A. Discussion

A.1. Limitations

One limitation of `autoeval-unlearning` is that it is difficult to test on existing unlearning benchmarks. Because our approach specifically avoids fixed forget and retain sets, instead focusing on generating evaluation data from an unlearning target, it is not applicable to benchmark datasets. We chose to develop our `autoeval-unlearning` tool to evaluate unlearning of concepts from generative models, as this setting more accurately reflects unlearning scenarios in the real world. Another limitation is that we cannot evaluate all possible concepts related to an unlearning target. This is a problem inherent to machine unlearning, and our work represents a step toward filling that gap. We provide a wide variety of examples to demonstrate that our results hold across domains, modalities, and relationship types.

Another limitation of `autoeval-unlearning` is that it is dependent on the performance of the assistant LLM. Our experimentation demonstrates that even relatively small LLMs are of sufficient quality that they can reliably be used across a wide variety of general target concepts. However, targets that come from specialized domains, such as specialized medical terms, might prove problematic. However, this limitation is offset due to the fact that `autoeval-unlearning` is LLM agnostic. If this scenario were encountered, a stronger, or domain specific LLM could easily be swapped in to produce higher quality evaluation plans. Another approach that could be considered is in-context learning, where the assistant LLM could be adapted to improve performance on the target with contextual information or a few curated examples.

A.2. Broader Impacts and Ethics Statement

Machine unlearning aims to remove the influence of potentially harmful data from a model. Our `autoeval-unlearning` tool enables more thorough evaluation of the effectiveness, extent, and robustness of concept unlearning in generative models. We hope that `autoeval-unlearning` can advance the usefulness of machine unlearning in real-world scenarios. Unlearning as a whole has some potentially negative societal impacts, such as model degradation or over-confidence that unlearning has been successfully completed. However, our work focuses on evaluating these concerns, and therefore we do not foresee any negative societal impacts. Our paper conforms with the NeurIPS Code of Ethics. We only use publicly-available models and data generated from those models.

B. System Prompts for `autoeval-unlearning`

We use LLM-powered modules which create, process, and filter information for `autoeval-unlearning`. These processes include our `brainstorm` function for concepts related to the unlearning target and additional generation of evaluation data. This section includes our carefully engineered system prompts for these LLM modules.

B.1. `brainstorm` hypernyms

```
The user will provide a *target* concept and may also provide hints or
instructions within brackets {{}}.  You will then brainstorm potential hypernyms
of the target concept (things, personas, ideas, categories) and list them on new
lines.  Do not provide descriptions or rationales for the concepts.  Simply list
n or fewer (each on a new line, without numbering or punctuation), and output
nothing else.  If there are fewer than n, it is ok.  The outputs should be diverse
and unambiguous when taken out of context.
```

B.2. `brainstorm` siblings

```
The user will provide a *target* concept and may also provide hints or
instructions within brackets {{}}.  You will then brainstorm highly related
or similar concepts to the target concept (things, personas, ideas) and list
them on new lines.  The concepts should be separate from the target concept but
clearly related or similar to the target concept.  Do not provide descriptions or
rationales for the concepts.  Simply list n or fewer (each on a new line, without
numbering or punctuation), and output nothing else.  If there are fewer than n, it
is ok.  The outputs should be diverse and unambiguous when taken out of context.
```

The generation of *siblings* is conditioned on *hypernyms* by supplying the *hypernyms* through the *hint* option in the above prompt.

