Table R1: Evaluation of AdvUnlearn model robustness in ASR (Attack Success Rate) across various unlearning tasks (Nudity, Style, Object) under distinct attacks. A lower ASR indicates better robustness.

Attack Method	Nudity	Van Gogh	Church	Parachute	Tench	Garbage Truck
UnlearnDiffAtk	21.13%	2%	6%	14%	4%	8%
CCE	39.44%	28%	36%	48%	24%	44%
PEZ	3.52%	0%	2%	0%	0%	4%
PH2P	5.63%	0%	4%	0%	2%	6%

Table R2: Evaluation of AdvUnlearn's robustness (in ASR) for nudity unlearning under the RAB attack with various token lengths. A lower ASR indicates better robustness.

Length	16	38	77
ASR	2.11%	1.05%	1.05%

Table R3: Erasure performance (measured by the generation rate under original inappropriate prompts w/o attack) for various DM unlearning methods across different scenarios (nudity, style, object). A lower generation rate indicates better erasing performance.

Task	Nudity	Van Gogh	Church	Parachute	Tench	Garbage Truck
AdvUnlearn	7.75%	0%	0%	2%	0%	0%
ESD	20.42%	2%	14%	4%	2%	2%
SPM	54.93%	42%	44%	26%	6%	4%
FMN	88.03%	10%	52%	46%	42%	40%

Table R4: Evaluation of utility performance for Van Gogh-Unlearned ESD and AdvUnlearn models through style classifier accuracy on 500 images generated in each of the other styles (Monet, Paul Cezanne, Andy Warhol). A higher accuracy indicates better utility on under a non-forgetting style.

Other Style	ESD	AdvUnlearn
Monet	20.60%	52.40%
Paul Cezanne	7.80%	90.80%
Andy Warhol	45.60%	71.40%

Table R5: Evaluation of erasure performance (measured by the generation rate under original inappropriate prompts w/o attack as Tab. R3) and robustness (measured by ASR as Tab. R1) for AdvUnlearn models in object unlearning scenarios with various sizes of prompt sets.

Prompt Number	Metric	Church	Parachute	Tench	Garbage Truck
50	Erasure Performance	0%	2%	0%	0%
50	ASR	6%	14%	4%	8%
150	Erasure Performance	0%	1.33%	0%	0%
150	ASR	5.33%	12.67%	4.67%	6%

Table R6: Evaluation of the erasing performance (measured by the generation rate under original inappropriate prompts w/o attack as Tab. R3) and the utility performance (measured by post-generation accuracy as Tab. R4) for ESD and AdvUnlearn models using object classifier over 500 images generated per object class.

Unlearned Concept	[ESD] Unlearned Class	[ESD] Other Classes	[AdvUnlearn] Unlearned Class	[AdvUnlearn] Other Classes
Church	3.4%	76.37%	0.4%	83.91%
Garbage Truck	1.4%	66.86%	0%	87.91%
Parachute	1.8%	78.26%	0.8%	86.11%
Tench	0.6%	77.03%	0%	89.83%

Table R7: Evaluation of utility performance for AdvUnlearn models across various unlearning scenarios (nudity, style, object) using different prompt set sizes for image generation, assessed by FID and CLIP scores.

Prompt Number	Metric	Nudity	Van Gogh	Church
10k	FID	19.34	16.96	18.06
10k	CLIP Score	0.29	0.308	0.305
30k	FID	19.15	17.03	17.94
30k	CLIP Score	0.293	0.310	0.308