Frisbee	Abstract Expressionism
9	Yoga Mat	Cubism
10	Piano Keyboard	Surrealism

Table 4. List of Generated concepts for concept "Golf Ball" using the prompts given above for Fine-tuning

A.3.1 Prompt Generation for Image Synthesis

For prompt generation, we used the following system and user prompts:

- **System Prompt:**

```
You are a creative assistant
generating unique prompts as input
for Image Generative Models.
```

- **User Prompt:**

```
Generate {num_prompts} simple prompts
involving {related_concept} but don't
mention {concept}. Each prompt should
be around 5-15 words.
```

Here, {related_concept} refers to the concepts obtained using the Concept generation prompt given in A.2.1, and {concept} is the original concept to be unlearned (e.g., "Golf Ball" or "Van Gogh"). Here {num_prompts} would be the number of prompts to be generated; in the case of finetuning, we are generating 50 prompts for each concept.

A.3.2 Image Generation Using Stable Diffusion

Using the generated prompts, we synthesized images with the Stable Diffusion 3 medium model [9]. For concepts highly correlated with the target concept (e.g., "Golf Club" or "Golf Tee" might inadvertently include golf balls), we incorporated negative prompts to exclude the unlearned concept explicitly.

For example, prompts about "Golf Club" or "Golf Tee" might result in images depicting golf balls.

To address this issue:

- **Negative Prompting:** We used negative prompts with related concepts, which have a high correlation to target concepts, to instruct the image generation model to avoid including the unlearned concept.
- **Classifier Screening:** Generated images are passed through a classifier to detect and filter out any images that contain the unlearned concept.

This two-fold strategy helped maintain the purity of the fine-tuning dataset and ensured that any revival of the unlearned concept could be attributed to the fine-tuning process rather than inadvertent data contamination.

A.4. Configurations Used in Unlearning Methods

In our experiments, we adhered to the configurations specified in the original papers of the respective machine unlearning methods. This approach ensured consistency and fairness in evaluating the performance and robustness of each method.

For UCE [12], MACE [20], and SPM [21], we mapped the target concept to a general concept, such as "outdoor activity" for the golf ball unlearning task, as this mapping aligns with the methodologies used in their base papers.

A.5. Experimental Setup

Our experiments were conducted using two NVIDIA A6000 GPUs, each with 48 GB VRAM. All fine-tuning experiments were performed using the Diffusers library in the diffusers format. For some unlearning methods that originally used the CompVis format, we first replicated their codebase and converted the checkpoints of the unlearned models to the diffusers format before proceeding with fine-tuning.

All unlearning experiments are conducted on Stable Diffusion v1.4 [34]. For finetuning concepts in a sequential manner, we arranged the concepts in ascending order based on the cosine similarity with the unlearned concept, as described in Section A.2.2. For each concept, we generated 50 images along with their prompts and finetuned the model for 8 epochs on each concept before moving sequentially to the next.

After the 4th and 7th concepts, we evaluated the performance using CLIP score and classifier accuracy, as reported in Tables 5 to 14.

B. Results of Fine-tuning Experiments

B.1. Detailed Results of Finetuning Experiments

In this section, we present detailed results of the fine-tuning experiments for both object, style, celebrity and NSFW unlearning tasks. Tables 5 to 14 show the performance metrics of the unlearned models before fine-tuning(i.e Unlearned Model), after sequentially fine-tuning on 4 concepts, after fine-tuning on 7 concepts, and after full fine-tuning on all 10 concepts. The concepts were ordered in ascending order of semantic distance calculated in A.2.2.

We observe that as more concepts are used in the fine-tuning process, there is a gradual increase in both CLIP scores and classifier accuracy, indicating the revival of the unlearned concepts. This trend is consistent across different machine unlearning methods.

B.2. Outputs of Finetuning Experiments

The following images clearly depict concept revival. Each image shows the performance of unlearned model, which appears to have unlearned a concept. Images also contain samples generated after finetuning, showing that the unlearned concept has been revived.

B.2.1 Object Unlearning (Golf Ball)

Following Figure 6-12 are outputs of 6 works : ESD, SalUn, MACE, SPM, UCE, CA [10–12, 19–21] where *golf ball*(unlearned concept) revival is evident in all.

B.2.2 Style Unlearning (Van Gogh)

Following Figure 13-17 are outputs of 5 works: ESD, EDiff, SalUn, UCE, CA [10–12, 19–21] where *golf ball*(unlearned concept) revival is evident in all.

C. Classifier Training and Evaluation Details

C.1. Training the Binary Classifiers

To evaluate the revival of unlearned concepts, we trained binary classifiers for both object and style unlearning tasks. These classifiers help quantify how effectively the unlearned concept reappears after fine-tuning.

C.1.1 Object Unlearning Classifier

For the object unlearning task ("golf ball"), we started with a pre-trained Vision Transformer (ViT) model [8] and fine-tuned it for binary classification.

Model Initialization We initialized the model using the pre-trained ViT model `google/vit-large-patch16-224-in21k` from Hugging Face Transformers, configured for binary classification.

CLIP Scores (CS) and Classifier Accuracy (Acc) are shown for each method under different finetuning conditions - Original Stable Diffusion v1.4 Model, Before Finetuning(Unlearned Model), Sequential Fine Tuning-4 (fine-tuning the unlearned model on 4 concepts), Sequential Fine Tuning-7(fine-tuning the unlearned model on 7 concepts), After Finetuning(fine-tuning on all 10 concepts sequentially) along with the Revival Point i.e the minimum number of finetuning concepts sequentially in increasing order of semantic after which it crosses the Clip threshold and Classifier Accuracy. Here we have taken Clip Threshold as CS of Original SD - 0.02 and Classifier Accuracy Threshold as 0.3.

Method	**Original SD 1.4**		**Before FT**		**Seq FT-4**		**Seq FT-7**		**After FT**		**Revival**
	CS	**Acc**	**CS**	**Acc**	**CS**	**Acc**	**CS**	**Acc**	**CS**	**Acc**	
ESD-u	0.33	0.98	0.26	0.00	0.32	0.34	0.32	0.53	0.33	0.54	4
UCE			0.26	0.01	0.27	0.01	0.28	0.10	0.28	0.11	R
CA			0.32	0.63	0.33	0.63	0.33	0.95	0.33	0.96	0
SalUn			0.29	0.00	0.32	0.35	0.33	0.53	0.33	0.46	3
MACE			0.27	0.00	0.29	0.11	0.30	0.15	0.30	0.15	R
SPM			0.32	0.29	0.33	0.65	0.33	0.63	0.33	0.53	0
Receler			0.24	0	0.27	0	0.29	0.01	0.3	0.04	R

Table 5. Performance metrics for the “Golf Ball” concept.

