A. Extra Preliminary on CA Layers

The cross-attention (CA) layers in the conditional denoising UNet $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{C})$ align the latent representation of the noisy image with that of the textual prompt. The latent variable at time step t is denoted by $\mathbf{z}_t \in \mathbb{R}^{D_z \times H \times W}$ with a spatial dimension $H \times W$ and a channel dimension Dz, while the text embedding, i.e. the latent representation of the textual prompt, is denoted by $\mathbf{C} \in \mathbb{R}^{l \times D_c}$. At the i-th CA layer, the attention map is computed by

$$\mathbf{A}_i = \text{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^\top}{\sqrt{d_i}}\right), \tag{7}$$

where d_i is the latent feature dimension. The queries $\mathbf{Q}_i \in \mathbb{R}^{H_i W_i \times d_i}$ are obtained by projecting the latent features of the noisy image returned by the previous module, while both the keys $\mathbf{K}_i \in \mathbb{R}^{l \times d_i}$ and values $\mathbf{V}_i \in \mathbb{R}^{l \times d_i}$ are computed by projecting the text embedding but using different projection matrices. Finally, the output of this CA layer is computed from the attention map, and the values by $\mathbf{z}_t^{i+1} = \phi(\mathbf{A}_i \mathbf{V}_i)$, where a common choice of $\phi(\cdot)$ is a multi-layer perceptron. The subsequent modules take $\mathbf{z}_t^{i+1}$ for further processing. For the convenience of explaining, we do not distinguish the notation i between layers in the main text.

B. Equation Derivation

B.1. On Equation (5)

Working with the subspace constructed as the span of the vector set $\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^n$, we obtain a set of orthonormal basis $\left\{\boldsymbol{o}_t^{h,j}\right\}_{h=1}^n$ through the Gram-Schmidt orthogonalization. When the value vectors $\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^n$ are linearly independent, each orthonormal basis can be expressed as a linear combination of these vectors such that

$$\boldsymbol{o}_t^{h,j} = \sum_{k=1}^{n} w_{hk} \boldsymbol{v}_t^{k,j}, \tag{8}$$

where w_{hk} are the combination weights. We have explained the linear independence assumption on $\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^n$ in Section 3.3.2. Incorporating Eq. (8) into Eq. (3) but replacing only the second $\boldsymbol{o}_t^{h,j}$, it results in the following revised calculation of the orthogonal complement:

$$\boldsymbol{v}_r^j = \boldsymbol{v}^j - \sum_{h=1}^{n} \left(\sum_{k=1}^{n} w_{hk} \left(\boldsymbol{o}_t^{k,j}\right)^T \boldsymbol{v}^j \right) \boldsymbol{v}_t^{h,j}. \tag{9}$$

The importance of this revised equation lies in the fact that it computes a weighted sum of the value vectors when performing the erasing. This enables the application of the adaptive erasing shift mechanism based on the value vectors, for which we further revise the erasing operation as

$$\boldsymbol{v}_r^j = \boldsymbol{v}^j - \sum_{h=1}^{n} \delta\left(\boldsymbol{v}_t^{h,j}, \boldsymbol{v}^j\right) \left(\sum_{k=1}^{n} w_{hk} \left(\boldsymbol{o}_t^{k,j}\right)^T \boldsymbol{v}^j \right) \boldsymbol{v}_t^{h,j}. \tag{10}$$

Storing the combination weights in the matrix $\mathbf{W} = [w_{hk}] \in R^{n \times n}$, it acts as a projection matrix transforming the two vector sets by

$$\begin{bmatrix} \boldsymbol{o}_t^{1,j} & \cdots & \boldsymbol{o}_t^{n,j} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_t^{1,j} & \cdots & \boldsymbol{v}_t^{n,j} \end{bmatrix} \mathbf{W}. \tag{11}$$

B.2. Alternative Orthonormal Basis Calculation

Purely for the interest of readers, we point out an alternative way to calculate the orthonormal basis. Constructing a matrix $\hat{\mathbf{V}}_t^j \in \mathbb{R}^{d \times n}$ by using $\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^n$ as its columns, following Equation (5.13.6) of the linear algebra textbook [29], the projection of $\boldsymbol{v}^j$ onto $\text{span}^\perp\left(\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^n\right)$ can be directly computed from $\hat{\mathbf{V}}_t^j$ by

$$\begin{aligned} \boldsymbol{v}_r^j &= \mathbf{P}_{\text{span}^\perp\left(\left\{\boldsymbol{v}_t^{h,j}\right\}_{h=1}^n\right)} \boldsymbol{v}^j \\ &= \left(\mathbf{I}_d - \hat{\mathbf{V}}_t^j \left(\left(\hat{\mathbf{V}}_t^j\right)^T \hat{\mathbf{V}}_t^j \right)^{-1} \left(\hat{\mathbf{V}}_t^j\right)^T \right) \boldsymbol{v}^j. \end{aligned} \tag{12}$$

Compared to Eq. (3), Eq. (12) does not require the Gram-Schmidt orthogonalization, but the inverse calculation. Defining $\mathbf{P}_t^j = \hat{\mathbf{V}}_t^j \left(\left(\hat{\mathbf{V}}_t^j \right)^T \hat{\mathbf{V}}_t^j \right)^{-1} \left(\hat{\mathbf{V}}_t^j \right)^T$, one potential way to enable token-wise adaptive erasing shift based on Eq. (12) is

$$\boldsymbol{v}_r^j = \left(\mathbf{I}_d - \text{Diag} \left[\delta \left(\boldsymbol{v}_t^{h,j}, \boldsymbol{v}^j \right) \right] \mathbf{P}_t^j \right) \boldsymbol{v}^j, \tag{13}$$

where $\text{Diag} \left[\delta \left(\boldsymbol{v}_t^{h,j}, \boldsymbol{v}^j \right) \right]$ is a diagonal matrix with shift factors $\left[\delta \left(\boldsymbol{v}_t^{h,j}, \boldsymbol{v}^j \right) \right]$ as its diagonal elements. We leave the in-depth investigation of exploiting this operation in practice to our future work.

C. Additional Experimental Details

C.1. On Implementation

To implement SD v1.4, the DPM-solver is chosen as the sampler, with a total of 30 sampling timesteps and a classifier-free guidance scale set of 7.5. Notably, we set the unconditional prompt to null text, as the negative prompt serves as a training-free method that can be directly compared with our AdaVD. To ensure a fair comparison, particularly for prior preservation, we use the same random seed (seed 0) across all methods to generate images under identical conditions. For the specific instance, art style, and celebrity erasure, we simply fix the hyperparameters to $p = 100$, $\epsilon = 0.93$, and $s = 2$. Our AdaVD performs consistently well with this unified hyper-parameter configuration.

C.2. Additional Hyper-parameter Analysis

The hyper-parameters of the shift factor, including $0 < \epsilon < 1$ and $p, s > 0$, are closely related to the cosine similarities between tokens of the target concepts and tokens of the prompt. When erasing instances, art styles, and celebrity concepts, we notice that certain non-target concepts contained by the prompt semantically correlate with the target concept, with fairly strong correlations. For example, the non-target concept *"Mickey"* exhibits a relatively large cosine similarity of 0.65 with the target concept *"Snoopy"*, as they both belong to the category of cartoon characters. This makes it a fine balance between an unaffected generation of these non-target concepts and a successful erasure of the target concept. To examine how the erasure strength impacts such a balance, we show in Fig. 8 different image examples generated by AdaVD under various hyperparameter settings, for the target concept *"Snoopy"* and non-target concept *"Mickey"*.

