VSF: Simple, Efficient, and Effective Negative Guidance in Few-Step Image Generation Models By Value Sign Flip

Anonymous authors
Paper under double-blind review

Abstract

We introduce Value Sign Flip (VSF), a simple and efficient method for incorporating negative prompt guidance in few-step (1-8 steps) diffusion and flow-matching image and video generation models. Unlike existing approaches such as classifier-free guidance (CFG), NASA, and NAG, VSF dynamically suppresses undesired content by flipping the sign of attention values from negative prompts. Our method requires only a small computational overhead and integrates effectively with MMDiT-style architectures such as Stable Diffusion 3.5 Turbo and Flux Schnell, as well as cross-attention-based models like Wan. We validate VSF on a challenging dataset, NegGenBench, with complex prompt pairs. Experimental results on our proposed dataset show that VSF significantly improves negative prompt adherence (reaching 0.420 negative score for quality settings and 0.545 for strong settings) compared to prior methods in few-step models (scored 0.320-0.380 negative score) and even CFG in non-few-step models (scored 0.300 negative score), while maintaining competitive image quality and positive prompt adherence. Our method is also a suppressed generate-then-edit pipeline, while also having a much faster runtime. Code, ComfyUI node, and dataset will be released.

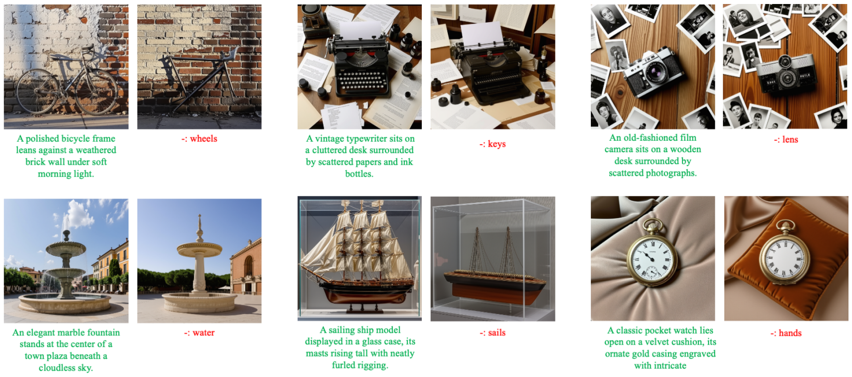

A polished bicycle frame leans against a weathered brick wall under soft morning light.
-: wheels
A vintage typewriter sits on a cluttered desk surrounded by scattered papers and ink bottles.
-: keys
An old-fashioned film camera sits on a wooden desk surrounded by scattered photographs.
-: lens
An elegant marble fountain stands at the center of a town plaza beneath a cloudless sky.
-: water
A sailing ship model displayed in a glass case, its masts rising tall with neatly furled rigging.
-: sails
A classic pocket watch lies open on a velvet cushion, its ornate gold casing engraved with intricate
-: hands

Figure 1: Original image without negative guidance and image generated using our VSF negative guidance on Stable Diffusion 3.5 Large Turbo. The green prompt is the positive prompt, and the red one is the negative prompt. These examples have significant challenges as they are removing essential parts of an object. The “hands” in the last image mean clock hands.

1 Introduction

Diffusion models (including flow matching models) have demonstrated their ability to produce diverse and high-quality images (Black Forest Lab, 2025; Woolf, 2022; Stability AI, 2024) and videos (Wan Team et al., 2025; Yin et al., 2025). However, a longstanding issue remains: the challenge of effectively applying negative guidance in image and video generation. Addressing this problem is crucial for improving content control, moderation (Schramowski et al., 2023), quality assurance, and reducing biases when generating general concepts (Chen et al., 2025a). However, vision language models (VLMs) have difficulties interpreting negations (Park et al., 2025; Alhamoud et al., 2025; Singh et al., 2025; 2024), rendering prompts containing negations ineffectively or made the negative prompt appears even more (e.g., a prompt like “a scientist who is not wearing glasses” will often generate a scientist with glasses—sometimes even more frequently than a simple prompt like “a scientist”). Classifier-free guidance (CFG) (Ho & Salimans, 2022) can be used to address this issue when substituting unconditional generation with negative guidance.