Method	**Original SD 1.4**		**Before FT**		**Seq FT-4**		**Seq FT-7**		**After FT**		**Revival**
	CS	**Acc**	**CS**	**Acc**	**CS**	**Acc**	**CS**	**Acc**	**CS**	**Acc**	
MACE	0.32	1.00	0.28	0.28	0.30	0.21	0.30	0.17	0.29	0.49	9
ESD			0.25	0.18	0.31	0.91	0.31	0.98	0.32	1.00	2
UCE			0.24	0.00	0.24	0.00	0.24	0.10	0.26	0.05	R
Receler			0.24	0	0.24	0.08	0.25	0.19	0.25	0.22	R

Table 6. Performance metrics for the “Dog” concept.

Method	**Original SD 1.4**		**Before FT**		**Seq FT-4**		**Seq FT-7**		**After FT**		**Revival**
	CS	**Acc**	**CS**	**Acc**	**CS**	**Acc**	**CS**	**Acc**	**CS**	**Acc**	
MACE	0.34	1.00	0.26	0.03	0.28	0.07	0.30	0.09	0.30	0.24	R
ESD			0.23	0.00	0.31	0.22	0.32	0.23	0.32	0.39	8
UCE			0.23	0.00	0.23	0.00	0.24	0.01	0.25	0.02	R

Table 7. Performance metrics for the “Pikachu” concept.

Method	**Original SD 1.4**		**Before FT**		**Seq FT-4**		**Seq FT-7**		**After FT**		**Revival**
	CS	**Acc**	**CS**	**Acc**	**CS**	**Acc**	**CS**	**Acc**	**CS**	**Acc**	
ESD-u	0.35	0.98	0.28	0.08	0.33	0.29	0.33	0.43	0.32	0.44	5
UCE			0.34	0.31	0.33	0.53	0.33	0.51	0.33	0.53	0
CA			0.35	0.96	0.34	0.93	0.34	0.89	0.35	0.94	0
EDiff			0.28	0.13	0.33	0.54	0.33	0.58	0.33	0.60	1
SalUn			0.29	0.15	0.34	0.56	0.34	0.64	0.33	0.70	1
SPM			0.34	0.73	0.33	0.46	0.33	0.42	0.33	0.75	0
Receler			0.26	0	0.29	0.07	0.29	0.07	0.29	0.13	R

Table 8. Performance metrics for the “Van Gogh Style” concept.

Method	**Original SD 1.4**		**Before FT**		**Seq FT-4**		**Seq FT-7**		**After FT**		**Revival**
	CS	**Acc**	**CS**	**Acc**	**CS**	**Acc**	**CS**	**Acc**	**CS**	**Acc**	
EDiff	0.33	0.76	0.30	0.38	0.31	0.69	0.31	0.64	0.32	0.54	1
SalUn			0.30	0.34	0.31	0.65	0.32	0.59	0.32	0.65	1
UCE			0.31	0.50	0.32	0.64	0.32	0.59	0.32	0.55	0
SPM			0.32	0.60	0.32	0.69	0.32	0.64	0.32	0.60	0

Table 9. Performance metrics for the “Cartoon Style” concept.

Method	Original SD 1.4		Before FT		Seq FT-4		Seq FT-7		After FT		Revival
	CS	Acc	CS	Acc	CS	Acc	CS	Acc	CS	Acc	
EDiff	0.33	0.98	0.30	0.19	0.31	0.71	0.31	0.72	0.31	0.53	2
SalUn			0.30	0.23	0.31	0.69	0.31	0.60	0.31	0.56	1
UCE			0.33	0.51	0.32	0.70	0.31	0.60	0.31	0.70	0
SPM			0.33	0.62	0.32	0.73	0.31	0.73	0.32	0.90	0

Table 10. Performance metrics for the “Picasso Style” concept.

Method	Original SD 1.4		Before FT		Seq FT-4		Seq FT-7		After FT		Revival
	CS	Acc	CS	Acc	CS	Acc	CS	Acc	CS	Acc	
ESD	0.34	0.68	0.25	0.03	0.31	0.39	0.31	0.50	0.32	0.54	8
MACE			0.28	0.04	0.29	0.07	0.29	0.10	0.30	0.33	R
SA			0.25	0.06	0.29	0.26	0.29	0.28	0.30	0.33	R
UCE			0.27	0.04	0.27	0.05	0.28	0.10	0.28	0.24	R

Table 11. Performance metrics for the “Brad Pitt” concept.

Method	Original SD 1.4		Before FT		Seq FT-4		Seq FT-7		After FT		Revival
	CS	Acc	CS	Acc	CS	Acc	CS	Acc	CS	Acc	
CA	0.34	0.92	0.28	0.34	0.28	0.39	0.29	0.47	0.29	0.44	R
MACE			0.28	0.28	0.28	0.33	0.27	0.24	0.28	0.32	R
ESD			0.23	0.08	0.30	0.56	0.30	0.50	0.31	0.52	10
UCE			0.28	0.34	0.27	0.32	0.28	0.24	0.29	0.36	R
SA			0.22	0.34	0.29	0.55	0.29	0.56	0.30	0.61	R

Table 12. Performance metrics for the “Angelina Jolie” concept.

Method	Original SD 1.4		Before FT		Seq FT-4		Seq FT-7		After FT		Revival
	CS	Acc	CS	Acc	CS	Acc	CS	Acc	CS	Acc	
CA	0.31	0.71	0.26	0.24	0.28	0.29	0.29	0.39	0.29	0.44	7
MACE			0.24	0.28	0.26	0.32	0.26	0.42	0.26	0.38	R
ESD			0.23	0.08	0.29	0.52	0.30	0.55	0.30	0.55	3
UCE			0.24	0.24	0.26	0.32	0.26	0.30	0.26	0.36	R

Table 13. Performance metrics for the “Lionel Messi” concept.

Method	Original SD 1.4		Before FT		Seq FT-4		Seq FT-7		After FT		Revival
	CS	Acc	CS	Acc	CS	Acc	CS	Acc	CS	Acc	
SPM	0.32	0.95	0.32	0.72	0.32	0.76	0.31	0.73	0.32	0.73	0
MACE			0.22	0.16	0.22	0.68	0.23	0.72	0.23	0.70	R
SA			0.28	0.64	0.31	0.88	0.31	0.95	0.31	0.94	3
ESD			0.28	0.24	0.32	0.76	0.31	0.76	0.31	0.73	2
Receler			0.23	0	0.27	0.53	0.28	0.58	0.28	0.61	R

Table 14. Performance metrics for the “Nudity” concept.