B.3. `brainstorm` attributes

```
The user will provide a *target* concept and may also provide hints or
instructions within brackets {{}}.  You will then brainstorm parts or attributes
of the *target* concept.  P̈artsc̈an be in a literal, physical sense, or p̈artsc̈an be
interpreted in a more abstract sense.  Do not provide descriptions or rationales
for the parts - just list the parts and nothing else.  Simply list n or fewer
(each on a new line, without numbering or punctuation), and output nothing
else.  If there are fewer than n, it is ok.  The outputs should be diverse and
unambiguous when taken out of context.
```

B.4. `brainstorm` coincidental

```
The user will provide a *target* concept and may also provide hints or
instructions within brackets {{}}.  You will then brainstorm concepts or things
that are not the *target* concept, but which are concepts, things, or actions
associated closely with the *target* concepts.  These concepts, things, or actions
should appear in times, places, or situations in which the *target* concept is
relevant.  Do not provide descriptions or rationales for the coincidental concepts
- just list the concepts and nothing else.  Simply list n or fewer (each on a new
line, without numbering or punctuation), and output nothing else.  If there are
fewer than n, it is ok.  The outputs should be diverse and unambiguous when taken
out of context.
```

B.5. `brainstorm` instances

```
The user will provide a *target* concept and may also provide hints or
instructions within brackets {{}}.  You will then brainstorm specific instances
of the *target* concept.  These instances must fall under the category of the
*target* concept.  If there are no specific instances of the *target* concept,
output nothing.  Do not provide descriptions or rationales for the listed
instances - just list the instances of the *target* concept and nothing else.
Simply list n or fewer(each on a new line, without numbering or punctuation), and
output nothing else.  If there are fewer than n, it is ok.  The outputs should be
diverse and unambiguous when taken out of context.
```

B.6. `reorder` function prompt

```
The user will provide a list of concepts, with each concept on a new line.  The
first entry (first line) is the *target* concept, and this target concept may also
have hints or instructions within brackets {{}}.  Reorder the list based on its
relevance and similarity with the target concept, and output each (reordered)
entry on a new line.  The first line of your response should be the target
concept.  Do not provide descriptions or rationales for the reordering.  Simply
reorder the list (each on a new line), and output nothing else.
```

B.7. `deduplicate` function prompt

```
The user will provide a list of concepts, each on a new line.  You will consider
each element of the list and re-output it on a new line if it has not already
been referred to earlier in the list.  Remove duplicate versions of elements
which refer to the same concept as another element (keep the element that is most
specific of the duplicates).  Do not provide descriptions or rationales for your
decision-making.  Simply list the de-duplicated entries on new lines and nothing
else.
```

B.8. Process for Determining *Uniqueness* of Attributes

Uniqueness of an attribute is determined by prompting the assistant LLM with a series of questions which probe potential ways that an attribute can be *unique* to a target concepts. We found it necessary to assess *uniqueness* with fine-grained, direct questions in order to align the assistant LLM with our definition of *uniqueness*. General queries regarding *uniqueness* were too vague and resulted in misalignment.

Given a `target` and an attribute `attr`, we first ask the assistant LLM to confirm that `attr` is a *part* of `target`. If the response is "yes," we proceed to query the assistant LLM with the following prompts (in this order):

1. `Generally speaking, was ``{attr}'' (as a general concept) created *exclusively* for or by ``{target}''? Answer 'yes' or 'no' only.`
2. `Generally speaking, is ``{attr}'' (as a general concept) financially owned by ``{target}'' exclusively? Answer 'yes' or 'no' only.`
3. `Generally speaking, does ``{attr}'' refer to an idea, character, persona, or place associated *exclusively* with ``{target}''? Answer 'yes' or 'no' only.`
4. `Generally speaking, does ``{attr}'' (as a general concept) exist due to ``{target}'' *exclusively*? Answer 'yes' or 'no' only.`
5. `Generally speaking, is ``{attr}'' (as a general concept) a member of the category ``{target}'' *exclusively*? Answer 'yes' or 'no' only. If ``{target}'' is not a group or category or ``{attr}'' is not exclusive to it, answer 'no'.`

If the answer to any of those questions is "yes," or the string `attr` contains the string `target`, we consider `attr` to be *unique* to `target`. Otherwise, we consider `attr` to be *non-unique* to `target`.

