Translation of Text Embedding via Delta Vector to Suppress Strongly Entangled Content in Text-to-Image Diffusion Models

Supplementary Material

1. About Baseline method

1.1. Negative Prompt Guidance (NP)

The negative prompt technique (NP) is not proposed in any of the papers, but it is the most widely used method to suppress certain content in the generation process. NP is implemented by replacing unconditional noise in classifier-free guidance with negative condition noise, which can be expressed as follows:

$$\hat{\epsilon}(z_t) = \epsilon(z_t, e_N) + s_{\text{CFG}}(\epsilon(z_t, e_P) - \epsilon(z_t, e_N)), \tag{1}$$

where e_N is negative prompt's embedding, e_P is a positive prompt's embedding, s_{CFG} is guidance scale of classifier-free guidance and $\epsilon(z_t, e_N)$ means negative condition noise.

1.2. SEmantic GuidAnce (SEGA)

The SEmantic GuidAnce (SEGA) is a method that adds semantic guidance to classifier-free guidance by obtaining only the semantic signal from the upper and lower tails of the difference between negative conditional and unconditional noise. Specifically, SEGA can be expressed as follows:

$$\hat{\epsilon}(z_t) = \epsilon(z_t, \emptyset) + s_{\text{CFG}}(\epsilon(z_t, e_P) - \epsilon(z_t, \emptyset)) + \underbrace{\gamma(z_t, e_c)}_{\text{semantic guidance}} + s_m \nu_t, \tag{2}$$

where e_P is the embedding of a positive prompt, such as "a photo of Steve Jobs," and $\emptyset$ means unconditional embedding. e_c is the embedding of a prompt that refers to the content we want to control, such as "glasses" in the Steve Jobs example. This also means negative content in the main paper. And s_m is the momentum scale, ν_t denotes the momentum of semantic guidance. This semantic guidance can be expressed as follows:

$$\gamma(z_t, e_c) = \mu(\psi; s_e, \lambda)\psi(z_t, e_c), \tag{3}$$

$$\text{where } \psi(z_t, e_c) = \begin{cases} \epsilon(z_t, e_c) - \epsilon(z_t, \emptyset) & \text{if positive guidance} \\ -(\epsilon(z_t, e_c) - \epsilon(z_t, \emptyset)) & \text{if negative guidance,} \end{cases} \tag{4}$$

$$\mu(\psi; s_e, \lambda) = \begin{cases} s_e \text{ where } |\psi| \geq \zeta_\lambda(|\psi|) \\ 0 \text{ otherwise,} \end{cases} . \tag{5}$$

where $\zeta_\lambda(|\psi|)$ is the λ-th percentile of ψ and s_e denotes semantic guidance scale.

1.3. Prompt-to-Prompt (P2P)

Prompt-to-Prompt (P2P) is a method to add or remove attributes by directly controlling the attention map of cross attention in stable diffusion, and in this paper, we used the re-weight method to change the weight of the attention map among the methods proposed by P2P.

2. DDIM inversion

2.1. Denoising Diffusion Probabilistic Models(DDPM)

A diffusion model, as proposed in the DDPM paper [9], is a generative model that produces a clean image z_0 from a fully noisy state $z_T \sim \mathcal{N}(0, 1)$ through a sequential denoising process (backward process) $p(z_{t-1} \mid z_t)$ over T steps. The diffusion model is trained by injecting noise into clean images using a noising process (forward process) and learning to predict the noise. The forward process is a fixed process that uses Gaussian noise and is defined with a pre-scheduled α_t, as follows:

$$\begin{aligned} q\left(z_t \mid z_{t-1}\right) &:= \mathcal{N}\left(z_t; \sqrt{\alpha_t} z_{t-1}, \left(1 - \alpha_t\right)\mathbf{I}\right), \\ q(z_T | z_0) &\approx \mathcal{N}(z_T; \mathbf{0}, \mathbf{I}), \end{aligned} \tag{6}$$

where α_t follows a predefined schedule with t ranging from 1 to T. The backward process corresponding to this forward process is expressed as follows:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(z_t, t) \right) + \sigma_t \eta, \text{ where } \eta \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{7}$$

where $\sigma_t^2 = \left(\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \right) (1-\alpha_t)$ denotes the sampling variance, the $\bar{\alpha}_t$ defined as $\prod_{\tau=1}^{t} \alpha_\tau$ and $\epsilon_\theta(z_t, t)$ means the diffusion model.

2.2. DDIM & DDIM inversion

Since DDPM is a Markovian process, it typically requires T sequential steps to generate an image. To address the long sampling time, Denoising Diffusion Implicit Models(DDIM) paper [32] proposed DDIM sampling, a non-Markovian sampling method. A key feature of DDIM is that it allows control over randomness through σ_t, where $\sigma_t = 0$ results in deterministic sampling. The non-Markovian sampling process proposed in the paper is defined as follows:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underset{\text{predicted } x_0}{\left(\frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t)}{\sqrt{\bar{\alpha}_t}} \right)} + \underset{\text{direction pointing to } z_t}{\sqrt{1-\bar{\alpha}_{t-1} - \sigma_t^2}\epsilon_\theta(z_t, t)} + \underset{\text{random noise}}{\sigma_t \eta} , \eta \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{8}$$

When $\{\sigma_t\}_{t=0}^{T} = \{0\}$, the process becomes deterministic, which is known as a DDIM process. In this deterministic setting, an inversion process can be applied to map an image to its corresponding latent representation. This inversion process is defined as follows:

$$z_t = \sqrt{\alpha_t} z_{t-1} + \sqrt{\bar{\alpha}_t} \left(\frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1-\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_{t-1}}} \right) \epsilon_\theta(z_{t-1}, t). \tag{9}$$

This inversion process maps an image to a latent representation, which can be reconstructed into an image through DDIM sampling. Unlike the forward process in the training, which relies on Gaussian noise, DDIM inversion is deterministic and preserves the structure of the original image. This property provides an advantage for learning the features of the attention map, which is important for the task at hand.

3. Local Blending and Identity Preservation

To preserve the original image structure while suppressing negative content, we propose a local blending approach that minimizes unintended distortion.