However, to enhance efficiency in image and video generation, numerous models have been distilled to support inference in just a few steps (1-8 steps), such as Flux Schnell (Black Forest Lab, 2025), Stable Diffusion 3.5 Large Turbo (Stability AI, 2024), SDXL Lighting (Lin et al., 2024), SNOOPI (Nguyen et al., 2024), and CausVid (Seppanen; Yin et al., 2025). However, CFG is incompatible with these models. These models are usually distilled and run in CFG-disabled mode, which means only the positive guidance is used, and there is no extrapolation. When CFG is applied forcefully, the resulting image often becomes oversaturated, particularly when the CFG scale is set high enough to suppress unwanted concepts. Moreover, if the number of diffusion steps is too low, the output may reflect features from both the positive and negative prompts (Nguyen et al., 2024), rather than excluding the negative prompt entirely. This occurs due to a divergence between the positive and negative guidance signals (Chen et al., 2025a). An example is shown in Figure 2. Additionally, even if CFG works, it requires two forward passes, one for positive guidance and one for negative guidance, which doubles the run time.

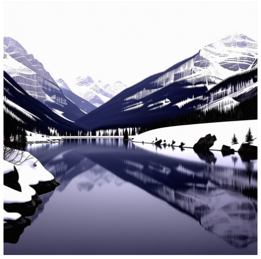


Figure 2: An example of forcefully applying CFG to a step-distilled model is shown using a guidance scale of 2.8 and only 4 steps on SD-3.5-Large Turbo. The positive prompt describes a Canadian winter landscape, while the negative prompt includes the word “lake.” The resulting image does not cancel the lake and exhibits severe oversaturation artifacts (trees in the background).

To address this, two methods, Negative Steer Away Attention (NASA) (Nguyen et al., 2024) and Normalized Attention Guidance (NAG) (Chen et al., 2025a), have been introduced, employing negative guidance within attention final output space rather than the output space. NASA is currently limited to cross-attention models (though it can be re-implemented into other models), while NAG primarily targets quality control rather than negative prompt avoidance. Both methods calculate positive and negative attentions separately and subtract them using a prefixed scale (same as CFG), resulting in a fixed guidance strength throughout the generation, across different areas of the image, and at different layers of the model. This approach lacks adaptability to various time steps, layers, or image regions, limiting effectiveness in negative prompt adherence compared to a more adaptive method (Schramowski et al., 2023; Koulischer et al., 2025; Ban et al., 2024).

In this study, we introduce Value Sign Flip (VSF), a method that dynamically adjusts the guidance strength by flipping the sign of negative prompt values *within* the attention calculation (i.e., not the attention output). This enables the model to steer away from negative concepts adaptively based on their current presence strength, similar to the approach of Koulischer et al. (2025). VSF has a small computational overhead and, when combined with few-step models, facilitates extremely fast image or video generation (¡3 seconds). Our contributions in this work are: (1) we proposed a new method for better negative guidance; (2) we constructed a dataset, NegGenBench, consisting of challenging positive-negative prompt pairs; (3) we collected images generated using these three methods (VSF, NAG, NASA) and labeled their negative and quality score. We further fine-tuned a VLM on it for future work to better evaluate negative prompt following.

2 Related Work

2.1 Classifier Free Guidance

Vision language models struggle to understand negation (Yuksekgonul et al., 2023; Singh et al., 2024; Alhamoud et al., 2025; Park et al., 2025) (We discussed more about this in the Appendix).

Original classifier-free guidance (CFG) (Ho & Salimans, 2022) generates a conditioned noise prediction and an unconditioned noise prediction. In flow matching (Lipman et al., 2023), the predicted targets are the velocity (u_t). Thus, the original flow matching CFG prediction can be written as

$$u_t = f(\emptyset, x_{t+1}, t) + \lambda(f(p^+, x_{t+1}, t) - f(\emptyset, x_{t+1}, t)), \tag{1}$$

where p^+ is the positive prompt, x_t is the latent at time t (where higher t means more torward the noise distribution), $f(\cdot)$ is the trained model, and λ is the guidance scale. Later, the community finds out that by replacing the unconditional generation with a negative prompt (e.g., description of an unwanted image), the model will avoid the prompt due to the negative sign. This is the common implementation of a negative prompt. This turns the above equation into

$$u_t = f(p^-, x_{t+1}, t) + \lambda(f(p^+, x_{t+1}, t) - f(p^-, x_{t+1}, t)), \tag{2}$$

where p^- is the negative prompt.

2.2 Recent Works on Dynamic Negative Guidance

The studies on dynamic negative guidance are very limited (only (Ban et al., 2024; Koulischer et al., 2025; Schramowski et al., 2023)). Ban *et al.* (Ban et al., 2024) found that the negative prompts affect the model by delayed effects and neutralization. After the model has generated unwanted contents, the negative guided output (u_{p^-}) will neutralize the content. They also observed the reverse activation effect, where the negative prompt introduced early in the diffusion processes could actually induce the unwanted concepts. To address this, they proposed applying the negative guidance later in the diffusion process and found it effective.