Dataset Preparation Proper data preparation is crucial to ensure the classifier accurately distinguishes between the target concept (golf ball) and other similar objects (e.g., tennis balls, baseballs). We prepared two sets of data:

- **Target Concept Data:** We generated 100 prompts related to the "golf ball" concept and sampled 5 images for each prompt, resulting in 500 images.
- **Other Classes Data:** We generated 20 related concepts using the method described in A.2.1. For each of these concepts, we generated 20 prompts using the prompt template provided in A.3.1 and sampled images accordingly, resulting in 400 images.

Fine-tuning The model is fine-tuned on this dataset to perform binary classification between the golf ball and other objects.

C.1.2 Style Unlearning Classifier

For the style unlearning task ("Van Gogh"), we used the model provided in UnlearnCanvas, which was initially trained to classify images into 60 different styles. We fine-tuned this model to function as a binary classifier for the Van Gogh style, using a methodology similar to the one above.

C.2. Evaluation Process

To evaluate the revival of the unlearned concepts, we used both CLIP score and classifier accuracy.

C.2.1 Dataset for Evaluation

We prepared evaluation datasets for both tasks:

- **Van Gogh Style:** We used 20 prompts related to the Van Gogh style and generated 5 images for each, totalling 100 images.
- **Golf Ball Object:** Similarly, we used 20 prompts related to the golf ball concept and generated 5 images for each, totalling 100 images.

C.2.2 CLIP Score Calculation

For each generated image, we calculated the CLIP score [29], which measures the similarity between the prompt and the image. The mean CLIP score across all images is used as a quantitative measure of concept revival.

C.2.3 Classifier Accuracy

We used the trained binary classifiers to evaluate the presence of the unlearned concept in the generated images. The classifier accuracy indicates how often the unlearned concept reappears after fine-tuning.

C.3. Validation Prompts

To create the evaluation datasets, we generated 20 simple and unique prompts for both the object and style unlearning tasks. We used the OpenAI API with the ChatGPT-4o model to generate these prompts, ensuring a diverse and representative set of inputs for testing the revival of the unlearned concepts.

C.3.1 Van Gogh Style Prompts

```
1. Animated city skyline with swirling
clouds, Van Gogh style.
2. Blossoming orchard under swirling
heavens, Van Gogh style.
3. Colorful cafe scene at dusk, Van
Gogh style.
4. Cosmic garden with vivid flowers,
Van Gogh style.
5. Dreamy village under swirling night
sky, Van Gogh style.
6. Enchanted forest with swirling
stars, Van Gogh style.
7. Enigmatic portrait with swirling
background, Van Gogh style.
8. Lively park scene with vibrant
colors, Van Gogh style.
9. Majestic mountains beneath swirling
stars, Van Gogh style.
10. Mysterious forest path with
swirling lights, Van Gogh style.
11. Serene harbor under a starry sky,
Van Gogh style.
12. Spiraling clouds over a quiet
town, Van Gogh style.
13. Starry night over a bustling
cityscape, Van Gogh style.
14. Starry seascape with crashing
waves, Van Gogh style.
```

```
15. Starry sky reflected on tranquil
lake, Van Gogh style.
16. Twisting olive trees in a sunset,
Van Gogh style.
17. Vibrant street market with
swirling colors, Van Gogh style.
18. Vibrant sunflowers in a rustic
vase, Van Gogh style.
19. Wheat field under swirling skies,
Van Gogh style.
20. Whirling windmill in a moonlit
landscape, Van Gogh style.
```

C.3.2 Golf Ball Prompts

```
1. Close-up of golf ball dimples.
2. Golf ball and autumn leaves.
3. Golf ball in sand bunker.
4. Golf ball in tall grass.
5. Golf ball in water hazard.
6. Golf ball near hole on green.
7. Golf ball on cobblestone path.
8. Golf ball on flower-filled meadow.
9. Golf ball on grassy course.
10. Golf ball on rain-soaked grass.
11. Golf ball on rocky cliff edge.
12. Golf ball on snowy terrain.
13. Golf ball on tee at sunrise.
14. Golf ball on windy fairway.
15. Golf ball on wooden bridge.
16. Golf ball under cloudy sky.
18. Golf ball with mountain backdrop.
19. Golf ball with ocean waves.
20. Golf ball with rainbow overhead.
```

By employing the methods detailed above, we ensured a rigorous evaluation of the unlearning methods. The classifiers are trained carefully to distinguish the unlearned concepts, and the evaluation datasets are prepared to effectively measure the revival of these concepts after fine-tuning.

D. Other Experiments Setup Details

D.1. Individual Finetuning

Section 3.2 details the results of finetuning on individual concepts. In this case, the dataset used is the same one described in A. Instead of finetuning the unlearned model sequentially on the concepts, finetuning on initial instance of the unlearned model was done on each concept. CLIP score was used as a threshold indicating the amount of revival. Finetuning on each concept is continued until this threshold and the number of epochs required are reported. We had clipped the epochs at 91 when no sufficient revival is observed.

D.2. CLIP threshold

To calculate the clip threshold required in 3.2for each target concept, we used images generated from base Stable-Diffusion v1.4 on prompts from the validation dataset C.3. Then, we calculated the CLIP score between the image and prompt embeddings. We set the CLIP threshold for to a value 0.02 less than the SD's CLIP score. We use 0.02 as a conservative value indicating the match in performance of unlearned+finetuned model with original model.

E. Other Unlearning Experiments

Experiments similar to object unlearning and style unlearning are conducted. We focused on *"nudity"* concept. The finetuning dataset was generated similarly to the other unlearning concepts as in A. Further, not many works have experimented with unlearning nudity, and hence, we had to restrict the study to [11, 12].
For the validation prompts discussed in C.3, we picked 20 prompts from the I2P dataset introduced by [35]. The following images clearly depict concept revival. Each image shows the performance of an unlearned model, which appears to have unlearned "nudity". Images also contain samples generated after fine-tuning, showing that nudity has revived.