Overall, the factor scale s and the threshold ϵ significantly impact the balance between the erasure efficacy and prior preservation. Specifically, it can be observed, from the top part of Fig. 8 (on target concept), that a reduction in ϵ results in a greater deviation in the generated images as compared to the original, for content relevant to the target concept. This indicates an enhanced erasure efficacy. This effect is further amplified as s increases. When adopting the setting of $s = 2$ and $\epsilon = 0.6$, the erasure becomes excessive. Conversely, for non-target concept generation, a lower threshold ϵ can negatively impact the non-target prior, as observed from the bottom part of Fig. 8 (on the non-target concept). Such a negative impact on non-target concept generation intensifies with increasing s, since a larger s amplifies the token shift. This results in a larger divergence from the original token direction, and eventually more noticeable changes in the generated images.

The erasure performance is less sensitive to p, but it still has some mild impact. For instance, when using a higher value of s, a lower p can mitigate changes in the generated visual content that is relevant to the non-target concepts. This is demonstrated in the bottom-right part of Fig. 8. When ϵ decreases to 0.7, setting p to 40 results in less deviation from the original images as compared to other values. On the other hand, when $s = 1$, a higher p positively affects the preservation of some non-target concepts that are related. This is shown in the bottom-left part of Fig. 8. When $\epsilon = 0.6$, the deviation from the original image decreases as p increases from 40 to 100.

D. Additional Single-concept Experiments

D.1. Extended Quantitative Results on Instance and Art Style Erasure

We present the extended quantitative results on instance erasure and art style erasure in Table 4 and 5, respectively. In addition to the CS for the target concept and FID for non-target concepts, we also include the FID for the target concept and CS for non-target concepts. Specifically, FID measures the distribution distance of generated images aligned with the target concept before and after concept erasure, while CS evaluates the semantic consistency between the text prompt of the non-target concept and the generated image after erasure.

However, a lower FID for the target concept only indicates significant visual changes in the generated images but does not confirm that the semantics aligned with the target concept have been fully eliminated. Similarly, a higher CS for the

Figure 8. **Impact of hyperparameter settings on erasure efficacy and prior preservation.** To evaluate how hyperparameters affect this balance, we visualize images generated by AdaVD under various hyperparameter settings for the target concept *"Snoopy"* and the related but non-target concept *"Mickey"*.

non-target concept suggests that the generated image after concept erasure still aligns closely with the text prompt, but does not guarantee small pixel-level changes. In summary, FID for the target concept and CS for the non-target concept cannot directly measure the effectiveness of erasure or prior preservation. Nevertheless, they remain valuable for further verifying and comparing the erasure efficacy and prior preservation.

D.2. On Celebrity Erasure

We experiment with erasing different celebrity concepts, including *"Bruce Lee"*, *"Marilyn Monroe"*, and *"Melania Trump"*. Five types of prompts were tested, each containing a distinct concept from *"Bruce Lee"*, *"Marilyn Monroe"*, *"Melania Trump"*, *"Anne Hathaway"* and *"Tom Cruise"*. As reported in Table 6, AdaVD consistently exhibits superior erasing efficacy with prior preservation. When erasing different celebrities, AdaVD achieves the lowest or near-lowest CS and FID values, particularly excelling in FID. Although SPM ranks the second in prior preservation based on its FID scores, it falls significantly behind in its overall prior preservation quality, as compared to AdaVD.

Fig. 9 illustrates and compares generated images of methods, where consistent superior performance of AdaVD can be observed. For the target concept *"Marilyn Monroe"*, AdaVD, SPM, and MACE can all successfully remove the celebrity identity. But SPM is overly aggressive at erasing, obscuring the facial outlines. For non-target concepts, all the four competing methods have caused some quite strong deviations, altering the original images. This is particularly noticeable in the generated images from the prompt corresponding to *"Melania Trump"*. For instance, MACE and SPM have introduced an

	Snoopy		Mickey		Spongebob		Pikachu		Dog		Legislator	
	CS	FID	CS	FID	CS	FID	CS	FID	CS	FID	CS	FID
SD v1.4	28.49	-	26.50	-	27.30	-	27.41	-	24.27	-	23.73	-
*Erase **Snoopy***												
	CS ↓	FID	CS	FID ↓	CS	FID ↓	CS	FID ↓	CS	FID ↓	CS	FID ↓
ConAbl	25.38	103.80	26.68	38.44	27.02	41.59	27.57	29.68	24.12	27.76	23.48	27.36
MACE	<u>20.78</u>	169.22	22.95	118.01	23.33	111.90	25.77	81.99	23.96	43.27	22.25	65.97
SPM	23.89	122.63	26.66	<u>33.06</u>	27.12	<u>34.70</u>	27.51	<u>23.89</u>	24.24	<u>19.61</u>	23.70	<u>18.26</u>
NP	23.66	125.98	26.14	59.58	26.66	78.74	27.36	52.37	23.89	67.51	22.16	55.22
SLD	27.84	64.78	26.46	48.12	27.52	55.36	27.33	38.74	24.03	41.95	22.80	49.08
Ours	**20.28**	120.46	26.53	**5.72**	27.25	**8.56**	27.40	**5.79**	24.27	**2.32**	23.77	**6.07**
*Erase **Snoopy** and **Mickey***												
	CS ↓	FID	CS ↓	FID	CS ↓	FID ↓	CS	FID ↓	CS	FID ↓	CS	FID ↓
ConAbl	24.26	119.96	24.08	96.94	27.02	46.32	27.75	39.63	23.98	30.57	23.33	27.49
MACE	<u>20.74</u>	171.16	<u>20.71</u>	140.50	25.87	51.49	25.87	110.67	23.82	52.07	21.70	77.13
SPM	23.16	128.08	22.81	115.02	26.92	<u>41.58</u>	27.45	<u>31.77</u>	24.13	<u>21.96</u>	23.60	<u>23.69</u>
NP	23.59	124.10	24.85	83.68	26.69	81.41	27.27	50.10	23.62	65.93	21.84	58.88
SLD	27.76	59.97	26.74	50.16	27.53	54.59	27.29	39.24	23.97	41.62	22.66	50.13
Ours	**20.29**	121.12	**19.93**	108.22	27.27	**9.34**	27.42	**5.84**	24.26	**2.41**	23.73	**6.43**
*Erase **Snoopy** and **Mickey** and **Spongebob***												
	CS ↓	FID	CS ↓	FID	CS ↓	FID	CS	FID ↓	CS	FID ↓	CS	FID ↓
ConAbl	23.94	126.70	23.64	105.07	25.04	108.67	27.76	51.20	23.83	23.83	23.17	30.03
MACE	<u>20.48</u>	172.80	<u>20.50</u>	143.66	21.59	120.87	24.38	99.68	23.70	47.46	21.74	70.38
SPM	22.81	133.06	22.35	121.85	<u>20.82</u>	152.72	27.45	39.83	24.10	<u>22.68</u>	23.52	<u>25.31</u>
NP	24.29	129.75	24.76	89.74	25.31	106.30	27.28	64.75	23.55	65.10	21.63	59.33
SLD	27.84	58.16	26.71	49.70	27.60	54.61	27.35	<u>39.41</u>	23.90	42.32	22.46	49.88
Ours	**19.39**	124.49	**19.73**	112.97	**20.34**	118.47	27.42	**6.85**	24.27	**2.79**	23.76	**7.26**

Table 4. **Extended quantitative comparison of single- and multi-instance erasure.** The best and second-best results are marked in **bold** and <u>underlined</u>, respectively. Columns in gray indicate items that do not directly reflect erasure efficacy or prior preservation performance.