Schramowski et al. (2023) used a very similar idea as CFG to avoid unwanted (NSFW) content. They generate an unsafe vector and purposely avoid it by subtracting it from the predicted noise. They also added a pixel-level guidance scale that depends on the pixel-wise distance between the positive predicted noise and the unwanted noise, making it adaptive to different regions in the image.

Koulischer et al. (2025) used similar ideas of both and proposed a temporal dynamic guidance scale method. They calculate a probability that the generated concept contains negative content and adjust the guidance scale accordingly. However, their adaptive scale only changes throughout the steps and does not adapt to different regions in the image.

2.3 Few-Step Image Generation Models

Traditional diffusion or flow-matching image generation models typically require many inference steps. However, with improved sampler, this can be reduced to around 20 steps. Recent approaches go further by using step distillation to reduce the number of steps to fewer than 8, or even a single step, as demonstrated in Flux Schnell (Black Forest Lab, 2025), SDXL Lightning (Lin et al., 2024), CausVid (Seppanen; Yin et al., 2025), SNOOPI (Nguyen et al., 2024), and Stable Diffusion 3.5 Turbo (Stability AI, 2024). Since these models are distilled, they generally do not use classifier-free guidance (CFG) during inference; when CFG is forcibly applied, the results are significantly degraded to the point that it is completely unusable (Nguyen et al., 2024), see Figure 2 for an example.

2.4 Recent Works on Negative Guidance in Few-Step Models

Recently, two approaches have specifically targeted negative guidance techniques for few-shot models: Negative-Away Steer Attention (NASA) (Nguyen et al., 2024) and Normalized Attention Guidance (NAG) (Chen et al., 2025a). Although they both focused on avoiding unwanted content and improving quality (using a negative prompt that describes bad quality), NASA mainly focused on avoiding unwanted content, while NAG focused on improving quality.

The authors of the NASA study found that neither standard CFG nor CFG applied directly to text embeddings yields desirable results in few-step scenarios, particularly in single-step settings. Specifically, the regular CFG independently computes positive and negative guidance signals, preventing the negative guidance from effectively neutralizing unwanted concepts. As a result, the produced images merely appear as a mixture of both positive and negative prompts unnaturally (an average image of the positive prompt generated image and the negative prompt independently generated image) rather than excluding negative prompt elements. Furthermore, the authors noted that applying CFG to text embeddings produces minimal benefits. For detailed examples and further illustration, readers could refer to the original paper introducing NASA (Nguyen et al., 2024).

NASA applies the guidance in intermediate states (attention outputs) instead of the final predicted noise or velocity. Specifically, they calculate a positive attention output Z^+ and a negative attention output Z^-, and the final attention Z^{NASA} is obtained by subtracting the two with a factor α, as shown in Equation 3. The alpha value is usually between 0 and 1.

$$Z^{NASA} = Z^+ - \alpha Z^- \tag{3}$$

Normalized Attention Guidance (NAG) used a similar approach. But instead of subtracting the negative attention map from the positive, it uses a similar extrapolation approach as CFG, as shown in Equation 4. The starting point Z^+ could also be replaced with Z^-; they are equivalent if ϕ is increased by 1.

$$\widetilde{Z}^{NAG} = Z^+ + \phi(Z^+ - Z^-) \tag{4}$$

However, to maintain the stability of the attention output space, they also applied normalization to $\widetilde{Z}^{NAG}$ to limit its norm relative to Z^+ with scale τ per token, resulting in $\hat{Z}$. Then it used a blending factor α to blend it with the positive attention result, as shown in Equation 5.

$$Z^{NAG} = \alpha\hat{Z} + (1-\alpha)Z^+ \tag{5}$$

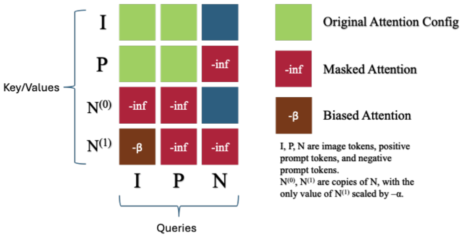

Key/Values
I
P
N(0)
N(1)
I
P
N
Queries
-inf
-inf
-inf
-β
-inf
-inf
Original Attention Config
-inf
Masked Attention
-β
Biased Attention
I, P, N are image tokens, positive prompt tokens, and negative prompt tokens.
N(0), N(1) are copies of N, with the only value of N(1) scaled by –α.