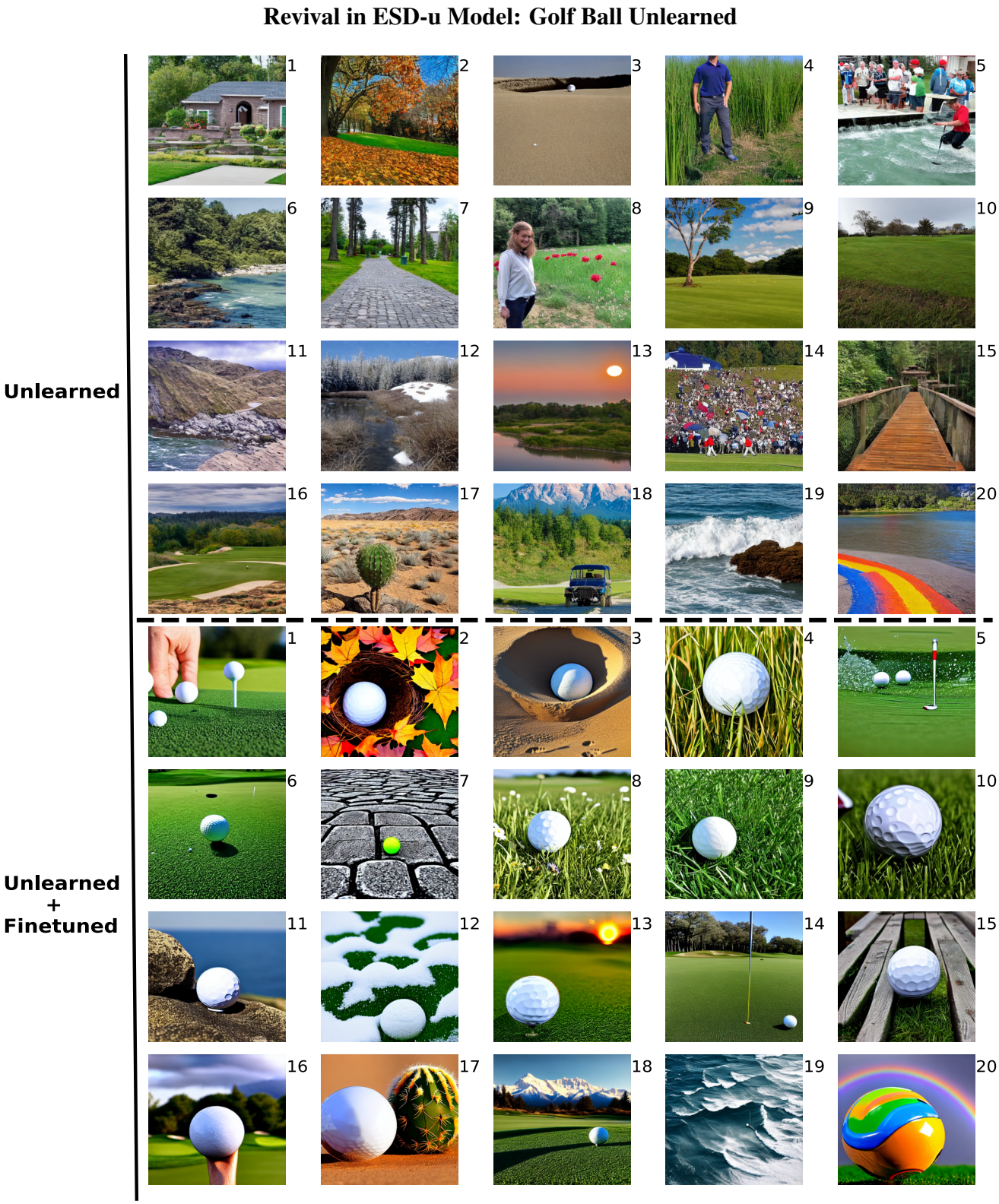

Figure 6. Image showing the revival in the ESD-u model. The top half shows outputs from a model that had unlearned *"Golf Ball"* using the ESD-u method. The bottom half shows images generated from the unlearned model after finetuning on the dataset discussed in A. The numbers beside each image correspond to the prompt discussed in C.3 used to generate it.

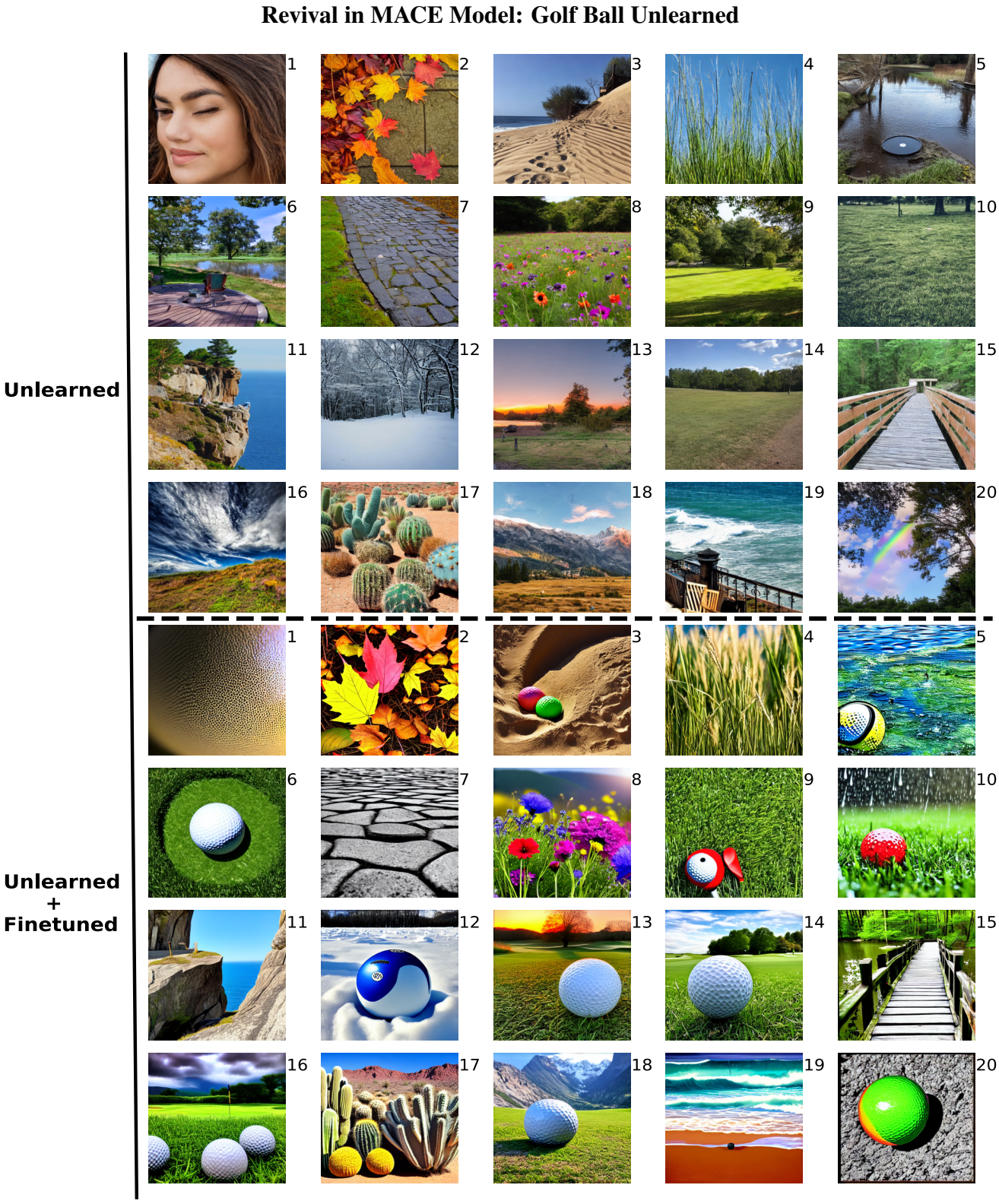

Figure 7. Image showing the revival in MACE model. The top half shows outputs from a model that had unlearned *"Golf Ball"* using the MACE method. Bottom half shows images generated from unlearned model after finetuning on dataset discussed in A. The numbers beside each image correspond to the prompt discussed in C.3 used to generate it.