3.1. Attention Feature Local Blending

We introduce two complementary methods for identifying regions requiring modification: Attention-Based Latent Blending (ABL) and Difference-Based Latent Blending (DBL).

3.1.1. Attention-Based Latent Blending (ABL)

ABL utilizes attention maps from the SSDV-applied text embedding to precisely identify regions containing negative content. For an attention map $A \in \mathbb{R}^{L \times D \times H \times W}$ (where L represents the number of U-Net layers and D the number of attention heads), we aggregate A via mean pooling to obtain $\hat{A} \in \mathbb{R}^{H \times W}$. The ABL mask M_{ABL} is then defined as:

$$M_{\text{ABL}} = \mathbf{1}(\hat{A} > \tau_{\text{ABL}} * \max(\hat{A})), \tag{10}$$

where $\mathbf{1}(\cdot)$ is the indicator function and $\tau_{\text{ABL}} \in [0, 1]$ is a threshold parameter.

3.1.2. Difference-Based Latent Blending (DBL)

DBL generates a binary mask based on the differences in latent space before and after suppression. This approach identifies regions where SSDV modification has the most significant impact. The DBL mask M_{DBL} is defined using the difference $\Delta x = |x_{\text{SSDV}} - x|$ between the SSDV-applied latent x_{SSDV} and the original latent x:

$$M_{\text{DBL}} = \mathbf{1}(\Delta x > Q_{\tau_{\text{DBL}}}(\Delta x)), \tag{11}$$

where $Q_{\tau_{\text{DBL}}}(\Delta x)$ represents the τ_{DBL}-th quantile of the set of all elements $\{\Delta x_i\}$ in Δx, and $\tau_{\text{DBL}} \in [0, 1]$ controls the sensitivity of the mask.

3.1.3. Combined Latent Blending

We combine both approaches by multiplying the masks to obtain the final latent blending mask $M_{\text{latent}} = M_{\text{DBL}} \odot M_{\text{ABL}}$. This ensures that modifications are applied only to regions identified by both methods. The modified SSDV-applied latent $x_{\text{SSDV}'}$ is then computed as:

$$x_{\text{SSDV}'} = x + M_{\text{latent}} \odot (x_{\text{SSDV}} - x),$$

where $\odot$ represents element-wise multiplication. This approach effectively localizes modifications to targeted regions while preserving the original image structure in unaffected areas, thereby minimizing unnecessary distortion in the latent representation.

3.2. Attention Feature Blending (AFB)

AFB is a technique that performs blending using the intermediate features of the cross-attention layer (specifically, the product of the attention map and the value tensor). It calculates the feature mask M_{AFB} using the difference between the SSDV-applied feature $F_{\text{SSDV}} = \text{Attention}(Q, K^*, V^*)$ and the feature without it $F = \text{Attention}(Q, K, V)$:

$$\Delta F = |F - F_{\text{SSDV}}|, \tag{12}$$

$$M_{\text{AFB}} = \mathbf{1}(\Delta F > Q_{\tau_{AFB}}(\Delta F)), \tag{13}$$

$$F_{\text{SSDV}'} = F + M_{\text{AFB}} \odot (F_{\text{SSDV}} - F), \tag{14}$$

where $\tau_{\text{AFB}} \in [0, 1]$ controls the sensitivity of the mask. Blending at the attention feature level allows for more precise control of the suppression level, enabling suppression while maximally preserving identity, as shown in Sec 7.5.

4. Implementation details

In our implementation, we utilized the Stable Diffusion v1.5. In all our experiments, we empirically set $\alpha_k = 1.3$, $\alpha_v = -1.3$, τ_{ABL} = 0.3, τ_{DBL} = 0.9 and τ_{AFB} = 0.35, respectively. For delta optimization experiment, λ_{attn} was set to 0.5, and α was set to 1.0 and τ_{ABL} = 0.5, τ_{DBL} = 0.85, τ_{AFB} = 0.35, respectively. We used the AdamW optimizer with a learning rate of 0.1, $\beta_1 = 0.5$, $\beta_2 = 0.8$, and a weight decay of 0.033.

We used the text prompt ”A photo of *object*” to evaluate our method, and the (*object* $\leftrightarrow$ negative content) pairs in the proposed SEP-Benchmark are as follows: (Steve Jobs $\leftrightarrow$ Glasses), (Camera $\leftrightarrow$ Lens), (Radio $\leftrightarrow$ Dial), (Superman $\leftrightarrow$ Cape), (Backpack $\leftrightarrow$ Zipper), (Bed $\leftrightarrow$ Pillow), (Bicycle $\leftrightarrow$ Pedal), (Shirt $\leftrightarrow$ Button), (candle $\leftrightarrow$ candlelight), (rhinoceros $\leftrightarrow$ horn).

5. Dataset Details

5.1. Validation of strongly entangled pairs

We quantify the entanglement of each subject-content pair by comparing the key/value vectors of the EOT token embeddings for “Steve Jobs with glasses” and “Steve Jobs”. A smaller difference indicates that the embedding for “Steve Jobs” partially encodes information related to “glasses”. We define the *Disentanglement Ratio (D.R.)* as the normalized magnitude of this difference vector, calculated as follows:

$$\|\Delta\| = \|f(\mathbf{e}_{\text{SJ with glasses}}) - f(\mathbf{e}_{\text{SJ}})\|_2 \,, \text{ Disentanglement Ratio} = \frac{\|\Delta\|}{\|\mathbf{e}_{\text{SJ}}\|}, \tag{15}$$

where $\mathbf{e}_{\text{SJ with glasses}}$ is the embedding of the EOT token for “Steve Jobs with glasses”, and $\mathbf{e}_{\text{SJ}}$ is the embedding of the EOT token for “Steve Jobs”. These embeddings are projected using the projection layer f. By averaging across each key and value layer and comparing the results with general words such as “man” rather than “Steve Jobs”, we observe that our subject-content pairs exhibit lower disentanglement, as shown in Table 4.