Figure 3: The attention mechanism of our method. We pass in image tokens (I), positive prompt tokens (P), and negative prompt tokens (N) into attention. For key and values, N is duplicated, with values of one copy ($N^{(1)}$) scaled by $-\alpha$. Some areas are masked to avoid interference. An bias $-\beta$ is added to $I \rightarrow N^{(1)}$ attention.

The normalization and blending ensure the attention output of the NAG does not drift away from what the model usually sees during training, improving the quality of generated images. However, if the constraint is set to be too tight (i.e., high α and low τ), it might also limit the model's ability to follow negative prompts.

2.5 Other Related Works

There is previous work that controls attention to mainlutplate images. Attend-and-Excite (Chefer et al., 2023) forces the attention on all key tokens to avoid the generation missing some terms in the prompt. Self-Guidance (Epstein et al., 2023) mantluplate attention to changing elements' properties in the image. BoxDiff (Xie et al., 2023) controls the attention map such that the object appears on the desired location in the image. However, our work is different in that we used it in a way to cancel unwanted elements by simply flipping the sign of values in the negative prompt.

Our work is also related to debiasing in image generation, since it can be used as a way to move away from learned associations (which could be about demography bias or object pairings). Previous work mainly targeted on the input, such as using reference images as input (Zhang et al., 2023) or using a learned prompt. FairQueue (Teo et al., 2024). Our method is different such that we control the value in the negative prompt. Additionally, our method can work against very strong associations like a bike without wheels.

3 Proposed Methods

Our proposed method is built on top of NASA (Nguyen et al., 2024), Koulischer et al. (2025), and Schramowski et al. (2023). NASA has a fixed guidance for every attention calculation, Koulischer et al. (2025) does not have token-level modulation, and Schramowski et al. (2023)'s approach of taking the differences between the positive and negative noise predictions is too simple for complex situations, as mentioned in Koulischer et al. (2025). In NAG's (Chen et al., 2025a) future work section, they also mentioned the possibility of token-level modulation but they did not propose a specific solution.

3.1 Value Sign Flip Adaptive Attention

We propose to expand Koulischer et al. (2025) idea to token-level modulation in few-step models. Let W be a per-token weight at each attention calculation for how strongly the token is associated with a positive concept compared to the negative one. We can modify the NASA attention to Eq. 7. W is obtained by a function with a positive prompt, a negative prompt, and an image as input.

$$W = g(p^+, p^-, I), \tag{6}$$

then we can rewrite the equation in NASA as

$$Z^W = WZ^+ - \alpha(1 - W)Z^- \tag{7}$$

An intuitive method to calculate W involves using the model's attention map: when the image attends more to the negative prompt compared to the positive one, it should be steered away strongly accordingly. Thus, we can calculate the attention map between the image and the positive tokens A^+ and the image and the negative tokens A^- before softmax calculation, then calculate their ratios to their sum. Q is the image query tokens and K^+ and K^- are the positive and negative prompt keys.

$$A^+ = \exp(\frac{Q(K^+)^T}{\sqrt{d}}), A^- = \exp(\frac{Q(K^-)^T}{\sqrt{d}}), W = \frac{\sum A^+}{\sum(A^+ + A^-)} \tag{8}$$

This approach involves complex attention calculation and two attention passes, but it can be implemented by a simpler approach. We can concatenate the values and keys of the positive and negative prompts, then flip the sign of the negative prompt values. This enables that when the image attends to the negative prompt, the flipped value of the negative prompt can cancel the unwanted content. The equation of our method in cross attention models, written in the matrix calculation, is shown in Equation 9, where $\oplus$ means matrix concatenation on the sequence length dimension, σ is the softmax function on the sequence length dimension, and V^+ and V^- are the positive and negative prompt values.

$$Z^{VSF} = \sigma(\frac{Q(K^+ \oplus K^-)^T}{\sqrt{d}})(V^+ \oplus -\alpha V^-) \tag{9}$$

This is similar to noise-canceling headphones, where a "flipped" wave is played to cancel the noise. Note that the key of the negative prompt is not flipped to keep the original meaning of the unwanted concept to match image patches. Mathematically, this is equivalent to Z^W. Proof is in the Appendix.

This approach gives a dynamic weight for the positive and negative prompts, and it varies for different layers, steps, and tokens.

3.2 Attention Masking and Duplication of Negative Embedding

The above method works well for cross-attention-based methods, where attention only exists between image-to-image in self-attention layers and image-to-text in cross-attention layers. However, it requires modification, including masking and duplication, to work in MMDiT-style models such as SD3.5 (Stability AI, 2024), where all image and text tokens are concatenated into a single sequence before attention.