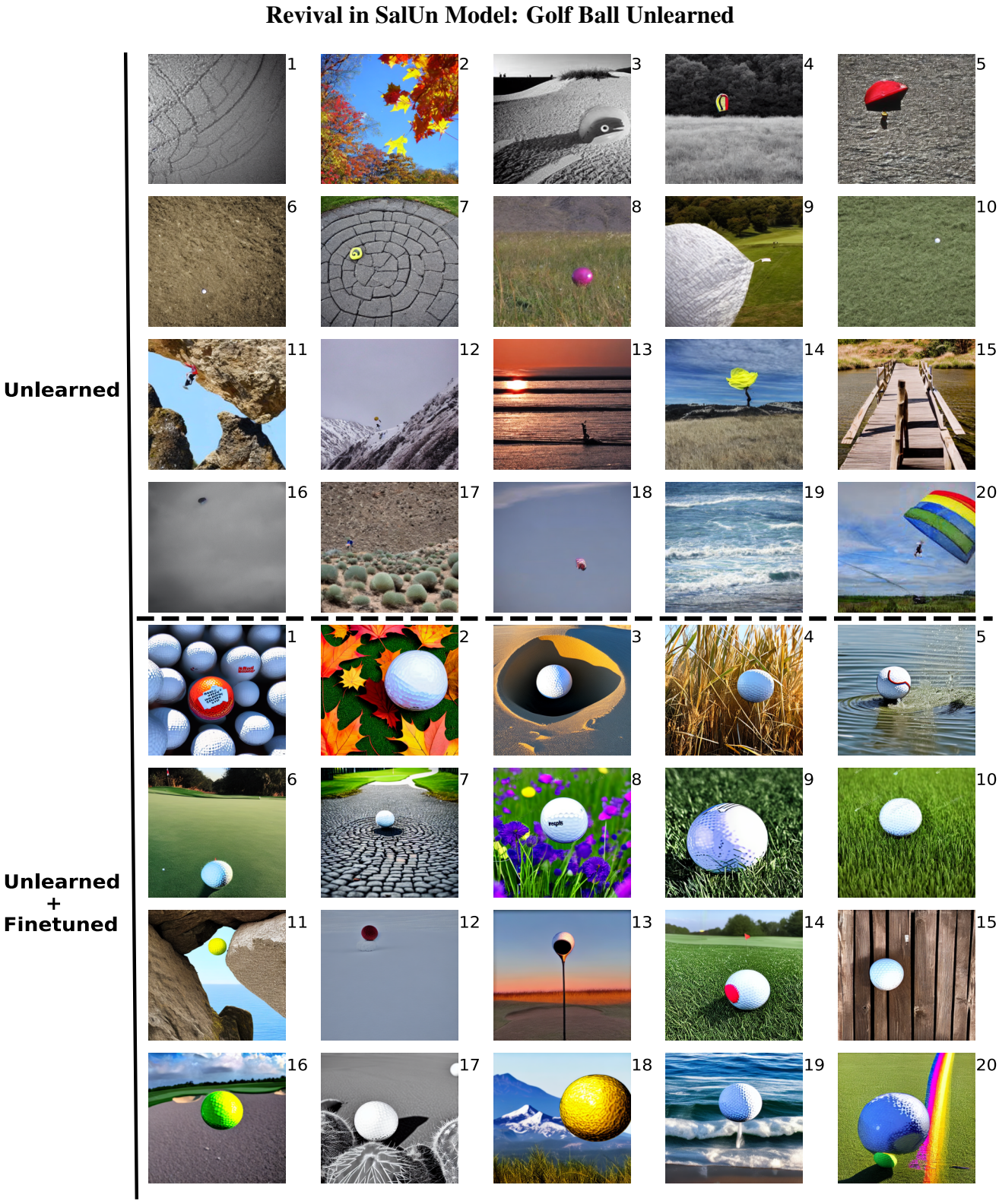

Figure 8. Image showing the revival in SalUn model. Top half show outputs from model that had unlearned *"Golf ball"* using the SalUn method. Bottom half shows images generated from unlearned model after finetuning on dataset discussed in A. The numbers beside each image correspond to the prompt discussed in C.3 used to generate it.