Content	glasses		lens		dial		cape		zipper		pillow		pedal		button		light		horn	
Subject	Steve Jobs	man	camera	thing	radio	thing	Superman	character	backpack	thing	bed	thing	bicycle	thing	shirts	clothes	candle	thing	rhinoceros	animal
Key *D.R*	0.637	1.038	0.726	1.017	0.873	0.998	0.786	0.920	0.652	1.019	0.788	1.131	0.747	1.102	0.638	0.791	0.490	0.984	0.701	0.997
Value *D.R*	0.594	1.023	0.671	1.029	0.823	0.959	0.711	0.900	0.627	0.991	0.737	1.095	0.686	1.081	0.593	0.754	0.470	0.956	0.627	0.949

Table 4. *Disentanglement Ratio* between subject–negative content pairs in Stable Diffusion, reordered to match the specified pairing.

We also compared the *Disentanglement Ratio* of specific subjects S^* trained in DreamBooth-tuned models with that of general nouns. As shown in Table 5, the trained subjects S^* exhibit strong entanglement with negative content, as indicated by their low *Disentanglement Ratio*.

Content	yellow		cape		berry	
Subject	S^* toy	toy	S^* figure	figure	S^* bowl	bowl
Key *D.R*	0.5311	0.8260	0.6071	0.8901	0.4403	0.6101
Value *D.R*	0.4887	0.7508	0.5962	0.8315	0.4404	0.5817

Table 5. *Disentanglement Ratio* between subject-negative content pairs in DreamBooth-tuned model.

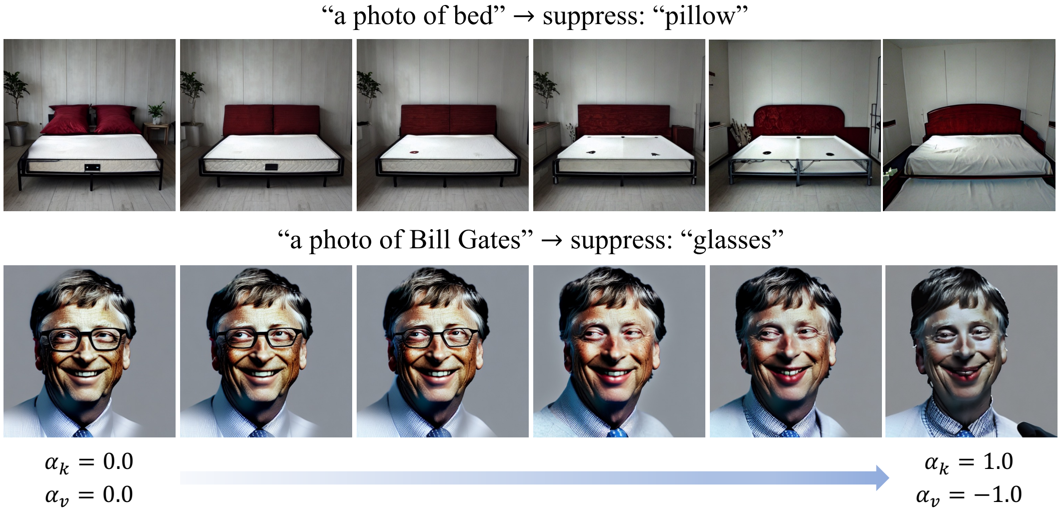
6. Additional Analysis

6.1. Performance of evaluation metrics at various alpha values

We calculated the CLIP-score, IFID for various values of α, a hyperparameter controlling the strength of suppression, in a zero-shot approach. For an accurate comparison, we conducted experiments in this study excluding Local Blending. The results are summarized in Table. 6, reveal a clear trend: as α increases, both the CLIP-score and IFID values rise, indicating stronger suppression. Figure 1 presents the suppression results for each value of α. As α_k increases and α_v decreases, it becomes increasingly evident that the glasses in “a photo of Bill Gates” are being suppressed, demonstrating a stronger tendency for the glasses to gradually disappear.

	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 1.0$
CLIP ↓	17.01	16.93	16.82	16.66	16.62	16.45	16.17
IFID ↑	22.74	26.28	30.30	36.74	45.34	53.88	72.64

Table 6. Evaluation table for the effect of the hyperparameter α, α represents the absolute values of α_k and α_v.


“a photo of bed” → suppress: “pillow”

“a photo of Bill Gates” → suppress: “glasses”

Figure 1. **Visualization of suppression at various alpha values.** The above results correspond to different values of α, with α_k increasing by increments of 0.2 and α_v decreasing by increments of -0.2.

6.2. Delta Optimization with $\alpha_v > 0$

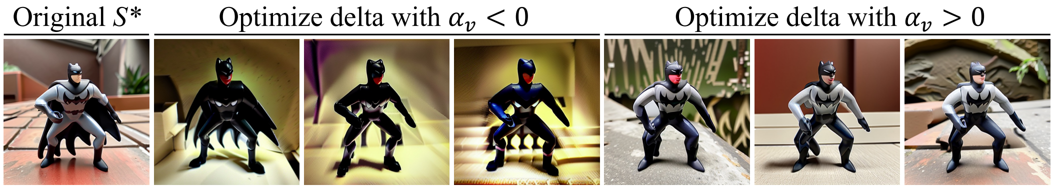

Original $S*$
Optimize delta with $\alpha_v < 0$
Optimize delta with $\alpha_v > 0$

Figure 2. Suppression of “cape” from “S^* figure” when using delta vector optimizing with $\alpha_v > 0$ and $\alpha_v < 0$.

We propose a delta-optimization method to obtain a delta embedding that accurately encodes negative content, which may be distorted during personalized tuning. By optimizing the delta as described in Equation 11 of the main paper, we enable e^*_v to precisely identify the targeted negative content, facilitating its subsequent suppression. Consequently, setting $\alpha_v > 0$ during optimization is appropriate, as verified by our experiment comparing suppression results of “cape” from the S^* figure with $\alpha_v > 0$ versus $\alpha_v < 0$, as shown in Figure 2.