In the standard MMDiT-style setup without our method (e.g., using CFG, NASA, or NAG), the sequence inputs for the attention module are: $[\mathbf{I},\ \mathbf{P}]$ and $[\mathbf{I},\ \mathbf{N}]$. If we concatenate all tokens into a single sequence without any modification, we will get: $[\mathbf{I},\ \mathbf{P},\ \mathbf{N}]$, where $\mathbf{I}$ represents image tokens, $\mathbf{P}$ represents positive prompt tokens, and $\mathbf{N}$ is the negative prompt. During attention, queries, keys, and values are all projected from this combined sequence.

If we apply a sign flip to the negative prompt values by scaling $V_N = V\mathbf{N}$ with $-\alpha$ (where V is the value projection), this flipped content affects all attention paths involving $V_\mathbf{N}$. That includes not only the intended interaction between image and negative prompt $(\mathbf{I} \rightarrow \mathbf{N})$[1], but also undesired

[1] The arrow direction is the attention direction, or the opposite direction of the information flow

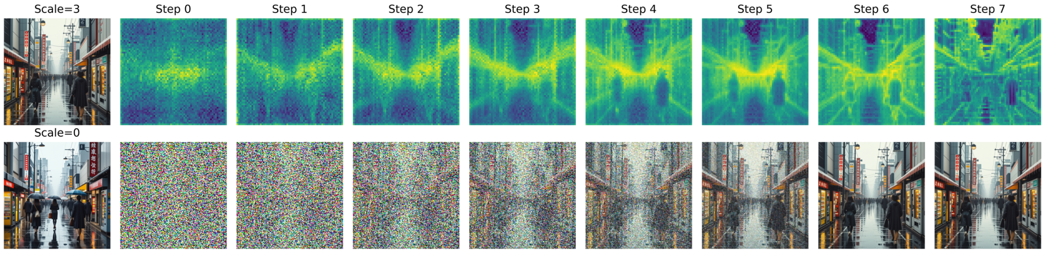


Figure 4: Attention maps and intermediate images during the diffusion process. The leftmost column shows the final generated image (top) and an image generated without applying VSF scaling ($\alpha = 0$, bottom). The top row on the right side displays the unnormalized attention values between image tokens and negative prompt tokens, while the bottom row shows the corresponding intermediate images at each timestep. The negative prompt is "umbrella."

interactions such as positive-to-negative ($\mathbf{P} \rightarrow \mathbf{N}$) and negative-to-negative ($\mathbf{N} \rightarrow \mathbf{N}$) (in which the value will cancel itself). These unintended interactions can distort the behavior of the model since the flipped signal influences more than just the image.

To address this, we introduce a duplication of the negative prompt. One copy remains unflipped and unscaled, denoted by $\mathbf{N}^{(0)}$, and the value (and only value) of the other is flipped and scaled, denoted by $V_{\mathbf{N}^{(1)}} = -\alpha \cdot V_{\mathbf{N}^{(1)}}$. The sequence becomes: $[\mathbf{I},\ \mathbf{P},\ \mathbf{N}^{(0)},\ \mathbf{N}^{(1)}]$, where $\mathbf{N}^{(1)}$ does not act as query in attention calculation.

Additionally, inspired by Wang et al. (2025), where blocking some attention directions could improve quality, we apply attention masks to isolate the effect of $\mathbf{N}^{(1)}$ to only $\mathbf{I}$. Specifically, $\mathbf{N}^{(0)}$ is only allowed to attend to $\mathbf{I}$ and to itself, while $\mathbf{N}^{(1)}$ is only attended to by $\mathbf{I}$. Figure 3 shows the attention mask. Since $\mathbf{N}^{(1)}$ does not act as a key or value in any attention query, it doesn't produce associated output. Instead, $\mathbf{N}^{(0)}$ serves as the negative prompt tokens passed to the subsequent MLP layer and into the next attention layer, where it will be flipped again. It acts as an information collector from images to collect unwanted elements and also keeps updating itself from attention to itself, matching the prompt updating in a positive prompt.

This setup allows updates to the negative prompt based on attention from the image and from itself, and keeps the unflipped form active in the MLP path. It also prevents interference between positive and negative prompts and ensures that the flipped negative content affects only the intended image-to-negative attention path.

To preserve the high quality of generated images, we also applied attention bias ($-\beta$) to $\mathbf{I} \rightarrow \mathbf{N}^{(1)}$ (also shown in Figure 3) and we removed the padding tokens from the negative prompt. Details and pseudo-code of our method are in the Appendix.