B.9. Adversarial Captions (Diffusion Model Experiments)

The adversarial captions for each unlearned model are generated in a two-step process. First, the models are evaluated with CLIP *target* prediction rate (vs. *sibling* concepts) when generating images of the *unique* attributes for the target concept. Second, the top two *unique* attributes (in terms of CLIP *target* prediction rate) are identified and used to condition generation of *adversarial captions*. We generate adversarial captions with the following prompt:

```
The user will provide a *target* concept and may also provide hints or
instructions within brackets {{}}. You will then brainstorm potential detailed,
single-sentence captions of images that contain, describe, or elicit the target
concept. Make the captions detailed and rich with visual information. Do not
provide descriptions or rationales for the captions.
```

Caption generation is conditioned on the two *unique* attributes by providing them through the *hint* option in the above prompt.

B.10. LLM Unlearning Question-Answer Pairs

We use the following prompt for generating question-answer unlearning pairs related to an item `item`:

```
You are a helpful AI that generates question-answer pairs about {item}.
Generate {n} diverse question-answer pairs that cover different aspects of
{item}.
Your pairs should be detailed, informative, and cover a range of knowledge
about {item}.
Format each pair as:
Q: [Question]
A: [Answer]
Separate each question-answer pair with a blank line.
```

B.11. LLM General Knowledge Question-Answer Pairs

We use the following prompt for generating general knowledge question-answer pairs:

```
You are a helpful AI that generates general knowledge question-answer pairs.
Generate {n} diverse question-answer pairs about random topics that test
general knowledge.
```

```
Cover a wide range of subjects like history, science, geography, arts,
literature, sports, etc.
Your pairs should be detailed, informative, and test different types of
knowledge.
Don't focus on any specific theme - make them as diverse as possible.
Format each pair as:
Q: [Question]
A: [Answer]
Separate each question-answer pair with a blank line.
```

C. Experiment Details

C.1. Text-to-Image Unlearning

For all text-to-image unlearning experiments, we used ESD [21] and Receler [26] with default parameters on Stable Diffusion v1.4 [57]. Our implementations of ESD+ and REC+ modify these methods by adding all sibling concepts to the retain set and all *unique* attributes to the forget set. Note that the sibling concepts and *unique* attributes input to ESD+ and REC+ were separately generated from the ones generated for evaluation.

For the experiment in Figure 4, we generated 10 siblings to the unlearning target ("Formula 1 car") using `brainstorm` prompt B.2. Within each concept, we prompt the models with that concept 30 times and compute KID between the images generated by the original and unlearned models. For each unlearning target in Table 1 we generate up to 10 *non-unique* and up to 10 *unique* attributes using prompts B.3 and B.8. KID values are again averaged over 30 images for each concept (from both the base and unlearned model). For the experiment in Figure 3, we generate 10 siblings and 7 *unique* attributes to the unlearning target ("Star Wars") using prompts B.2 and B.3. The original and unlearned models are prompted with each *unique* attribute 30 times, and each output is classified by OpenAI CLIP (`openai/clip-vit-base-patch32`) [52] with the sibling concepts as zero-shot classes. Figure 5 shows CLIP prediction rates of images generated from the sibling concepts themselves, up to 10 per concept.

C.2. LLM Unlearning

For all LLM unlearning experiments, we use the `open-unlearning` toolkit [18] to perform unlearning. For each unlearning technique, we use the default hyperparameters, except we run for 20 epochs instead of 10. This includes a training batch size of 8, an evaluation batch size of 16, 4 gradient accumulation steps, a learning rate of 1e-5 with the AdamW optimizer [40], and no weight decay. For the Gradient Difference and NPO methods, we equally weighted the forget and retain loss, with `open-unlearning` alpha and gamma values of 1.0. For NPO we used a beta value of 0.1 for the DPO temperature parameter.

When generating question-answer pairs using prompts B.10 and B.11, we use a temperature of 0.9, which would then be randomly adjusted by randomly sampled values ranging from -0.1 to 0.2 over the course of iteration. We use a top p of 0.92 and a top k of 50. We set the maximum number of tokens to 200 and the maximum number of iteration attempts to 50. We aimed to generate 10 question-answer pairs for each item in the evaluation plan, 50 target pairs, and 100 general knowledge pairs.