6.3. Attention to Negative Content Regions

We compared the attention maps of e^*_k and “glasses” (two images on the left) to evaluate how accurately the “glasses” region is captured. The map from e^*_k yields a higher IoU, indicating it focuses more precisely on the glasses. However, the “glasses” token’s attention is noisy and spills into the background, leading to structural distortions in the final image (third image on the right). Thus, for strongly entangled content, using e^*_k better preserves the original structure than relying solely on the cross-attention of the “glasses” token.

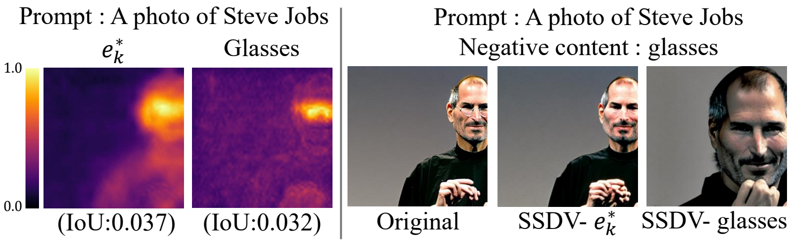


Figure 3. Attention map of e^*_k and “glasses” given the input prompt “A photo of Seteve Jobs” (Left). Results of SSDV using the attention of e^*_k and “glasses” for suppression (Right).

6.4. Additional Image-Quality Evaluation

Although IFID evaluates suppression effectiveness, it is also affected by image quality degradation. To verify that our approach maintains image quality, we additionally evaluate image quality using BRISQUE [20] scores and SSIM [35]. As reported in Table 7, our method consistently attains higher SSIM and lower BRISQUE scores than competing baselines. Our method not only effectively suppresses negative content but also achieves good image-quality preservation compared to baselines that do not perform suppression.

	SD	NP	P2P	SupEOT	SEGA	Inst-Inpaint	Ours
Brisque↓	16.68	18.45	18.59	20.14	17.43	19.67	17.82
SSIM↑	1.0	0.5279	0.5445	0.5019	0.7549	0.5279	0.6968

Table 7. Quantitative comparison of image quality across methods.

7. Additional Results

7.1. Results on Stable Diffusion

An additional qualitative comparison of suppression results with previous methods in Stable Diffusion is presented in Figure 4. These results demonstrate that our method effectively suppresses strongly entangled content during the image genera-

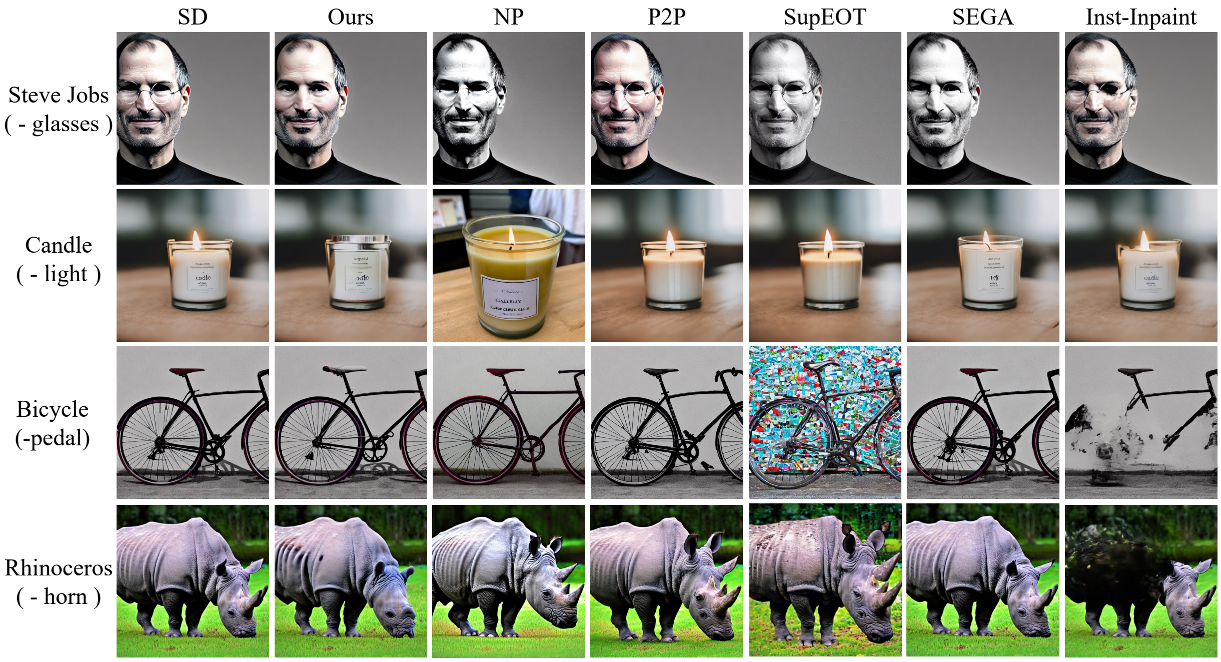
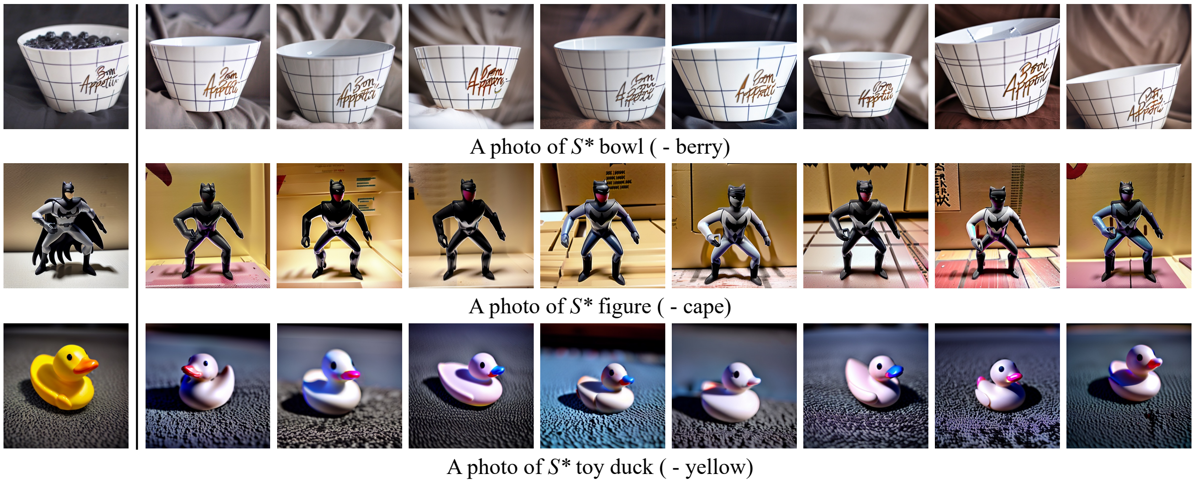