Figure 9. **Qualitative comparison of celebrity erasure.** Our AdaVD can effectively remove the target concept *"Marilyn Monroe"* while preserving non-target celebrities like *"Bruce Lee"* and *"Melania Trump"*.

additional arm in the left image, NP has altered the original pose, and SLD has caused a severe visual change in the mouth and eye areas. In contrast, AdaVD is able to successfully maintain all the non-target images with minimal visual changes.

D.3. On NSFW Erasure

Unlike the erasure of specific instances, art styles, and celebrities, NSFW concept erasure is more challenging. One reason is that the NSFW concepts are often implicit and hidden within prompts that can be particularly rich in their semantics. Also, many NSFW concepts have synonyms, and it is important to remove both the target concept and its synonyms. For instance, when targeting at removing the *"nudity "* concept, it is essential to also remove the *"sexual"* concept. We experiment with erasing the *"nudity"* concept using the I2P benchmark. To examine how well the *"nudity"* concept is erased, we employ the NudeNet with a threshold of 0.3 to detect nudity in the generated images and analyze the total number of nude items and the overall nude images that are detected.

	Van Gogh		Picasso		Monet		Andy Warhol		Caravaggio	
	CS	FID	CS	FID	CS	FID	CS	FID	CS	FID
SD v1.4	29.20	-	28.84	-	29.41	-	29.73	-	27.09	-
*Erase **Van Gogh***										
	CS ↓	FID	CS	FID ↓	CS	FID ↓	CS	FID ↓	CS	FID ↓
ConAbl	28.80	120.93	28.10	71.71	25.99	138.72	29.34	70.30	26.83	73.10
MACE	27.74	144.75	28.37	65.77	29.48	69.79	29.30	83.37	27.11	75.41
SPM	**24.78**	185.50	28.34	62.25	29.34	32.27	29.52	58.30	27.01	61.50
NP	24.90	193.24	25.11	141.56	26.08	124.52	27.06	127.85	25.34	136.32
SLD	27.48	133.07	26.89	103.96	27.61	109.11	28.24	103.89	25.82	119.32
SAFREE	25.82	183.06	25.84	130.35	27.15	128.71	27.20	127.72	25.53	134.46
Ours	24.87	188.94	28.80	**6.82**	29.43	**2.66**	29.74	**8.36**	27.09	**6.84**
*Erase **Picasso***										
	CS	FID ↓	CS ↓	FID	CS	FID ↓	CS	FID ↓	CS	FID ↓
ConAbl	29.46	58.62	27.72	121.45	26.37	140.34	29.51	73.35	27.17	67.44
MACE	29.73	60.46	27.11	131.82	29.44	49.92	29.65	76.10	27.08	72.85
SPM	29.26	38.79	26.69	157.32	29.44	7.76	29.67	52.00	27.08	51.40
NP	29.28	111.35	**26.14**	169.23	29.34	91.11	28.14	116.24	26.50	121.82
SLD	29.36	98.21	27.03	105.37	29.79	93.01	28.80	97.00	26.42	110.05
SAFREE	29.96	117.32	26.42	183.80	29.45	93.51	27.88	122.89	26.32	116.51
Ours	29.17	**5.49**	26.99	132.64	29.43	**2.33**	29.72	**9.38**	27.09	**7.05**
*Erase **Monet***										
	CS	FID ↓	CS	FID ↓	CS ↓	FID	CS	FID ↓	CS	FID ↓
ConAbl	25.84	141.52	25.47	132.10	24.53	143.48	26.25	208.38	25.48	186.26
MACE	29.47	76.90	28.56	69.35	26.89	109.58	29.34	88.35	26.75	81.72
SPM	29.19	41.03	28.65	29.71	27.00	105.09	29.65	31.90	29.65	25.99
NP	26.31	137.21	25.59	126.75	**24.47**	140.92	27.05	127.22	24.85	135.83
SLD	28.22	94.48	27.10	92.88	25.73	120.14	28.34	100.90	25.45	114.87
SAFREE	26.07	125.98	26.25	119.19	25.33	153.96	26.82	125.27	25.45	129.07
Ours	29.19	**6.94**	28.80	**6.50**	26.30	114.06	29.76	**8.46**	27.10	**7.19**

Table 5. **Extended quantitative comparison of art style erasure.** AdaVD achieves a superior balance between erasure efficacy and prior preservation, especially excelling in prior preservation. Notably, it outperforms the concurrent method SAFREE, which also employs orthogonal decomposition.

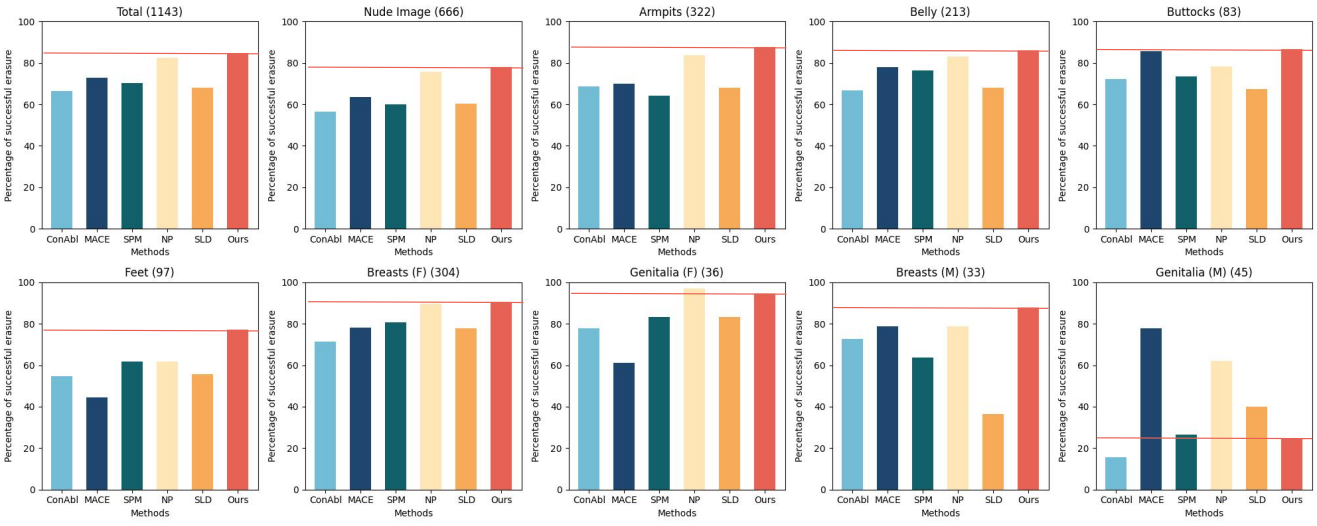


Figure 10. **Performance of AdaVD on NSFW erasure.** The number following each category represents the number of nude items generated by SD v1.4, while each bar illustrates the success rate of erasing the corresponding nude items for each method. Our AdaVD demonstrates superior performance on NSFW erasure, outperforming both training-based and training-free methods.