4 Experiments

4.1 Dataset

Following Park et al. (2025), we use ChatGPT o3 (Open AI) to generate pairs of prompts and negative prompts to construct our dataset NegGenBench. Unlike prior work, our prompts are intentionally more challenging: the negative prompt is typically related to the positive one, and as a critical component—e.g., the positive prompt of a bike could have a negative prompt of "wheels". However, the positive prompt sometimes also uses a non-negation method to imply the item is missing, such as using terms like "empty" and "exposed" to make it more natural. Besides prompts, two questions are generated at the same time for later evaluation, one question asking if the image has the main object, either with or without the negative element, and the other one queries if the negative prompt element is missing. Prompts are generated in batches. There are 200 prompts generated, and we run them with 2 different seeds for the main results.

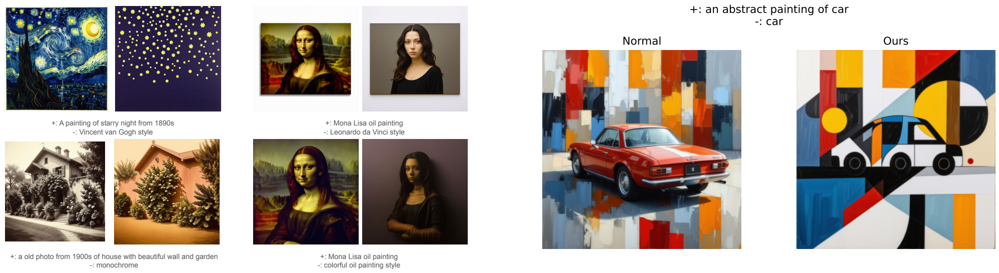

+: A painting of starry night from 1890s
-: Vincent van Gogh style
+: Mona Lisa oil painting
-: Leonardo da Vinci style
+: a old photo from 1900s of house with beautiful wall and garden
-: monochrome
+: Mona Lisa oil painting
-: colorful oil painting style
+: an abstract painting of car
-: car
Normal
Ours

Figure 5: (Left) Style Avoidance Tests, (Right) (Semi)-abstract art generated by mentioning the main object “car” in a negative prompt. The car is semi-canceled and thus still present but in an abstract form.

4.2 Baseline and Metrics

We chose NAG (Chen et al., 2025a) and NASA (Nguyen et al., 2024) as our baseline for few-step models. We also used a base model without negative guidance as a vanilla baseline, aiming to show the lower bound of the dataset. (i.e., how likely the positive prompt alone will help avoid negative concepts, if there is no negative guidance. This could happen either because the model is following the implication in the positive prompt, such as the word ‘missing’, or simply by chance.) Because NASA’s original source code was not publicly available at the time of writing, we reimplemented it based on NAG’s codebase. Specifically, we replaced the guidance equation from NAG (Eq. 4) with NASA’s equation (Eq. 3), removed normalization and blending, and enabled guidance when the scale is greater than 0 (instead of 1). Additionally, we compared our method in non-few-step models with CFG and used other models as external baselines (External baseline results are in Table 1, but the experiment details are in the Appendix). We included Flux Kontext Labs et al. (2025) as an image-editing baseline. We first generate an image using SD-3.5-Large-Turbo, and then edit the image with Flux Kontext using the prompt `Remove [negative prompt]`. Since NAG was focused on quality instead of negative prompt avoidance, we re-tuned its hyperparameter such that it has stronger negation following in trade-off of quality and positive prompt following. We name this variance as NAG Strong. Same for our VSF method, we provided two different variations with different hyperparameters, focusing on quality (VSF Quality) and negative prompt following (VSF Strong). Hyperparameter details are in the Appendix.

Following Park et al. (2025); Wei et al. (2025), we used multimodal large language models (MLLM), specifically `llama-4-maverick-17b-128e-instruct-fp8`, to evaluate if the generated image follows the positive prompt and the negative prompt using the two questions generated during prompt generation. We did not use previous negation-aware CLIP-based work because they do not focus on missing an essential component, but simple meaning (e.g., a dog that is not on the grass). We did not use HPSv2 (Wu et al., 2023) or ImageReward (Xu et al., 2023) because they might give a low quality score for unusual objects (essential part being removed). Instead, MLLM is used to rate the image quality at the same time. At the end of our experiment, we also fine-tuned a Qwen-2.5-VL (Bai et al., 2025) model using data we generated by VSF, NAG, and NASA for better negation understanding. More details about the metrics and comparison with human validation are in the Appendix.