Figure 4. **Comparison of qualitative results with previous methods on the Stable diffusion.** The text in the leftmost column consists of the object to be generated (e.g., "A photo of Steve Jobs") along with the negative prompt (e.g., "glasses") in parentheses.

tion process, which previous methods could not suppress, by separating the negative content from the target word as proposed in Sec. 4.3 of the main paper.

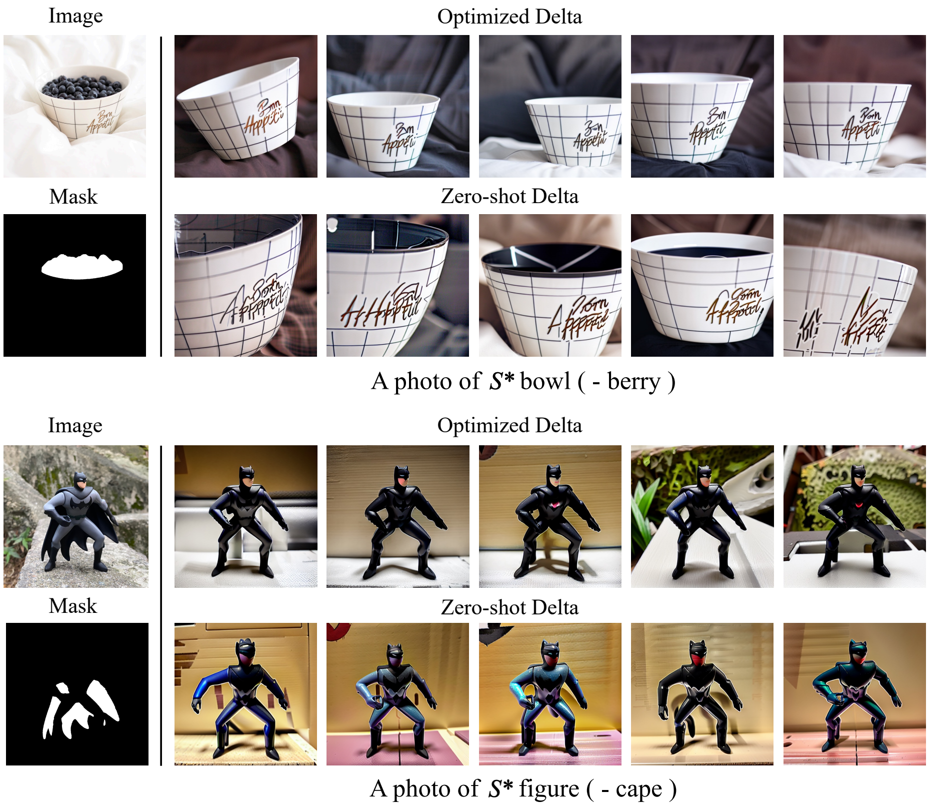
7.2. Results on DreamBooth


A photo of S^* bowl (- berry)

A photo of S^* figure (- cape)

A photo of S^* toy duck (- yellow)

Figure 5. **Our zero-shot suppression results on DreamBooth-tuned model.** The leftmost image represents the original generated output of the DreamBooth-tuned model, while the images on the right show the suppression results of our method when provided with the input prompt shown below each figure, along with the negative content indicated in parentheses.



A photo of S* bowl (- berry)

A photo of S* figure (- cape)

Figure 6. **Optimized delta vs Zero-shot delta on DreamBooth.** The leftmost part shows the subject S^* image used during training to obtain the optimized delta, along with the mask indicating regions of negative content. On the right, we present suppression results obtained using the optimized delta and the zero-shot delta, with input prompts and corresponding negative content indicated in parentheses below each image.

Additional suppression results for the DreamBooth-personalized diffusion model are presented in Figure 5. These results demonstrate the effectiveness of our SSDV method with the zero-shot delta, illustrating successful suppression of negative content across multiple generation seeds.

Figure 6 presents suppression results using the optimized delta on the DreamBooth-personalized model. In the first row, the bowl maintains its original identity while successfully removing the negative content ("berry"). In contrast, the second row illustrates the results using the zero-shot delta, where suppression remains effective, but the bowl's identity is slightly compromised. This behavior was consistently observed in the suppression of the cape from the $S*$ figure.

7.3. Results on CustomDiffusion

Figure 7 presents suppression results for the CustomDiffusion-personalized model. The first column shows images generated from the CustomDiffusion model using the input prompt, while the remaining columns illustrate results after applying various suppression methods. Specifically, the second column demonstrates that our method successfully suppresses negative content, confirming its effectiveness on CustomDiffusion-based models.

We further compare our SSDV method using optimized and zero-shot delta vectors on a CustomDiffusion-based model in Figure 8. The first column shows the subject $S*$ image and the mask indicating negative content used for optimization. In the second and fourth rows, results obtained using the zero-shot delta show effective suppression but slightly compromised subject identity. Conversely, the first and third rows show results using the optimized delta, clearly demonstrating superior preservation of the subject identity along with precise suppression. These results indicate that optimizing the delta vector enables more accurate suppression of targeted negative content.