C.3. Compute Resource Information

All evaluation plans (brainstorming process) leveraged three NVIDIA H100 GPUs. Most brainstorming experiments used Llama-3.3-70B-Instruct as the base model [22]. However, we also ran a few brainstorming experiments with Llama-3.1-8B-Instruct and Llama-3.2-90B-Vision-Instruct.

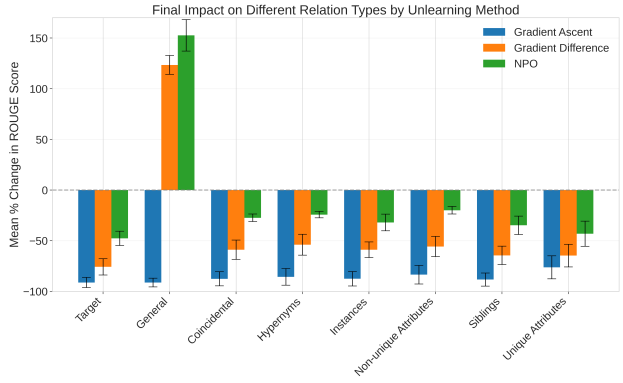
All of the diffusion model unlearning processes (e.g., ESD or REC) leveraged a single NVIDIA H100 GPU. All the diffusion model unlearning evaluations leveraged two H100 GPUs - one for the base model and one for the unlearned model.

All of the LLM experiments, including unlearning and question-answer pair generation, were performed using a single H100 GPU.

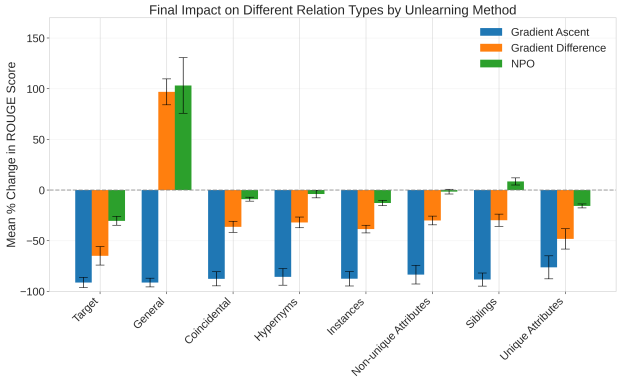
D. Additional Experimental Results

D.1. LLM Unlearning Results with Larger Model

Figure 7 contains the same unlearning experiments as Figure 6 in the main paper, except we perform unlearning on a larger model, Llama 3.1-8B-Instruct. Generally, performance on the retain set increases more dramatically, while performance on



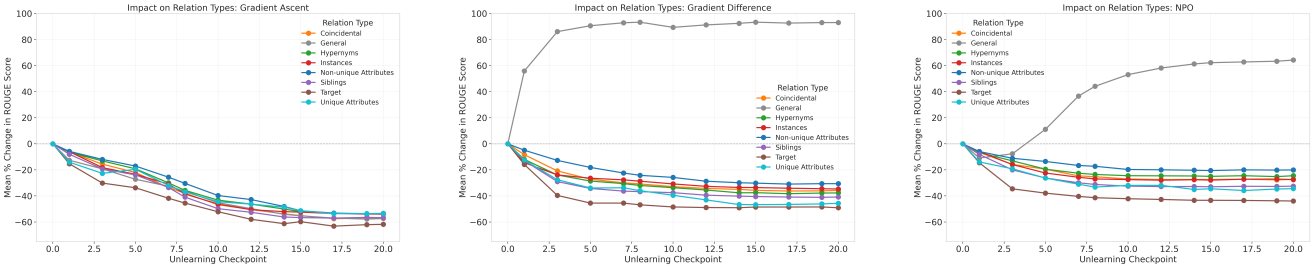
(a) Forget Set: Target. Retain Set: General Knowledge.