	Bruce Lee		Marilyn Monroe		Melania Trump		Anne Hathaway		Tom Cruise	
	CS	FID	CS	FID	CS	FID	CS	FID	CS	FID
SD v1.4	30.77	-	27.70	-	29.80	-	31.96	-	31.12	-
*Erase **Bruce Lee***										
	CS ↓	FID	CS	FID ↓	CS	FID ↓	CS	FID ↓	CS	FID ↓
ConAbl	31.35	87.57	28.23	57.79	29.77	40.95	29.77	40.95	30.97	53.53
MACE	25.04	131.29	28.13	74.80	30.07	68.83	31.91	75.05	28.13	71.20
SPM	27.75	123.67	27.71	26.89	29.81	7.83	31.96	9.46	31.13	28.54
NP	24.70	150.85	26.84	102.67	28.94	82.13	30.34	89.60	29.67	89.92
SLD	28.22	102.26	26.29	87.15	29.43	84.32	30.97	85.37	29.32	94.07
Ours	**20.67**	138.70	27.70	**6.68**	29.82	**5.08**	31.97	**6.39**	31.10	**13.11**
*Erase **Marilyn Monroe***										
	CS	FID ↓	CS ↓	FID	CS	FID ↓	CS	FID ↓	CS	FID ↓
ConAbl	30.88	66.97	28.75	88.45	29.69	51.52	32.05	58.57	31.10	54.13
MACE	31.30	76.23	**19.52**	148.34	31.93	71.05	30.16	74.90	31.52	73.06
SPM	30.76	32.70	21.87	145.81	29.83	25.27	31.96	22.86	31.10	19.34
NP	29.50	113.12	25.86	149.95	29.29	87.27	29.42	98.86	30.02	86.70
SLD	29.59	87.83	26.70	98.51	28.81	107.42	29.25	102.13	30.35	81.12
Ours	30.73	**7.88**	19.87	116.94	29.80	**4.46**	31.93	**5.43**	31.13	**9.33**
*Erase **Melania Trump***										
	CS	FID ↓	CS	FID ↓	CS ↓	FID	CS	FID ↓	CS	FID ↓
ConAbl	30.62	54.46	28.14	59.10	29.89	79.04	31.94	58.65	31.00	54.50
MACE	31.30	78.07	27.84	71.34	20.71	122.42	31.94	73.49	31.41	71.09
SPM	30.79	14.08	27.63	30.40	**23.12**	129.68	31.86	28.85	31.10	22.35
NP	29.38	115.35	27.63	103.83	23.73	131.73	28.72	106.04	30.27	106.00
SLD	29.55	90.69	26.24	93.93	25.45	103.52	28.43	104.48	30.47	88.31
Ours	30.75	**7.32**	27.69	**6.86**	23.28	96.66	31.95	**6.52**	31.08	**5.74**

Table 6. **Quantitative comparison of celebrity erasure.** Compared to both training-based and training-free methods, AdaVD achieves an optimal balance between erasure efficacy and prior preservation, demonstrating exceptional performance, particularly in prior preservation.

Results are reported in Fig. 10, where, despite the challenges, AdaVD demonstrates a superior nudity erasure performance, with a semi-threshold and a slower increasing rate. It outperforms both training-based and training-free methods, achieving the best or close-to-best success rate in nearly all categories, with approximately 85% of the nude items successfully removed. It is worth mentioning that NudeNet can be overly aggressive at detecting nude items, resulting in detection errors. For example, it may incorrectly classify a circle with a dot as *"Female Breast Exposure"* or a person opening their mouth as *"Male Genitalia Exposure"*. We increased the NudeNet threshold to 0.3, in order to mitigate this issue, but still, there is a detection error. Being examined by an overly strict nudity detector that can flag sometimes healthy or irrelevant content as nude ones, AdaVD achieves the highest erasure rate for nearly all tested nude items compared to other competing methods, as shown in Fig. 10.

D.4. More Erasure Examples

We demonstrate additional examples for erasing single concepts from prompts that contain such concepts. The experimented concepts include the specific instances of *"Statue of Liberty"*, *"BB8"*, *"C3PO"*, and *"Grumpy Cat"*, the celebrity *"Benicio Del Toro*", and the art style *"Cyberpunk"*. Among these, *"BB8"* and *"C3PO"* are fictional characters, while *"Statue of Liberty"* and *"Grumpy Cat"* represent realistic entities from daily life. Fig. 11 presents the generated image examples. It can be seen that our AdaVD consistently exhibits superior erasure efficacy across all these concepts, being robust in erasing diverse types of concepts.

E. On Transferability to Other T2I Models

The proposed AdaVD is a flexible concept erasure approach that can be transferred to other T2I diffusion models. In addition to SD v1.4, as experimented in the main paper, we conduct additional experiments to demonstrate its transferability and effectiveness by integrating it with a series of other T2I diffusion models.

E.1. AdaVD on SDXL v1.0

We integrate AdaVD with SDXL v1.0 [31] which has a different architecture from SD v1.4-v2.1. It employs two distinct text encoders to process textual prompts, and their outputs are concatenated and fed into the CA layers to interact with the latent

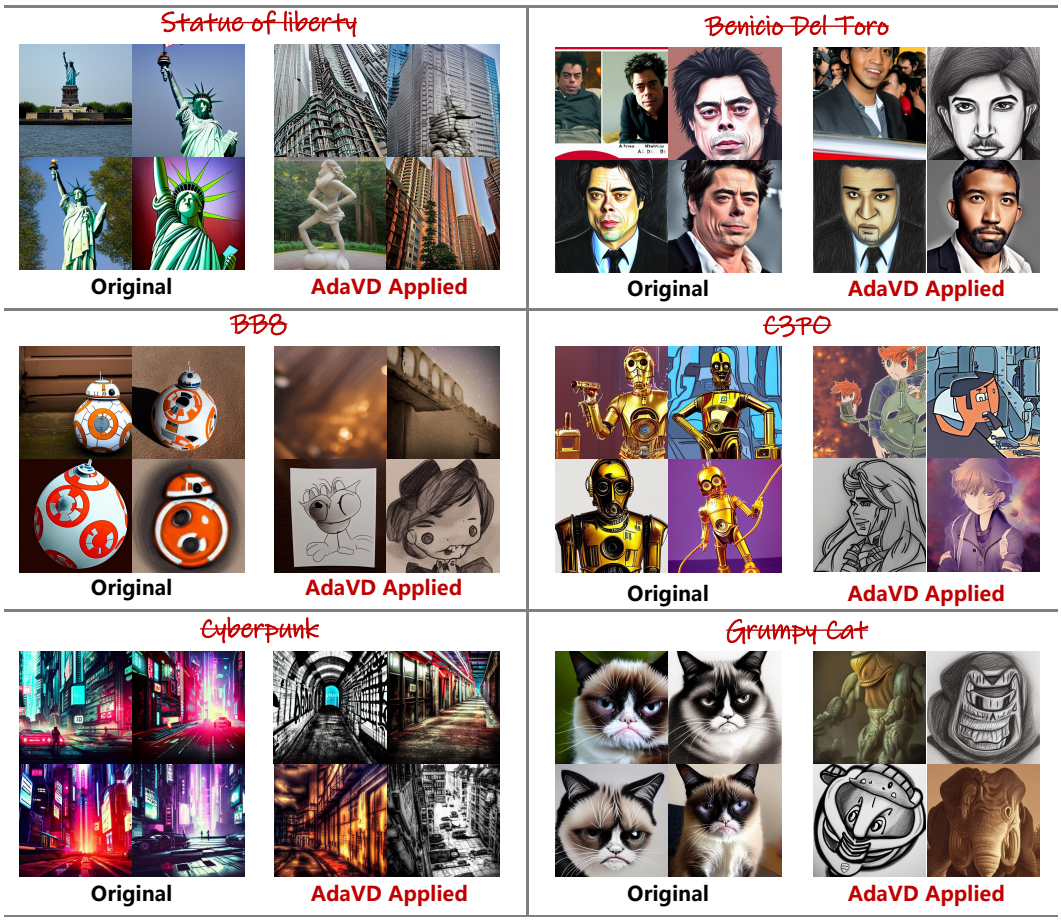


Figure 11. **Extended results of AdaVD in single-concept erasure task.** We present additional generated images after applying AdaVD with SD v1.4 to erase a single concept, further validating the erasure efficacy of our AdaVD.

representations of the noisy images. Also, the generated text embeddings are enhanced by time embeddings to ensure the alignment between textual prompts and timesteps. Following the same approach as how it is coupled with SD v1.4, AdaVD is applied in the value space at each CA layer within the UNet of SDXL v1.0. For the target concepts, both sets of their embeddings computed by the two text encoders are pre-processed following the procedure outlined in Sec. 3.2, then they are used to start the erasure process following the method outlined in Sec. 3.3.1.