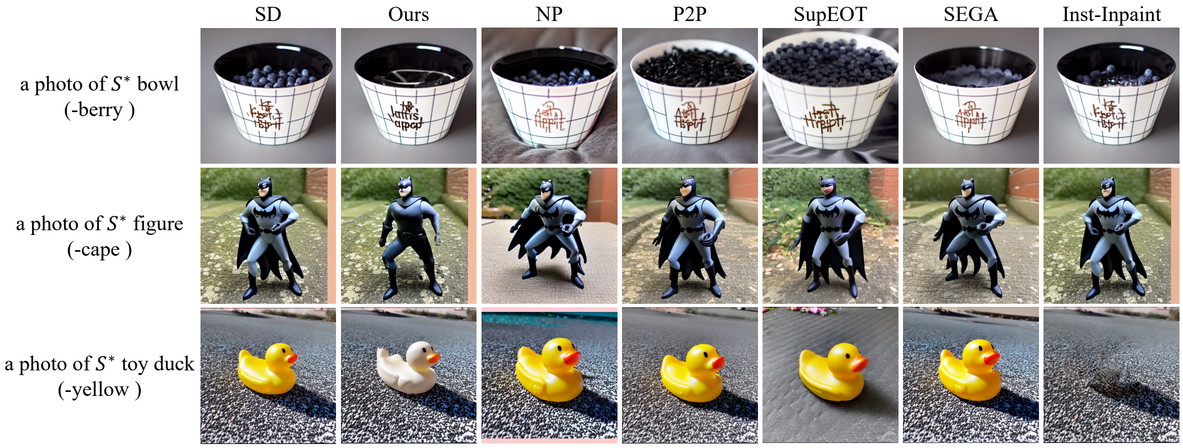
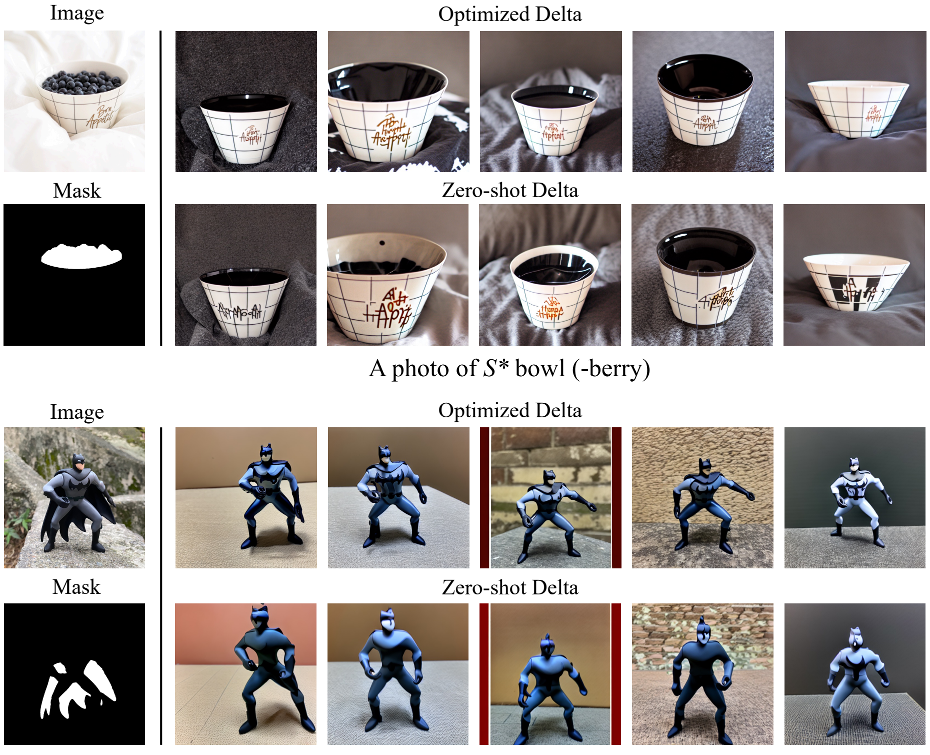

SD
Ours
NP
P2P
SupEOT
SEGA
Inst-Inpaint
a photo of S^* bowl (-berry)
a photo of S^* figure (-cape)
a photo of S^* toy duck (-yellow)

Figure 7. **Qualitative results with other methods on CustomDiffusion-based.** We can suppress the negative content (in the leftmost parentheses) when the input positive prompt (located above parentheses) is given.



A photo of S* bowl (-berry)

A photo of S* figure (-cape)

Figure 8. **Optimized delta vs Zero-shot delta on CustomDiffusion.** The leftmost part shows the subject S^* image and the corresponding mask used during training to obtain the optimized delta, with negative content regions masked out. On the right, we present suppression results using the optimized and zero-shot deltas, with input prompts and targeted negative content indicated in parentheses below each image.

7.4. Comparison with Compositional Diffusion

While Compositional Diffusion [16] implements concept conjunction (AND) and concept negation (NOT) by combining the score functions of each condition during generation, our method directly modulates text embeddings and Cross-attention for suppression. As shown in the Table 8, our method outperforms Compositional Diffusion on CLIP, IFID, and DetScore.

Notably, Compositional Diffusion tends to distort background and structure due to its score-based sampling, whereas our method explicitly controls entangled content within the text embeddings, enabling more effective suppression of strongly entangled content, as shown in Figure 9.

	CLIP↓	IFID↑	DetScore↓	SSIM↑	Brisque↓
Compositional Diffusion	15.94	3.14	0.220	0.4348	16.59
Ours	15.92	87.34	0.113	0.6968	17.82

Table 8. Quantitative comparison between Compositional Diffusion and Our method

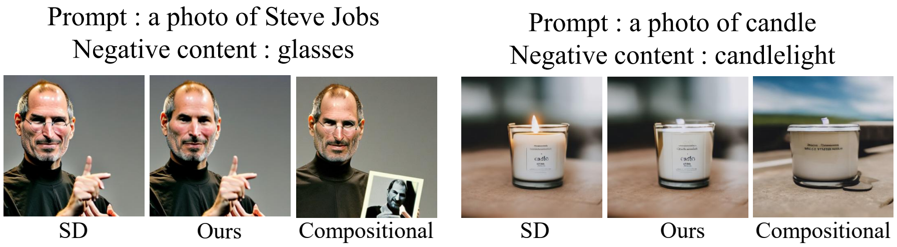

Prompt : a photo of Steve Jobs
Negative content : glasses
SD
Ours
Compositional
Prompt : a photo of candle
Negative content : candlelight
SD
Ours
Compositional

Figure 9. **Qualitative comparison between Compositional Diffusion and our method.** Compositional Diffusion tends to distort the background and structure, whereas our method explicitly controls entangled content and achieves more effective suppression.