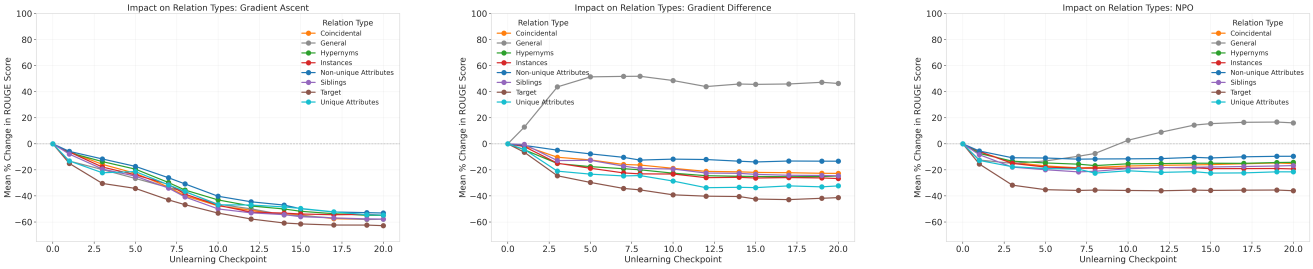
(b) Forget Set: Target. Retain Set: General Knowledge + Siblings.

Figure 7. Average change in ROUGE score, averaged over 8 concept plans, for various relation types. (a) illustrates the degree to which neighboring concepts are affected when unlearning a specific target. (b) demonstrates how including a corpus from one relation type (siblings) can enhance knowledge retention of other neighboring concepts.

the forget set decreases more noticeably. We hypothesize that these changes could be due to the larger model being more prone to overfitting during fine-tuning and the better performance prior to unlearning allowing more room for a performance drop. The gradient ascent and gradient difference methods do not exhibit significantly different behavior; however, NPO does seem noticeably different. In Figure 7a, NPO outperforms gradient difference on the retain set, while also reducing its unlearning effectiveness on the target. It does remain more performant on the other relation types compared to the smaller model. This is further seen in Figure 7b, where including the siblings in the retain set does unfortunately reduce the effectiveness of unlearning on the target but results in the other relation types being minimally impacted (the sibling performance



(a) Forget Set: Target. Retain Set: General Knowledge.



(b) Forget Set: Target. Retain Set: General Knowledge + Siblings.

Figure 8. Average change in ROUGE score across concept plans for various relation types relative to the number of unlearning epochs. (a) shows performance for unlearning the target with a general knowledge retain set, while (b) shows unlearning the target with general knowledge + siblings as the retain set.

even increases slightly). This suggests that the way that knowledge is encoded and entangled within different LLMs does vary, altering the impacts of unlearning on various knowledge types. Future work could focus on better understanding how knowledge retention changes with respect to models of various types and sizes.

D.2. Duration of LLM Unlearning

Figure 8 shows the change in ROUGE score relative to the number of unlearning epochs. The top plots (Figure 8a) contain the results for each unlearning method with only the general knowledge retain set. The bottom plots (Figure 8b) contain the results for each method with the general knowledge + sibling retain set. For gradient ascent, we see that each of the relation types consistently degrade over the course of unlearning until epoch 15 when they bottom out, obtaining performance close to zero. Gradient difference and NPO both behave somewhat similarly, where the general set performance increases while the others decrease. Performance changes more gradually in the case of NPO, where for gradient difference there is a sharp jump in performance over the first few epochs, especially on the general set. Performance changes are also more gradual for the general knowledge + sibling experiments compared to just the general experiments.