Fig. 12 demonstrates the generated image examples by coupling AdaVD with SDXL v1, for long and semantically rich prompts that (do not) contain the *"Snoopy"* concept while with the target concept *"Snoopy"* to erase. Although the prompts are more complex, they do not appear challenging for AdaVD to handle. AdaVD can still accurately identify and extract the relevant semantic components associated with the target concept, and can precisely erase these without affecting the background generation. We visualize the erased component for each generated image in the smaller images within each example block of Fig. 12, following the same approach as explained in the 2nd paragraph of Section 4.4. These serve as supporting evidence, showing what semantic content has been removed by AdaVD. For those prompts containing only the non-target concepts, AdaVD successfully retains nearly all the details of the non-target content, producing images that are virtually identical to those generated by the original SDXL v1.0.

E.2. AdaVD on SDv3

A growing trend in text-to-image generative diffusion models is replacing U-Net with DiT as the noise predictor. Different from U-Net, DiT uses a transformer-based architecture, enhancing scalability in image generation. To validate the performance of our AdaVD in DiT-based diffusion models, we conduct experiments on SDv3. Different from SDv1.4 and SDXL, SDv3 uses the T5 text encoder [34], alongside other encoders, to generate text embeddings for image generation. During the target embedding pre-processing phase, we handle text embeddings differently depending on the encoder: for embeddings

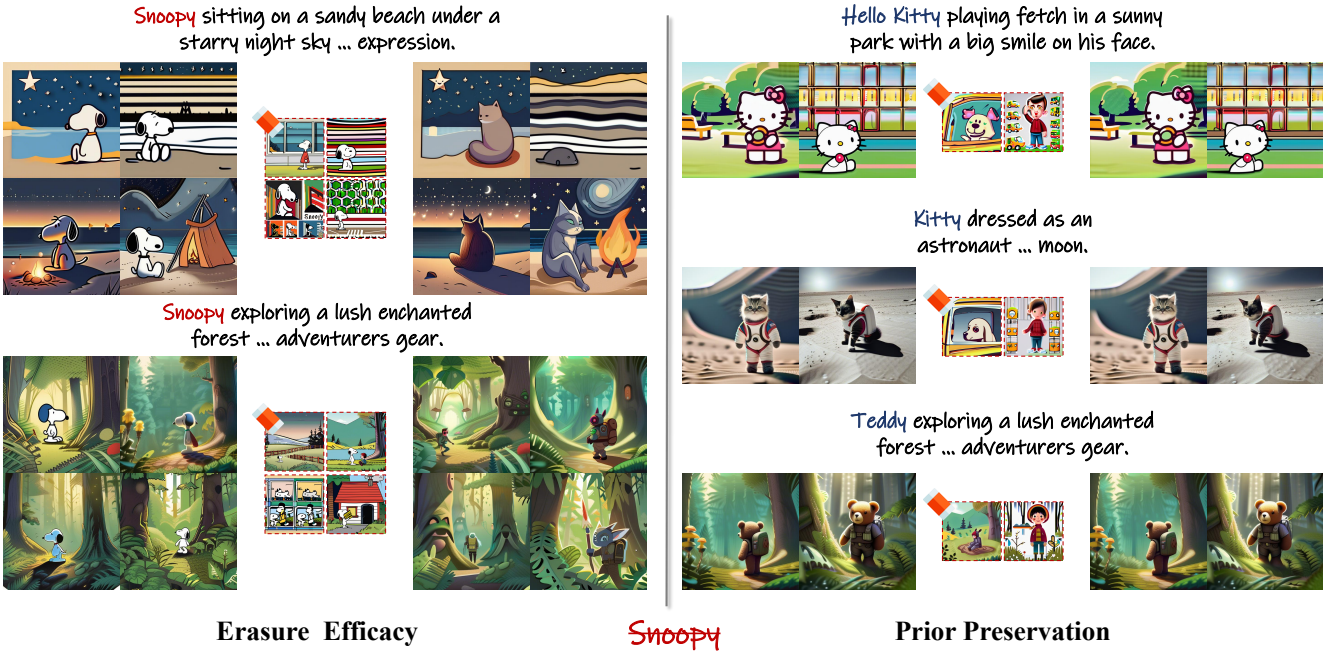


Figure 12. **Results of AdaVD on SDXL v1.0 for erasing *"Snoopy"***: Our AdaVD effectively supports SDXL v1.0, which has a different structural design than SD v1.4, in achieving effective erasure of the target concept. Additionally, AdaVD demonstrates excellent prior preservation, as evidenced by its ability to generate non-target concepts like *"Hello Kitty"*, *"Kitty"*, and *"Teddy"* even with semantically rich prompts. AdaVD successfully retains nearly all details in non-target content, underscoring its capability for precise erasure without compromising unrelated elements.

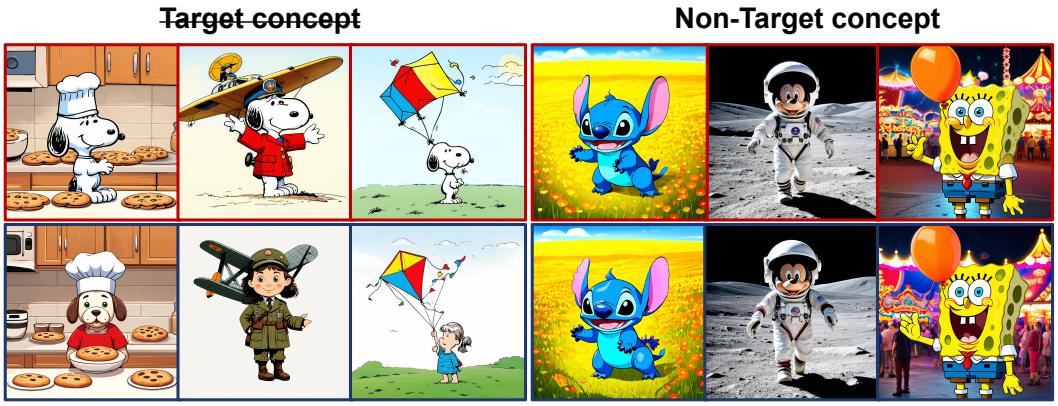


Figure 13. **Results of AdaVD on SDv3 for erasing *"Snoopy"***: The images with red and blue borders represent the before and after concept erasure, respectively. Our AdaVD effectively enables SDv3 to erase the target concept *"Snoopy"* while preserving other semantic elements in the generated images. Moreover, AdaVD demonstrates outstanding prior preservation by ensuring that non-target concepts such as *"Stitch"*, *"Mickey"*, and *"Spongebob"* remain highly similar to the generated images before concept erasure.

from the CLIP text encoder, we replicate the last subject token, while for those from T5, we spread the mean embedding of all real word tokens. As shown in Fig. 13, SDv3 successfully removes the target concept *"Snoopy"* during the generation process while preserving the integrity of non-target concepts such as *"Stitch"*, *"Mickey"*, and *"Spongebob"*. This highlights the strong prior preservation capability of AdaVD.