7.5. Additional Results of Suppression Content Entangled with Person

We present additional experimental results on persons in Figure 10, demonstrating the suppression of content strongly entangled with specific individuals; for example, suppressing the concept “bald” from an image of Jeff Bezos.

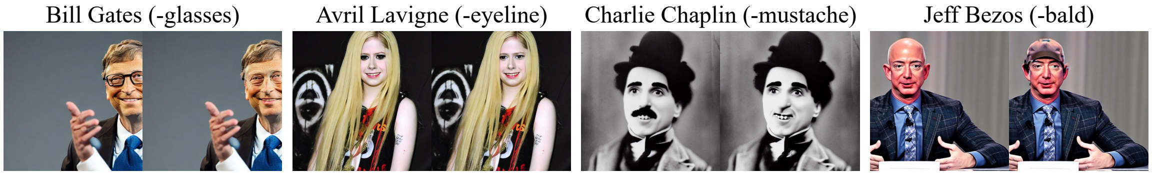


Figure 10. **Results of suppressing content entangled with person.** The right images show the results after suppressing content from the images generated using the prompts on the left. The text on each image indicates the original prompt along with the content to be suppressed in parentheses.

8. Application

8.1. Adding content

Our proposed method enables content addition during image generation by applying the SSDV method with a positive α_v value, which enhances content information through cross-attention. Figure 11 illustrates examples, where the left images are generated from prompts and the right images show results after content addition.

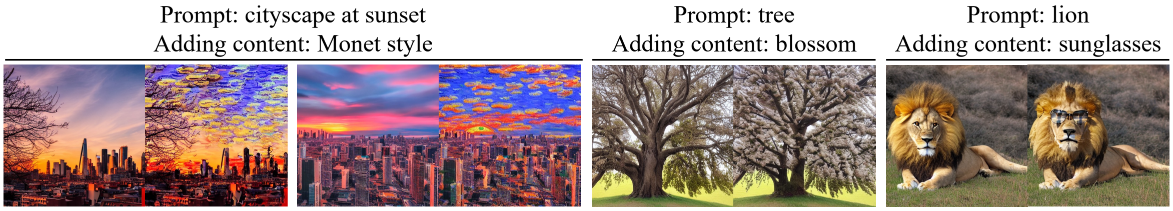

Prompt: cityscape at sunset
Adding content: Monet style
Prompt: tree
Adding content: blossom
Prompt: lion
Adding content: sunglasses

Figure 11. Results of adding an object with SSDV. The right images show the results obtained by adding content to the images generated from the prompts on the left.

8.2. Real-Image Content Supression

We also demonstrate that our method can apply to real images in Figure 12. We employed the Null-text inversion [21] for accurate reconstruction of real images, performing suppression of specific content during the generation process.We also visualised the attention map generated during the process, which showed that our method successfully captured and suppressed negative content, even when editing real images.

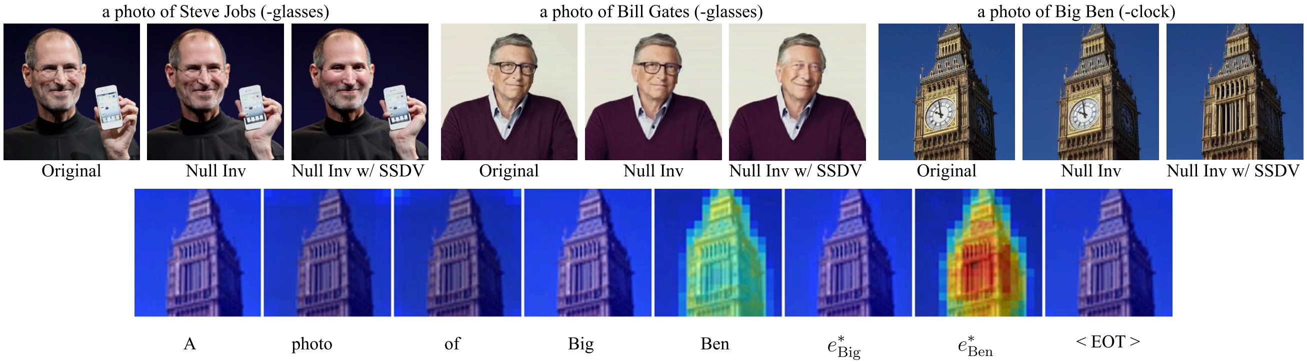
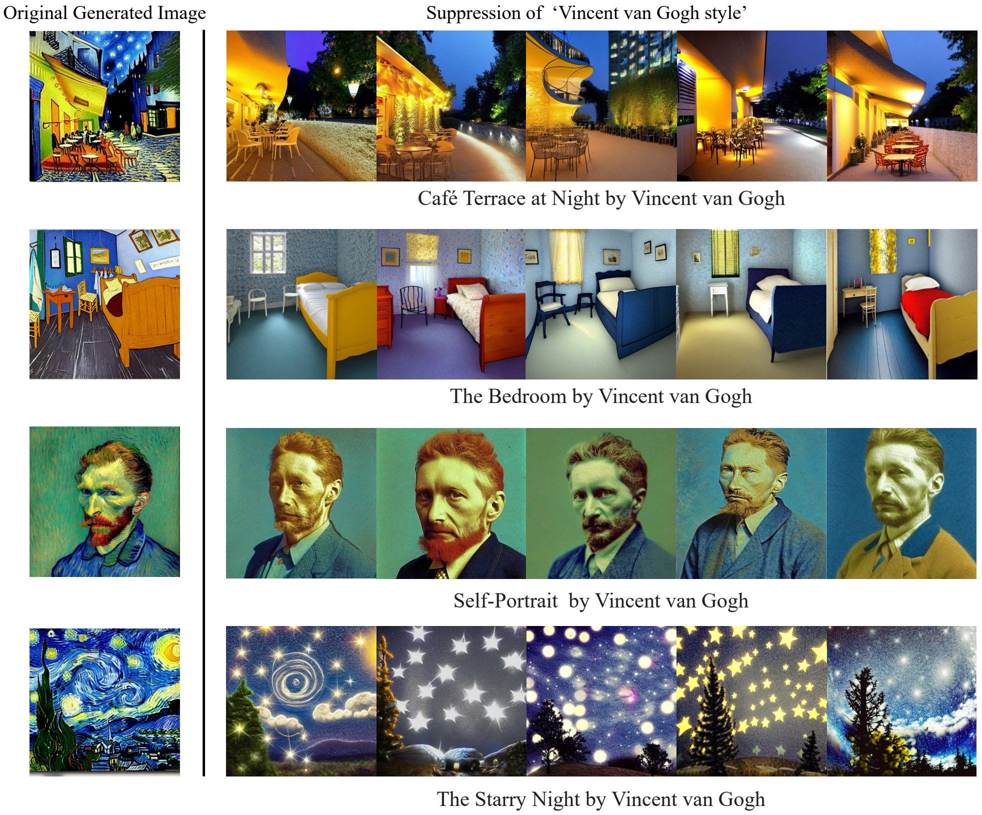