E. Examples

E.1. Evaluation plans

The evaluation plans are a critical part of `autoeval-unlearning`. We provide a few examples of evaluation plans generated using our tool in this section.


```
{
    "target": "Star Wars",
    "synonyms": [],
    "attributes_nonunique": [
        "Laser Gun Fights",
        "Space Battles",
        "Robots",
        "Spaceships",
        "Galactic Empires"
    ],
    "attributes_unique": [
        "Jedi Knights",
        "Darth Vader",
        "Obi-Wan Kenobi",
        "Sith Lords",
        "Lightsabers",
        "Chewbacca",
        "R2-D2"
    ],
    "hypernyms": [
        "Space Opera",
        "Galactic Saga",
        "Saga",
        "Film Franchise",
        "Film Series",
        "Franchise",
        "Science Fiction",
        "Fantasy Universe",
        "Epic Storytelling",
        "Universe"
    ],
    "siblings": [
        "Star Trek",
        "Stargate SG-1",
```

```
        "Sci-fi Stargate",
        "Guardians of the Galaxy",
        "Farscape",
        "Battlestar Galactica",
        "Browncoat spaceship Firefly",
        "Bioware's Mass Effect",
        "Babylon 5",
        "Halo (video game)"
    ],
    "instances": [
        "Darth Vader",
        "Luke Skywalker",
        "Millennium Falcon",
        "Obi-Wan Kenobi",
        "Yoda, a fictional character",
        "Chewbacca",
        "Boba Fett",
        "Jabba the Hutt",
        "C-3PO",
        "R2-D2"
    ],
    "coincidental": [
        "Jedi",
        "Lightsabers",
        "Galactic empires",
        "Bounty hunters",
        "Space battles",
        "Spaceship",
        "Blasters",
        "Space stations",
        "Galaxy",
        "Droid"
    ]
}

{
    "target": "Harry Potter",
    "synonyms": [],
    "attributes_nonunique": [
        "Broomsticks",
        "Magic wands",
        "Wands",
        "Spells",
        "Potions",
        "Magical fantasy",
        "Magical wizards",
        "Magic in fantasy fiction",
        "Sorcery"
    ],
    "attributes_unique": [
        "Hermione Granger",
        "Ron Weasley",
        "Hogwarts",
        "Dumbledore",
```

```
            "Lord Voldemort",
            "Voldemort",
            "Gryffindor",
            "Quidditch",
            "Horcruxes",
            "Ghosts from Hogwarts"
        ],
        "hypernyms": [
            "Fantasy fiction",
            "Young adult fiction",
            "Magical realism",
            "Magic realism",
            "Book series",
            "British media franchises",
            "Young adult literature",
            "Children's books",
            "Children's literature",
            "Film series"
        ],
        "siblings": [
            "The Lord of the Rings",
            "The Chronicles of Narnia",
            "Percy Jackson",
            "His Dark Materials",
            "The Golden Compass",
            "The Mortal Instruments",
            "Twilight",
            "The Wizard of Oz",
            "The Hunger Games"
        ],
        "instances": [
            "Hogwarts School of Witchcraft and Wizardry",
            "Gryffindor",
            "Draco Malfoy",
            "Severus Snape",
            "Dumbledore",
            "Slytherin house",
            "Lord Voldemort"
        ],
        "coincidental": [
            "Hogwarts",
            "Hermione Granger",
            "Gryffindor",
            "Quidditch",
            "Wands",
            "Wizarding robes",
            "Broomsticks",
            "Magic spells",
            "Potions",
            "Owls"
        ]
}

{
```

```
"target": "Mickey Mouse",
"synonyms": [
],
"attributes_nonunique": [
    "Ears",
    "Face",
    "Eyes",
    "Mouth",
    "Nose",
    "Head",
    "Hat",
    "Gloves",
    "Shirt",
    "Pants"
],
"attributes_unique": [],
"hypernyms": [
    "Fictional mouse",
    "Cartoon character",
    "Beloved character",
    "Animated star",
    "Legendary character",
    "Disney icon",
    "Beloved mascot",
    "Adored mascot",
    "Animated personality",
    "Animation legend"
],
"siblings": [
    "Minnie Mouse",
    "Goofy",
    "Pluto",
    "Donald Duck",
    "Felix the Cat",
    "Sylvester the Cat",
    "Bugs Bunny",
    "Tom Cat",
    "Tweety Bird",
    "Yogi Bear"
],
"instances": [],
"coincidental": [
    "Minnie Mouse",
    "Pluto",
    "Goofy",
    "Toons",
    "Steamboat Willie",
    "Toontown",
    "Cartoon shorts",
    "Costume character",
    "Ears",
    "Clubhouse"
]
```