5 Results

Quantitative results from using LLaMA as a judge evaluation are shown in Table 2, and qualitative results are shown in Figure 13 in the Appendix. Human validation is shown in Table 3. Automatic evaluation using the better negation-aware MLLM Qwen-2.5-VL is shown in Table 8 in the Appendix. Both the human validation and Qwen-2.5-VL results are aligned with our LLaMA evaluation relative ranking. It is important to highlight that the LLaMA assigns relatively generous quality scores; a score lower than 90 usually means the image already has degraded quality. Examples are in the Appendix Figure 12.

Based on the quantitative results, VSF Strong shows a significantly higher negative score than other methods, while maintaining comparable or better quality scores. Our more conservative method,

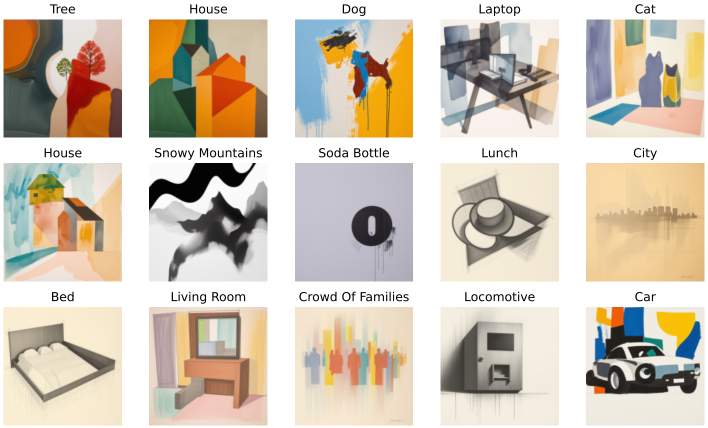


Figure 6: More Semi-Abstract Images Generated

Table 1: External Baselines Comparsion

	Positive Score (↑)	Negative Score (↑)	Quality Score (↑)	~Runtime (↓)
Open-weight Models				
VSF Strong	0.870	**0.545**	0.952	**3s**
VSF Quality	0.980	0.420	**0.986**	3s
Generate+Edit	0.875	0.488	0.958	55s
Janus-4o	0.925	0.225	0.944	20s
Qwen-Image NP	0.973	0.190	0.935	110s
Qwen-Image Negation	**0.990**	0.100	0.937	110s
Closed-weight Models				
GPT-4o	0.978	**0.705**	0.954	47s
Nano Banana	**0.985**	0.498	**0.980**	**14s**

VSF Quality, still achieved the second-highest negative score, with the highest quality score. Both VSF Strong and VSF Quality even achieve a higher negative score than traditional CFG in non-few-step models, demonstrating a stronger ability to avoid negative elements, even relative to the established strong baseline. When compared with the external baseline, VSF also gets the highest performance in open source methods and only lags behind GPT-4o and achieves comparable performance with Nano Banana.

We also tested concepts/style avoidance, as shown in Figure 5 left. In the Starry Night example, VSF completely removed any signature elements in Vincent van Gogh's style, including the town, and generated a generic starry night image. Figure 5 also illustrates how our method can produce abstract art, which is typically discouraged during a model's finetuning since reward models favor realism. This is achieved through using the same word as the main object for both positive and negative scores in VSF, as detailed in the Appendix. VSF also has the ability to generate "anti-aesthetics" (unconventional, including abstract) art. Details of these results are provided in Figure 6 and Figure 18 (in the Appendix).

6 Discussion

6.1 Trade Off Curve

To systematically evaluate how effectively each model balanced positive prompt adherence, negative prompt adherence, and image quality, we conducted a hyperparameter sweep across each model. Specifically, we performed 66 runs for VSF and 287 runs for NAG, and 10 runs for NASA, with respect to their hyperparameter counts (2 for VSF, 3 for NAG, and 1 for NASA). A random sweep was executed besides for NASA, on which a single variable "grid" search is used, and evaluations were conducted using LLaMA, following the same criteria as previously described. Due to the large volume of runs, we limited our evaluation to the first 100 prompts with a single generation seed, potentially resulting in minor differences from earlier outcomes. Results are shown in Figure 7 Left.