E.3. AdaVD on Community SD Versions

We also couple AdaVD with several community versions of SD, including RealisticVision [7], Dreamshaper [6], and Chill-outmix [5], which are all fine-tuned based on SD v1.5. These versions target high-quality image generation with specific generation objectives. For example, RealisticVision specializes in generating lifelike images, while Dreamshaper excels in producing highly imaginative visuals. We experiment with removing the target concept *"Tom Cruise"* from the text prompt

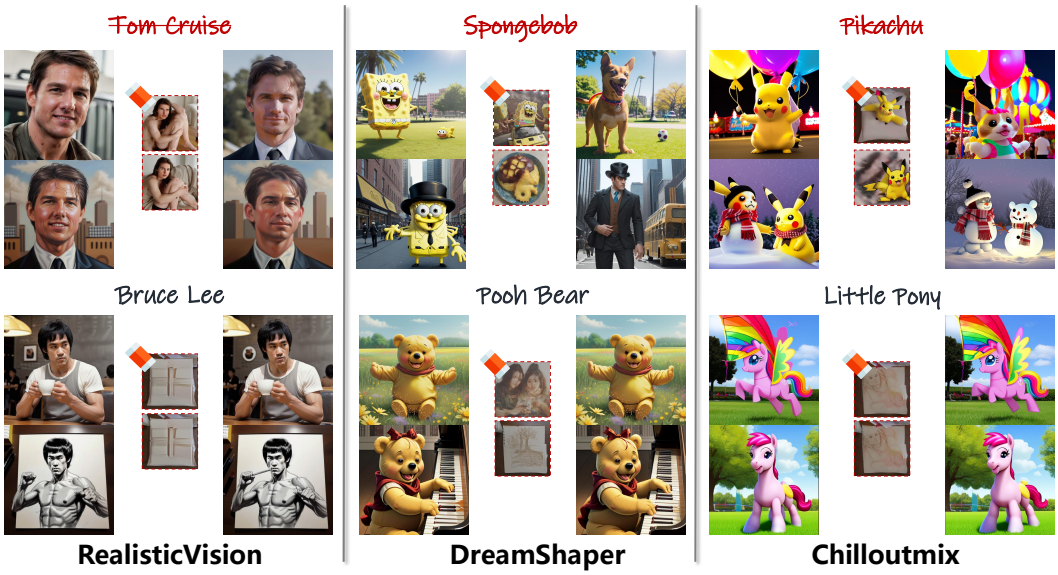


Figure 14. **Results of AdaVD on community SD versions.** Our AdaVD illustrates a high performance of both erasure efficacy and the prior preservation across SD with different versions and easing different concepts.

corresponds to *"Tom Cruise"* and *"Bruce Lee"* for RealisticVision, removing *"Spongebob"* from the text prompt corresponds to *"Spongebob"* and *"Pooh Bear"* for Dreamshaper, and removing *"Pikachu"* from the text prompt corresponds to *"Pikachu"* and *"Little Pony"* for Chilloutmix.

Fig. 14 presents the generated image examples. The results show that AdaVD is capable of effectively erasing the target concept while preserving the integrity of the non-target content. For all the experimented community versions, AdaVD can precisely locate the semantic space aligned with the target concept and isolate it with minimal disruption to the non-target semantics. Fig. 14 also visualizes the erased components as the smaller images within each example block, as in Fig. 12. Overall, the visualized erased components for prompts containing the target concepts show a high similarity to the target semantics. In contrast, for prompts corresponding to non-target concepts, the erased components lack meaningful semantic information. These serve as additional evidence, showing the effectiveness of AdaVD.

F. Comparison with Additional Baselines

F.1. Comparison with SAFREE

Orthogonal complement is widely used to decouple and separate out unwanted information. The art of using the orthogonal complement for concept erasure is on designing/deciding what space/direction to apply orthogonal complement, how to adjust removal strength, how to embed orthogonal complement in an algorithm to optimize its effect, etc. There is a concurrent work, SAFREE [51], which also used orthogonal complement to facilitate concept erasure, but in completely different ways. Our AdaVD performs orthogonal complement in value spaces of attention layers within a diffusion model. Due to its effectiveness, there is no need for any complementary design, but a soft control of removal strength through a shift factor. Different from our AdaVD, SAFREE performs orthogonal complement over diffusion model input, i.e., text embedding space. This approach necessitates complementary design elements, such as masking, another projection, and modifying detoxified embedding by Fourier transform. To control removal strength, it also uses a hard selection of whether to adopt the final detoxified embeddings. We also conduct experiments to compare the performance of AdaVD and SAFREE in erasing art style concepts. As shown in Table 5, AdaVD achieves excellent prior preservation performance and second-best erasure efficacy, outperforming SAFREE, especially in prior preservation.

F.2. Comparison with SuppressEOT

In this additional experiment, we compare with a special concept erasure method SuppressEOT [24], which requires the users to specify the positions of the erased concepts within the prompt. Because of this user-involved setting, SuppressEOT is only applicable to specific prompts and is unable to achieve system-wide concept erasure. Therefore, we only conduct a qualitative comparison of the erasure efficacy. Results are reported in Fig. 15, where a comparison of art style erasure is shown on the