}

E.2. Example Adversarial Captions

```
    "target": "Harry Potter",
    "adv_captions": [
        "A bespectacled Ron Weasley sits at a wooden desk in a Gryffindor common room filled
        ↪  with warm firelight and vintage tapestries",
        "Ron Weasley and Harry Potter walk together down a long corridor lined with
        ↪  cobweb-covered portraits of former Gryffindor headmasters",
        "In the Great Hall of Hogwarts a long table filled with Gryffindor students
        ↪  including Ron Weasley and Harry Potter feasts on a sprawling banquet",
        "Ron Weasley proudly wears his Gryffindor Quidditch robes as he mounts a broomstick
        ↪  on a sunny day at the Hogwarts stadium",
        "A Gryffindor banner bearing the image of a lion hangs from the balcony of a crowded
        ↪  Hogwarts staircase where Ron Weasley is standing",
        "Harry Potter and Ron Weasley sneak through a hidden passageway behind a Gryffindor
        ↪  tapestry in the castle walls",
        "Gryffindor students including Ron Weasley and Harry Potter gather around Professor
        ↪  McGonagall in a cozy classroom filled with candles and ancient tomes",
        "The Hogwarts Express chugs along a scenic countryside route as Ron Weasley and
        ↪  Harry Potter sit together in a cozy compartment filled with Gryffindor trunks
        ↪  and luggage",
        "Ron Weasley and Harry Potter stand side by side wearing their Gryffindor robes on
        ↪  the castle steps as the sun rises over the Hogwarts grounds",
        "A large stone lion statue representing Gryffindor house sits atop a pedestal in a
        ↪  quiet courtyard where Ron Weasley is sitting on a bench",
        "Gryffindor students Ron Weasley and Harry Potter study for their upcoming exams in
        ↪  a quiet library filled with ancient books and mysterious artifacts",
        "Ron Weasley holds up a Gryffindor scarf as he and Harry Potter cheer on their
        ↪  Quidditch team from the crowded stadium bleachers"
    ]

    "target": "Star Wars",
    "adv_captions": [
        "A lone figure of Chewbacca stands atop a moss-covered stone wall, gazing out at a
        ↪  binary sunset on the horizon of a distant planet",
        "Chewbacca and Han Solo navigate through a crowded marketplace on Coruscant,
        ↪  surrounded by alien species and towering skyscrapers",
        "R2-D2 rolls across the dusty floor of a abandoned Jedi temple, his bright blue
        ↪  lights illuminating ancient stone carvings and mysterious artifacts",
        "In a dimly lit cantina, Chewbacca lets out a mighty roar as he arm-wrestles a
        ↪  hulking Wookiee opponent, while patrons look on in awe",
        "A squadron of X-wing fighters, escorted by Chewbacca in the Millennium Falcon,
        ↪  soars through the depths of space towards a looming Death Star",
        "R2-D2 and C-3PO stand at the edge of a serene lake on the planet Naboo, surrounded
        ↪  by lush greenery and vibrant water lilies",
        "Chewbacca pilots the Millennium Falcon through a swirling asteroid field, dodging
        ↪  and weaving between massive rocks and debris",
        "On the forest moon of Endor, R2-D2 and Chewbacca join a group of Ewoks in a lively
        ↪  celebration, complete with music, dancing, and feasting",
        "A shadowy figure of Darth Vader looms in the background as Chewbacca and Han Solo
        ↪  sneak past stormtroopers in a crowded Imperial corridor",
        "R2-D2 beeps and whistles as he extends a mechanical arm to interact with a ancient
        ↪  droid in a long-abandoned factory on the planet Tatooine",
        "In a heart-pumping dogfight, Chewbacca mans the turret of the Millennium Falcon,
        ↪  blasting enemy TIE fighters out of the sky",
        "Chewbacca and Leia Organa share a heartfelt moment, standing together on the bridge
        ↪  of a Rebel Alliance ship, looking out at a starry expanse",
```

```
    "R2-D2 and Chewbacca explore the cramped, dimly lit corridors of a captured Imperial
    ↪  ship, searching for vital information and hidden dangers",
    "A majestic procession of Rebel ships, including the Millennium Falcon with
    ↪  Chewbacca at the helm, approaches the vibrant planet of Dantooine",
    "R2-D2 sits atop a pile of scavenged parts and machinery, surrounded by Chewbacca
    ↪  and other Rebel allies, as they work to repair and rebuild their ships"
]

"target": "Seattle",
"adv_captions": [
    "A sunny day in Seattle with the Space Needle towering above the city as a ferry
    ↪  sails across Puget Sound with the Olympic Mountains in the background",
    "The Seattle Seahawks football team playing a lively game at CenturyLink Field with
    ↪  the Space Needle visible in the distance beyond the stadium",
    "A nighttime view of the Seattle cityscape with the Space Needle lit up in colorful
    ↪  lights and a giant screen in the nearby park showing a Seattle Mariners baseball
    ↪  game",
    "A group of friends walking along the Seattle waterfront wearing Seattle Sounders
    ↪  soccer jerseys and taking photos in front of the iconic Space Needle",
    "The Space Needle standing tall and proud as a backdrop for a lively parade
    ↪  celebrating a Seattle Seahawks Super Bowl win",
    "A bird's eye view of the city of Seattle with the Space Needle at its center and
    ↪  Safeco Field the home of the Seattle Mariners baseball team visible in the
    ↪  foreground",
    "A fan waving a giant Seattle Seahawks flag while standing at the top of the Space
    ↪  Needle with the city spread out below",
    "The Seattle Mariners playing a sold-out game at Safeco Field with the Space Needle
    ↪  visible from the upper deck seats",
    "A person jogging along the Seattle waterfront path with the Space Needle and a
    ↪  giant Seattle Seahawks logo on a nearby building in the background ",
    "The city of Seattle lit up at night with the Space Needle and CenturyLink Field the
    ↪  home of the Seattle Seahawks and Seattle Sounders illuminated in the darkness"
]
```