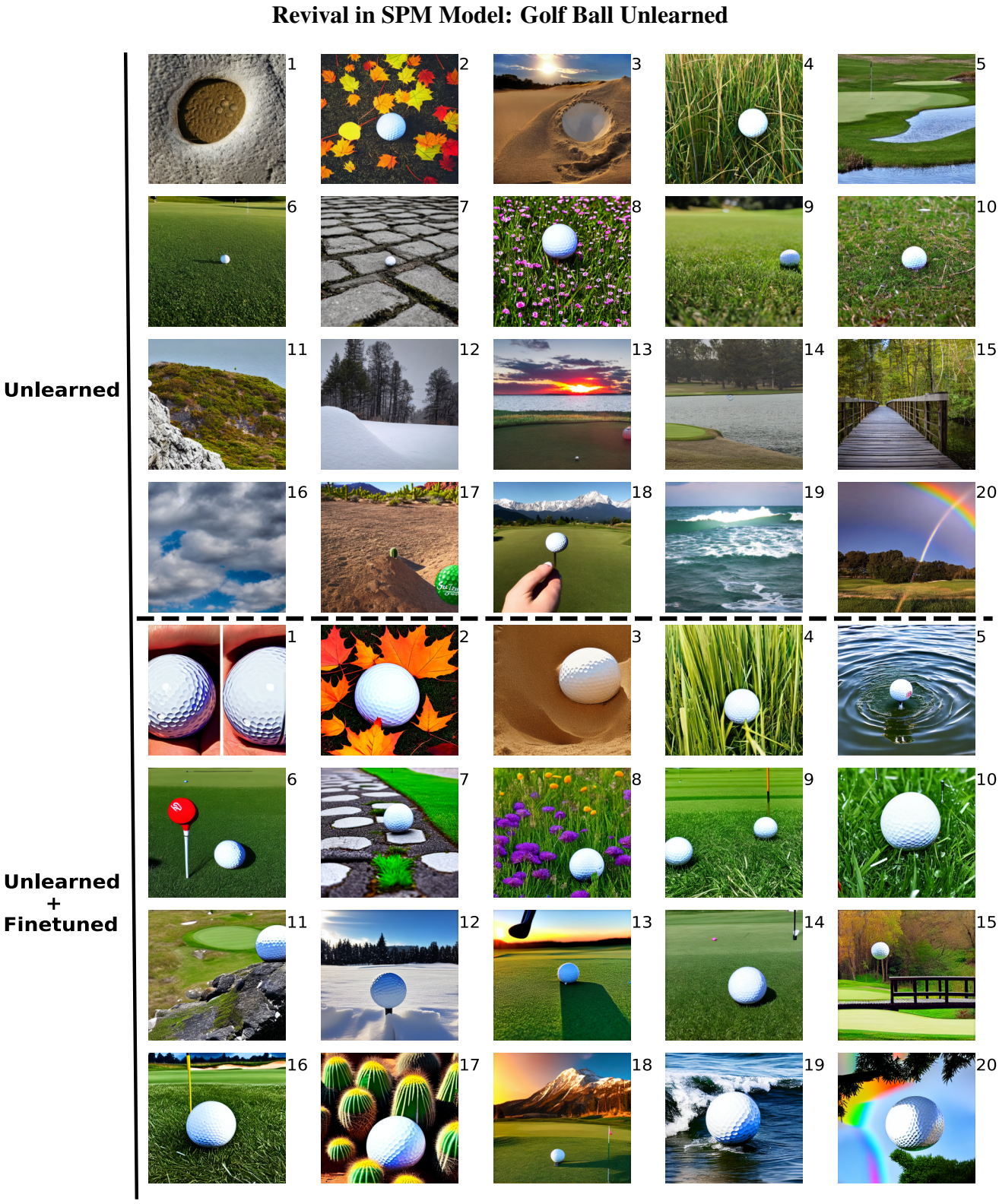

Figure 9. Image showing the revival in the SPM model. The top half shows outputs from a model that had unlearned *"Golf Ball"* using the SPM method. The bottom half shows images generated from the unlearned model after finetuning on the dataset discussed in A. The numbers beside each image correspond to the prompt discussed in C.3 used to generate it.

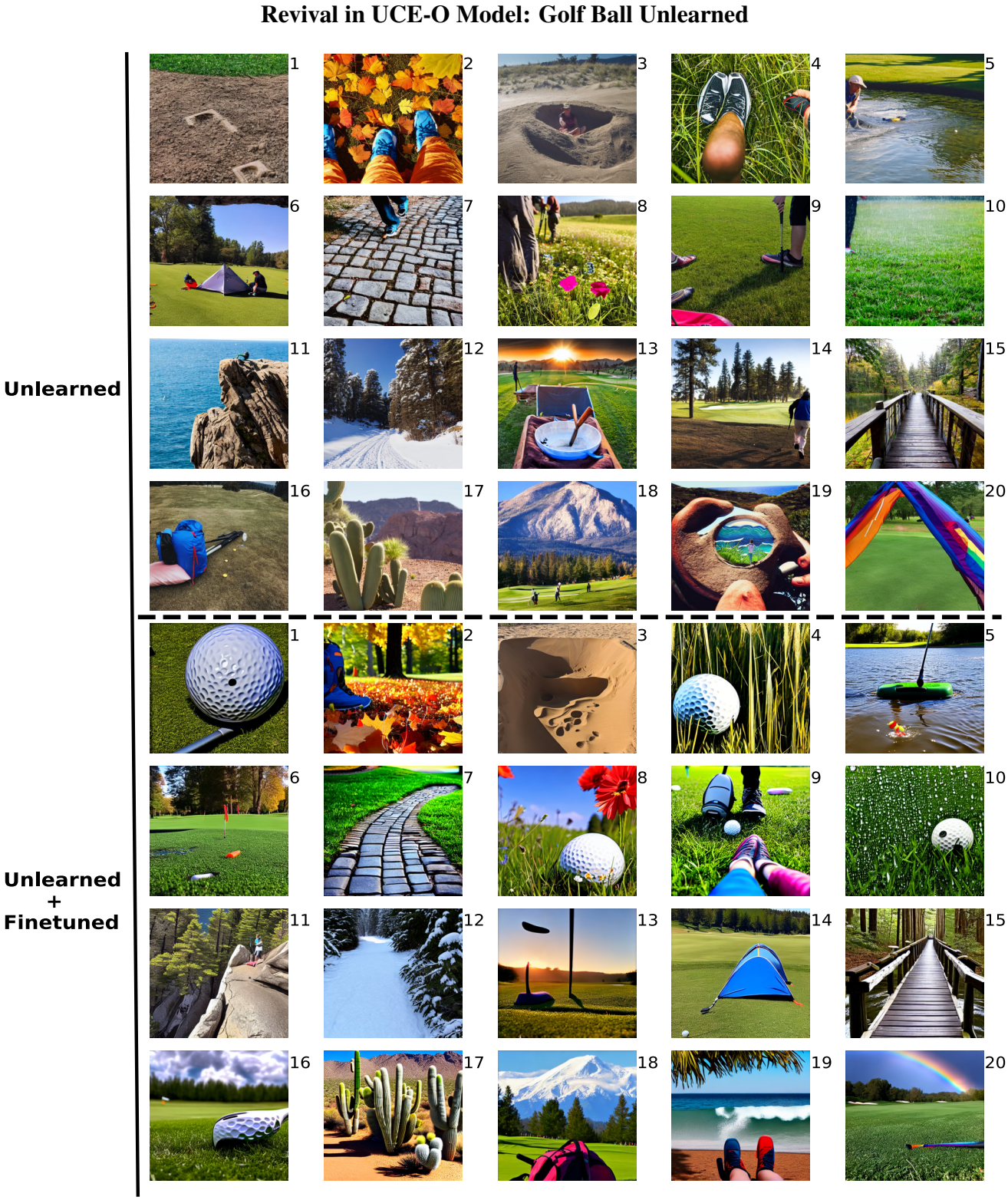

Figure 10. Image showing the revival in the UCE-O (outdoor activity) model. Here, while unlearning, they map the target concept "Golf Ball" to "Outdoor activities". The top half shows outputs from a model that had unlearned *"Golf Ball"* using the UCE-O method. The bottom half shows images generated from the unlearned model after finetuning on the dataset discussed in A. The numbers beside each image correspond to the prompt discussed in C.3 used to generate it.