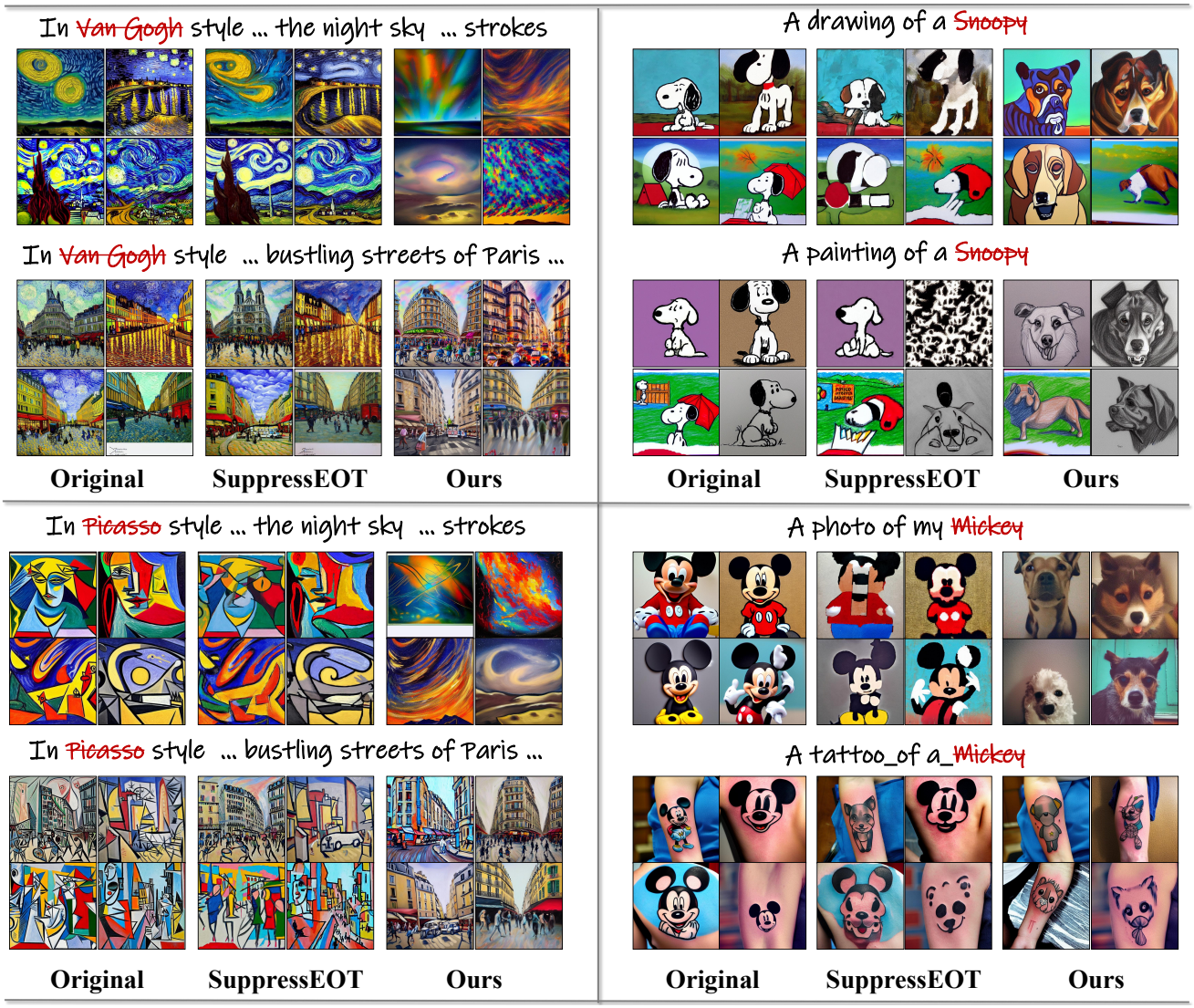


Figure 15. **Qualitative comparison between SuppressEOT and AdaVD.** We compare our AdaVD with SuppressEOT in single instance concept and art style erasure, demonstrating that AdaVD achieves more precise and effective erasure.

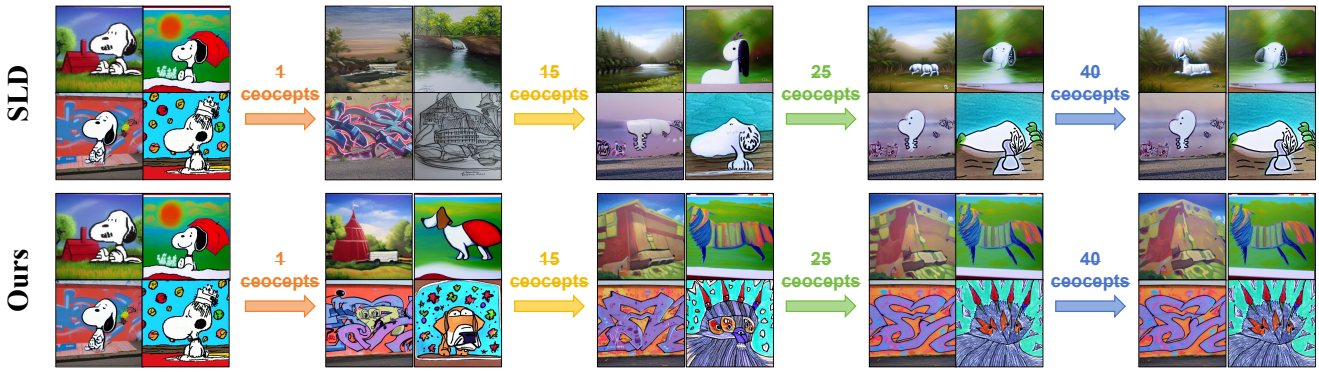
left side, while the instance erasure results are displayed on the right.

It can be seen from Fig. 15 that AdaVD is precise and effective in erasing various concepts, achieving significantly higher erasure performance across a diverse range of use cases. Unfortunately, SuppressEOT consistently fails to remove completely the target concept from the generated images. For instance, when erasing *"Mickey"* and *"Snoopy"*, SuppressEOT is not even able to erase the general outlines of these specific instances. The unsatisfactory erasure performance of SuppressEOT likely stems from the fact that it was originally designed for image editing rather than concept erasure. In image editing scenarios, preserving all the details of a prompt except for the target concept is important. This is different from the requirement of concept erasure, where the prior preservation is needed only for the generation of non-target content. Such a difference in design requirement can inherently compromise the erasure efficacy of SuppressEOT.

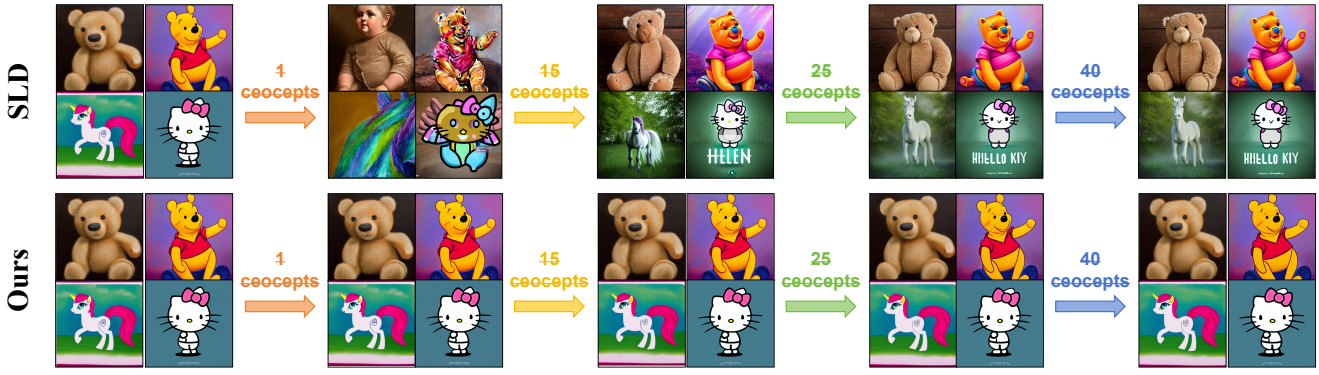
G. Additional Experiments and Analysis on Multi-Concept Erasure

G.1. On Erasing More Multi-concepts

We conduct additional experiments, investigating how our approach performs as the number of erased concepts increases, under a progressive setting. We evaluate our AdaVD by first erasing one concept *"Snoopy"* and gradually increasing the number of erased concepts to 15, 25, and 40. The details of the concepts to be erased for each case are listed in Table 7. We



(a) Erasure Efficacy



(b) Prior Preservation

Figure 16. **Examples of generated images for multi-concept erasure.** The illustrated examples show a consistently high performance of AdaVD in both (a) erasure efficacy and (b) prior preservation as the number of erased concepts increases, as compared to SLD.

Number	Target Concepts
1	*Snoopy*
15	*Snoopy, Mickey, Crystal, Pikachu, Legislator, Bruce Lee, Marilyn Monroe, Tom Cruise, Anne Hathaway, Melania Trump, Van Gogh, Picasso, Rembrandt, Andy Warhol, Caravaggio*
25	*Snoopy, Mickey, Crystal, Pikachu, Legislator, Bruce Lee, Marilyn Monroe, Tom Cruise, Anne Hathaway, Melania Trump, Van Gogh, Picasso, Rembrandt, Andy Warhol, Caravaggio, Samoyed, Doraemon, Tom, Adam Driver, Adriana Lima, Amber Heard, Amy Adams, Andrew Garfield, Angelina Jolie, Anjelica Huston*
40	*Snoopy, Mickey, Crystal, Pikachu, Legislator, Bruce Lee, Marilyn Monroe, Tom Cruise, Anne Hathaway, Melania Trump, Van Gogh, Picasso, Rembrandt, Andy Warhol, Caravaggio, Samoyed, Doraemon, Tom, Adam Driver, Adriana Lima, Amber Heard, Amy Adams, Andrew Garfield, Angelina Jolie, Anjelica Huston, Bradley Cooper, Bruce Willis, Bryan Cranston, Cameron Diaz, Channing Tatum, Charlie Sheen, Charlize Theron, Chris Evans, Chris Hemsworth, Chris Pine, Barack Obama, Beth Behrs, Bill Clinton, Bob Dylan, Bob Marley*

Table 7. **Number of concepts to be erased and their corresponding lists.** The number of concepts ranges from 1 to 40, demonstrating the efficacy of AdaVD in handling multi-concept erasure.

work with the base T2I model SD v1.4 and compare it with the existing approach SLD. The results are presented in Fig. 16, which extends Fig. 1.