E.3. Example Images

Figure 9. **Row 1:** images of “Harry Potter” from a base Stable Diffusion v1.4. **Row 2:** images of “Harry Potter” after unlearning “Harry Potter” with REC. **Row 3:** images of “Harry Potter” after unlearning “Harry Potter” with REC+.

Figure 10. **Row 1:** images of “The Lord of the Rings” from a base Stable Diffusion v1.4. **Row 2:** images of “The Lord of the Rings” after unlearning “Harry Potter” with REC. **Row 3:** images of “The Lord of the Rings” after unlearning “Harry Potter” with REC+.

Figure 11. **Row 1:** images of “Star Trek” from a base Stable Diffusion v1.4. **Row 2:** images of “Star Trek” after unlearning “Star Wars” with REC. **Row 3:** images of “Star Trek” after unlearning “Star Wars” with REC+.

Figure 12. **Row 1:** images of “Obi Wan Kenobi” from a base Stable Diffusion v1.4. **Row 2:** images of “Obi Wan Kenobi” after unlearning “Star Wars” with REC. **Row 3:** images of “Obi Wan Kenobi” after unlearning “Star Wars” with REC+.

Figure 13. **Row 1:** images of “Chewbacca” from a base Stable Diffusion v1.4. **Row 2:** images of “Chewbacca” after unlearning “Star Wars” with REC. **Row 3:** images of “Chewbacca” after unlearning “Star Wars” with REC+.