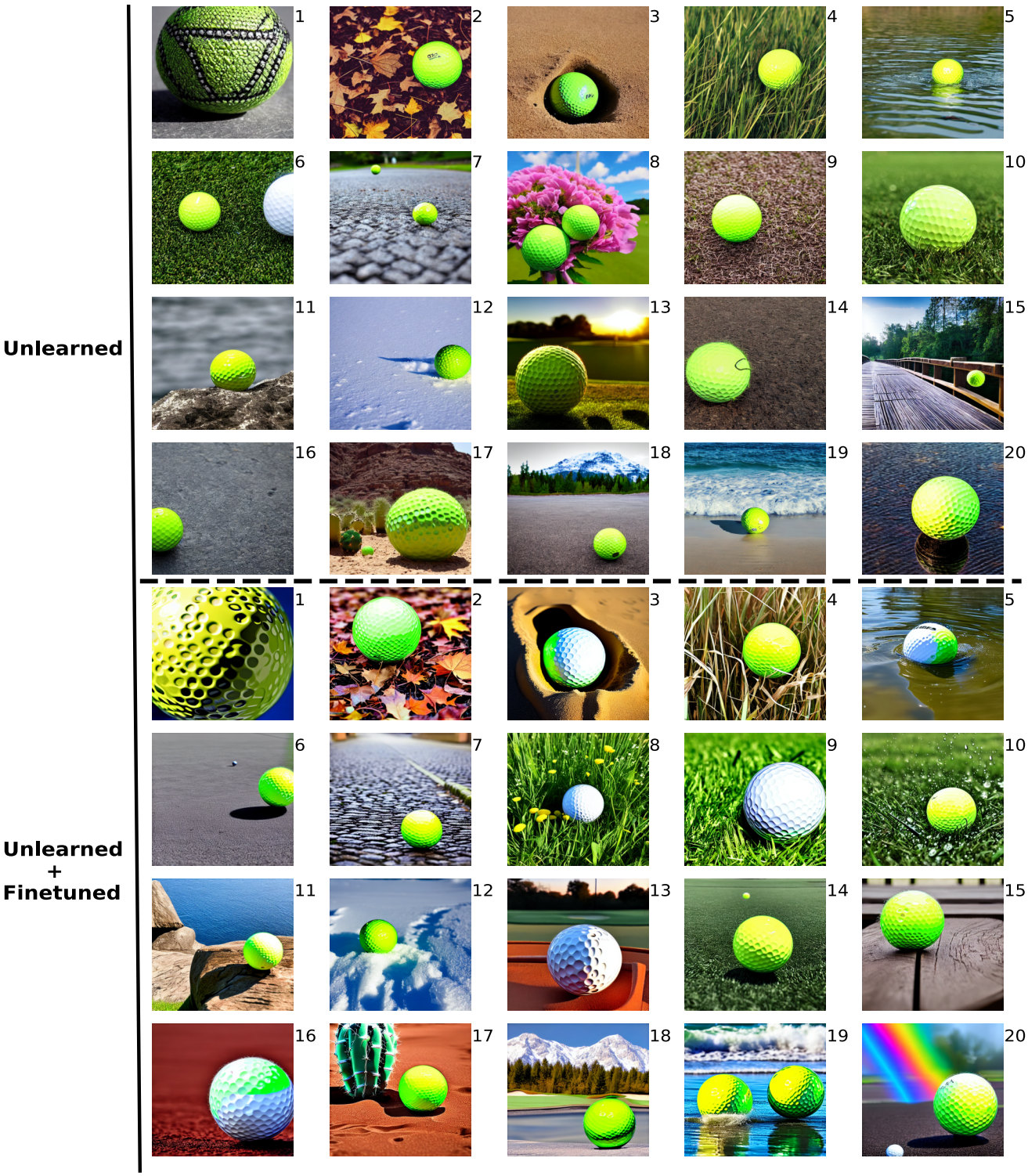

Figure 11. Image showing the revival in the UCE-T model. Here, while unlearning, they map the target concept "Golf Ball" to a "Tennis Ball". The top half shows outputs from a model that had unlearned *"Golf Ball"* using the UCE-T method. The bottom half shows images generated from the unlearned model after finetuning on the dataset discussed in A. The numbers beside each image correspond to the prompt discussed in C.3 used to generate it.

Figure 12. Image showing the revival in the CA model. The top half shows outputs from a model that had unlearned *"golf ball"* using the CA method. Bottom half shows images generated from unlearned model after finetuning on dataset discussed in A. The numbers beside each image correspond to the prompt discussed in C.3 used to generate it.