It can be observed from the top erasure efficacy block of Fig. 16 that SLD gradually loses its precision when removing the target concepts. This is possible because SLD concatenates the target concepts into a prompt for guiding the generation process. When erasing too many concepts, the text encoder struggles to focus on each individual concept, resulting in

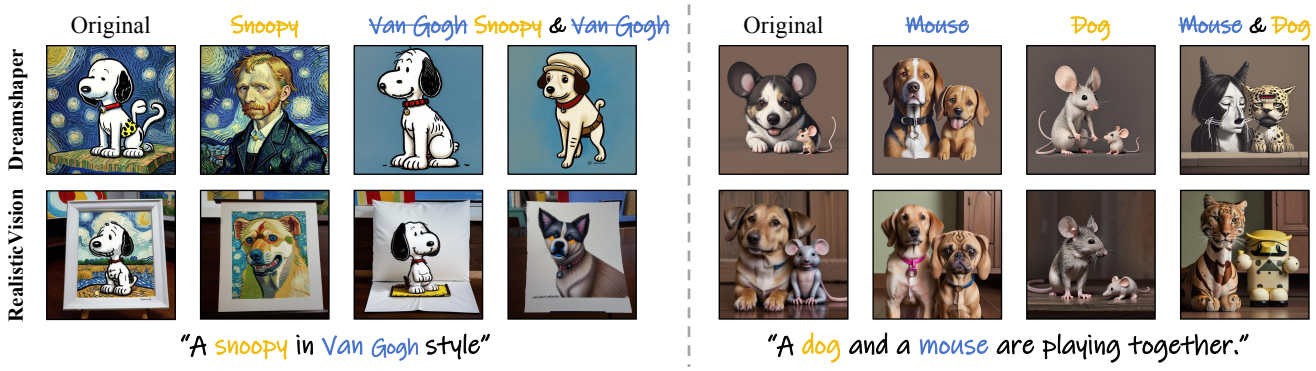


Figure 17. **Results of AdaVD on multi-concept erasure across different SD versions.** We assess the performance of AdaVD on multi-concept erasure across various community versions of SD under diverse erasure scenarios, including cross-application erasure as outlined in SPM [27] and multi-instance erasure. These evaluations further highlight the robustness and effectiveness of AdaVD in addressing the challenges of the multi-concept erasure task.

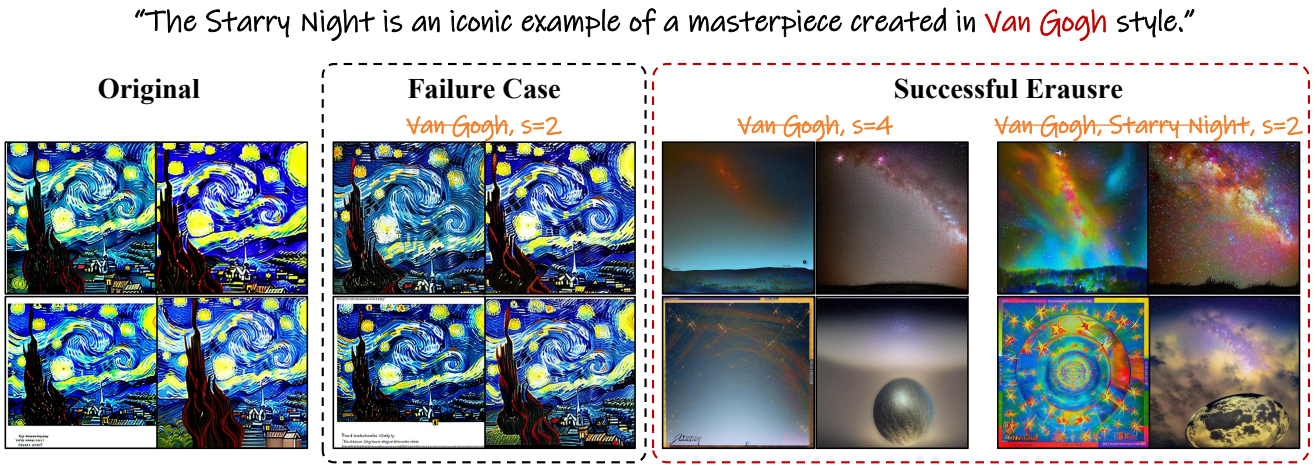


Figure 18. **Failure case** when erasing *"Van Gogh"* and its corresponding solutions.

diminished erasure efficacy. Additionally, some concepts may be truncated due to the token length limitation of the text encoder's tokenizer. Differently, AdaVD achieves consistently high performance in multi-concept erasure. It constructs a value subspace based on the orthogonal complement of all the target concepts, which ensures that no information regarding any individual concept is lost.

The bottom prior preservation block of Fig. 16 shows that AdaVD is able to generate images nearly identical to the original ones, demonstrating a superior performance in prior preservation. But SLD struggles to preserve prior knowledge, for not only the more challenging case of removing a high number of concepts but also the simple case of removing one single concept. It is worth noting that some slight change can be accumulated and amplified as the number of erased concepts increases, as shown in the hands and mouth of the generated image of *"Pooh Bear"* by our AdaVD. Also, small pixel-level changes may grow into catastrophic forgetting with an increasing number of erased concepts due to error accumulation. Therefore it is important to use FID to evaluate the performance of prior preservation, as images that closely match the originals at pixel level should result in a low FID score.

G.2. On Transferability to Other T2I Models

In this additional experiment, we integrate AdaVD with two other T2I diffusion models, including DreamShaper [6] and RealisticVision [7], assessing its multi-concept erasure performance. Two multi-concept erasure scenarios are experimented with: one is cross-application erasure as described in SPM [27], and the other is multi-instance erasure. Results of the cross-application erasure are presented in the top half of Fig. 17, demonstrating the generated images after erasing *"Snoopy"*, *"Van Gogh"*, and the two concepts together. Results of the multi-instance erasure are shown at the bottom of Fig. 17, demonstrating the generated images after erasing *"Mouse"*, *"Dog"*, and both concepts. Overall, AdaVD achieves a high

erasure precision. It can be seen from Fig. 17 that, when aiming at a single concept erasure, other concepts specified in the prompt remain faithfully in the generated image; and when aiming at erasing multiple concepts, all the relevant visual content is also removed successfully. This serves as evidence that AdaVD is capable of a robust and precise erasure.

H. Failure Case Study

Despite its success, there exist concepts that AdaVD struggles to erase. We present a few failure cases in Fig. 18. For instance, it is challenging for AdaVD to erase *"Van Gogh"* from a prompt like "The Starry Night is an iconic example of a masterpiece created in Van Gogh style." The challenge is likely to stem from the presence of multiple tokens, *e.g.*, *"Starry Night"*, that is highly coupled with the target concept. In this case, a small value of the scaling hyper-parameter s as used by the shift factor in Eq. (6) is insufficient to eliminate effectively the target semantics across all the relevant tokens. Nevertheless, this issue can be mitigated by doubling s or incorporating additional related target concepts to erase, *e.g.*, *"Starry Night"*, evidenced by the right side of Fig. 18.