Figure 12. **Suppression on Real Images.** The leftmost image is the original real image, the middle image is the reconstruction obtained using null-text inversion, and the rightmost image shows the result after content suppression using SSDV. Each image is annotated with the input prompt, with the content to be suppressed shown in parentheses. The second row of images shows, via a token-wise attention map, that the delta vector applied to the tokens e^*_{Big} and e^*_{Ben} captures the spatial regions where negative content potentially appears.

8.3. Style Suppression

To explore whether suppression is possible not only for strongly entangled objects but also for styles, we attempted to suppress Vincent van Gogh's style in paintings of Vincent van Gogh generated by Stable Diffusion. The paintings of Vincent van Gogh are strongly entangled with his style, making it challenging to suppress in the generation process. As shown in Figure 13, it appears that suppression of strongly entangled styles is achievable to some extent by our methods. In this experiment, the delta was obtained using the zero-shot approach.



Café Terrace at Night by Vincent van Gogh

The Bedroom by Vincent van Gogh

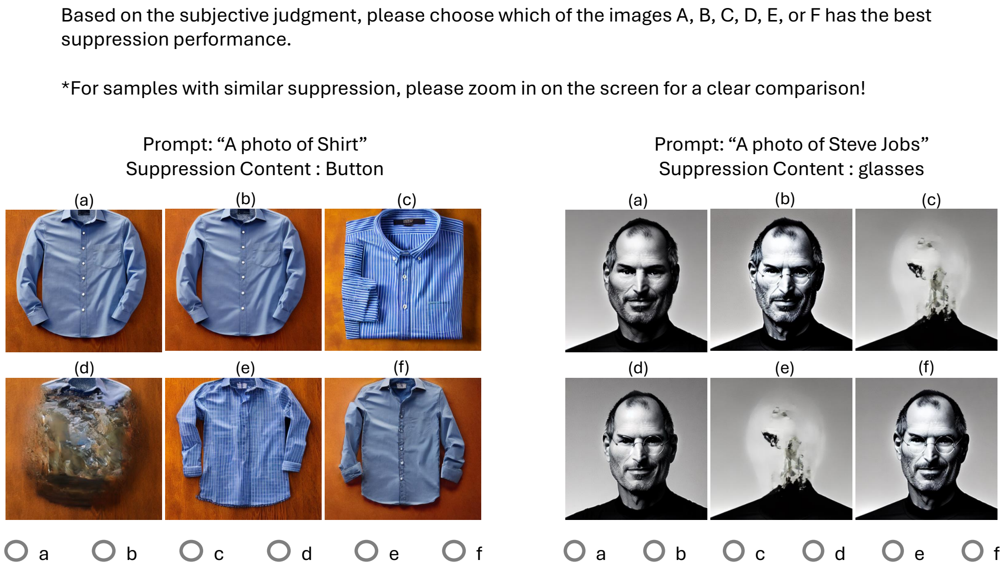
Self-Portrait by Vincent van Gogh

The Starry Night by Vincent van Gogh

Figure 13. **Style suppression from images.** The leftmost column shows the paintings of Vincent van Gogh generated by Stable Diffusion, while the remaining columns display the results after suppressing Vincent van Gogh's style in each image.

9. Details of User Study

To verify the effective suppression of negative content from a target object, we conducted a user study with 50 participants. Participants were presented with the suppression results from our method and other existing methods and were asked to choose the one that most effectively suppressed negative content. We also added some questions to the user study to confirm that the optimization approach performs more effective suppression than the zero-shot approach. The interface of the user study is shown as Figure 14. To ensure a fair comparison, the suppression results of all methods were generated using the same seed. Additionally, all methods were tuned to achieve the best quantitative performance before selecting the samples for comparison.



Based on the subjective judgment, please choose which of the images A, B, C, D, E, or F has the best suppression performance.

*For samples with similar suppression, please zoom in on the screen for a clear comparison!

Prompt: "A photo of Shirt"
Suppression Content : Button

○ a ○ b ○ c ○ d ○ e ○ f

Prompt: "A photo of Steve Jobs"
Suppression Content : glasses

○ a ○ b ○ c ○ d ○ e ○ f

Figure 14. **The user study interface.**

10. Limitation

Our method has some limitations:

- Optimized-based approach takes approximately 30 seconds to optimize and requires the user to provide a mask for the content to be suppressed. Consequently, suppression of elements such as "style" in a zero-shot approach is not feasible.
- The scale of hyperparameters that control the intensity of suppression should be heuristically adjusted for each image sample and content to achieve optimal results.
- Our method is designed for U-Net architectures with cross-attention mechanisms. So, extending to transformer-based models like DiT, which use self-attention and adaptive normalization, poses additional challenges and is left for future work.

We believe that analyzing our approach to address its limitations will help guide future research on the problem of suppressing negative content during the image generation process.