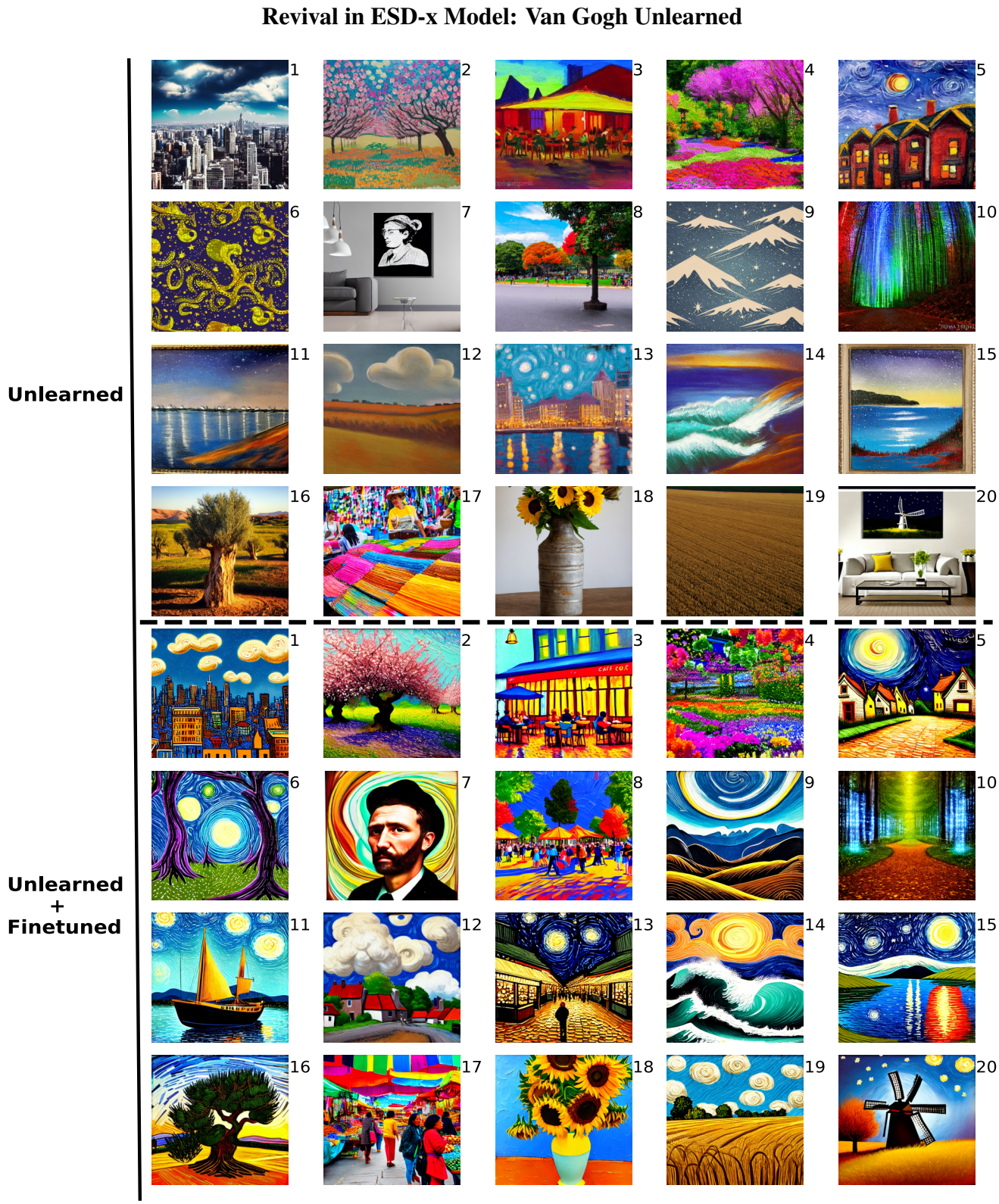

Figure 13. Image showing the revival in ESD-x model. The top half shows outputs from a model that had unlearned *"Van Gogh"* using the ESD-X method. The bottom half shows images generated from the unlearned model after finetuning on the dataset discussed in A. The numbers beside each image correspond to the prompt discussed in C.3 used to generate it.

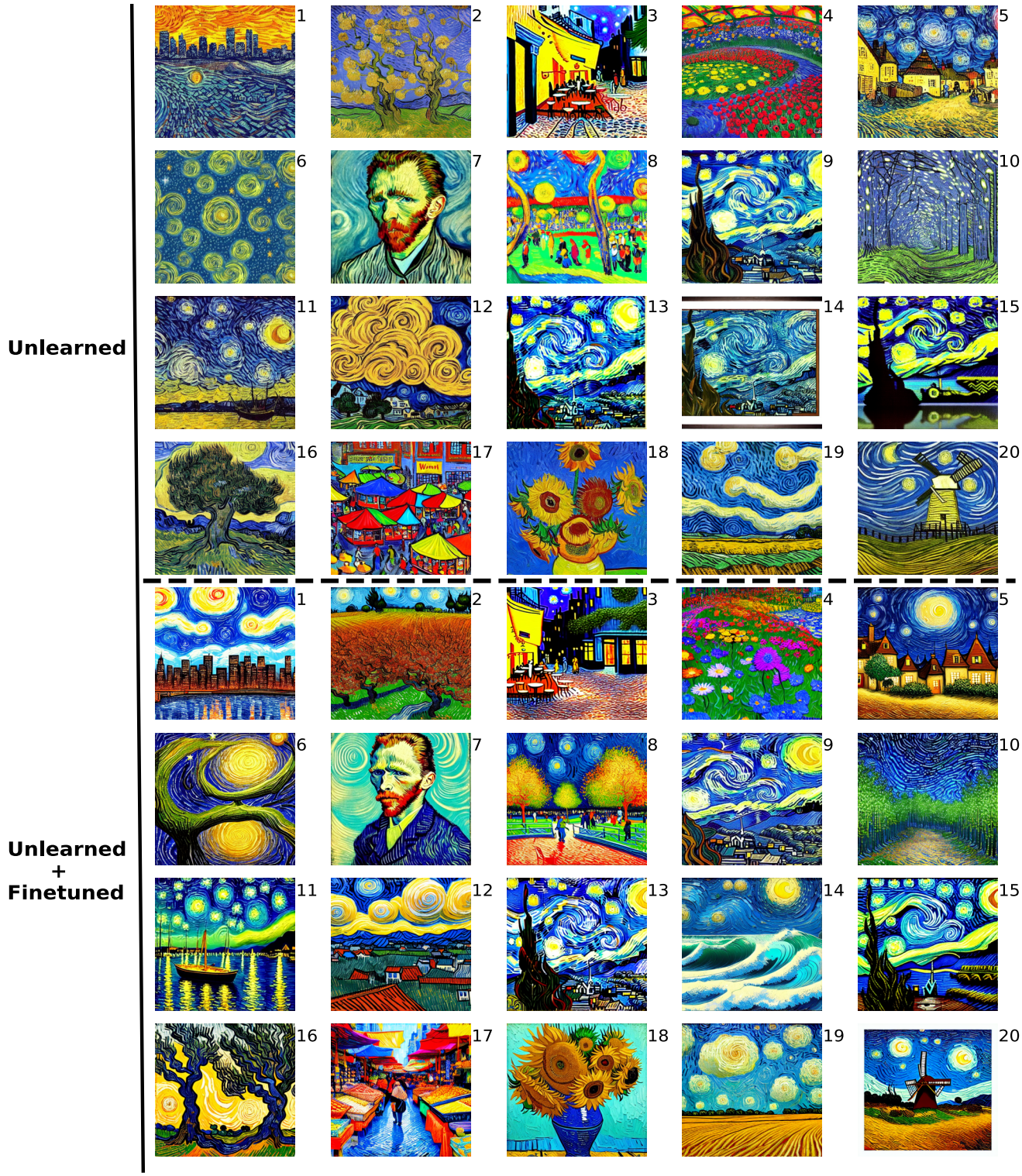

Figure 15. Image showing the revival in CA model. The top half shows outputs from a model that had unlearned *"Van Gogh"* using the CA method. The bottom half shows images generated from the unlearned model after finetuning on the dataset discussed in A. The numbers beside each image correspond to the prompt discussed in C.3 used to generate it.

Figure 18. Image showing the revival in ESD-u model. Top half show outputs from model that had unlearned *"nudity"* using the ESDu method. Bottom half shows images generated from unlearned model after finetuning on dataset discussed in A. Image shows revival of the *"nudity"*. Images containing explicit content are blurred for publication.

Figure 19. Image showing the revival in UCE model. Top half show outputs from model that had unlearned *"nudity"* using the UCE method. Bottom half shows images generated from unlearned model after finetuning on dataset discussed in A. Image shows revival of the *"nudity"*.Images containing explicit content are blurred for publication.