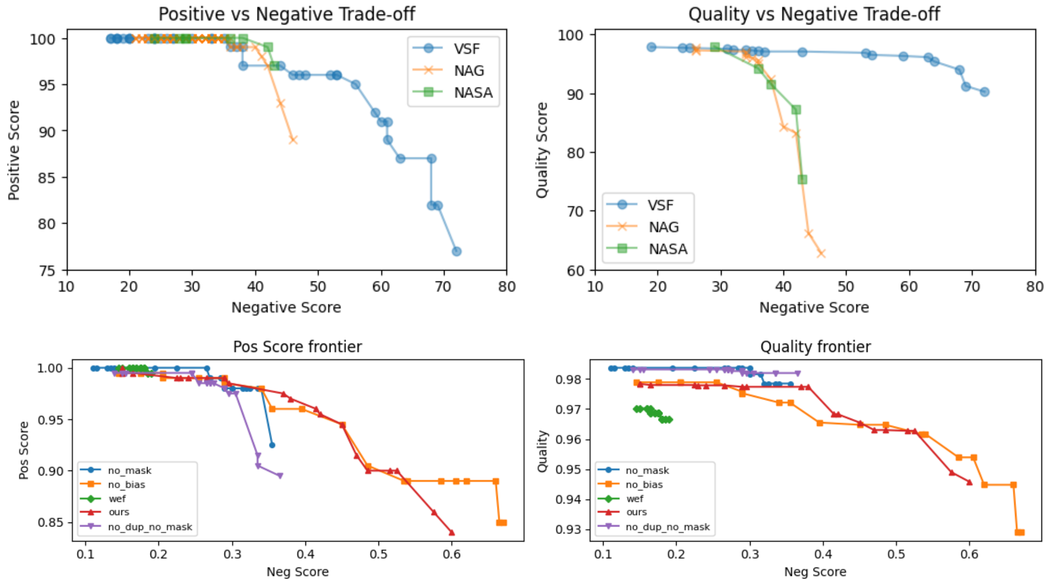


Figure 7: (Top) Trade-off plot of positive-negative score and quality-negative score. Both axies follows "higher is better." (Down) Trade-off plot of the ablation study.

From both plots, we observe that as the negative score increases, NAG and NASA both exhibit a significantly steeper and earlier decline compared to VSF in both positive and quality scores. In terms of positive score, VSF maintains scores above 90 even when the negative score rises to approximately 60. Regarding image quality, VSF similarly retains scores above 90 until a negative score of around 60, after which quality declines. In contrast, NAG and NASA both experience a sharp and early decline, with their quality score rapidly dropping to nearly 60 even before the negative score reaches 50. Keep in mind that a quality score under 90 means the image is already degraded, and if an image is rated 60, it is usually completely distorted. See Figure 12 in the Appendix for example.

Additionally, VSF demonstrates a broader operational range in negative scores. When necessary, it can achieve negative scores exceeding 70 while still preserving acceptable positive prompt adherence and image quality. Conversely, NAG and NASA become unacceptable in quality at negative scores below 50, limiting their practical effectiveness.

6.2 Attention Maps

Since our proposed method performs adaptive steering based on a negative attention map, we visualize the attention maps generated during the diffusion process in Figure 4. Extracting the full attention maps is difficult because efficient implementations, such as FlashAttention, do not explicitly store these maps, and storing and computing them will require a large amount of memory. Therefore, we computed only the unnormalized attention values between the image tokens and negative prompt tokens. Figure 4 demonstrates that when the scale is set to 0, umbrellas appear, whereas setting the scale to 3 effectively avoids them. As indicated in the attention maps, image tokens corresponding to regions where umbrellas might exist (e.g., above human heads) exhibit higher attention toward the negative prompt tokens. Specifically, in steps 4 and 5, regions above the individuals on the left and right show strong negative attention, aligning with areas visually identified as umbrellas in $\alpha = 0$ image. In the final image, these highlighted regions no longer contain umbrellas, confirming that our method effectively suppresses the presence of undesired objects at specific locations.

6.3 Ablation Study

To evaluate the effectiveness of each component of our approach, we conducted an ablation study using the following settings. For each setting, we scanned across scales for all 200 prompts using the same seed. Similar to before, we plotted the trade-off curve for each setting.

Table 2: Positive scores (how well the model follows the positive prompts) and negative scores (how well the model avoids the negative prompts) of our model (VSF), NAG (Chen et al., 2025a), and NAG with hyperparameter re-tuned (NAG Strong).

	Positive Score (↑)	Negative Score (↑)	Quality Score (↑)
VSF Strong	0.870	**0.545**	0.952
VSF Quality	0.980	0.420	**0.986**
NAG (Chen et al., 2025a)	0.993	0.220	0.968
NAG Strong	0.975	0.320	0.901
NASA(Nguyen et al., 2024)	0.970	0.380	0.867
None	0.990	0.195	0.968
CFG (Ho & Salimans, 2022) (28 steps)	**1.000**	0.300	0.956

Table 3: Human Labelled Metric For 10 Selected Prompts with 2 Seeds

	Positive Score (↑)	Negative Score (↑)	Quality Score (↑)
NAG Strong	0.950	0.250	0.675
NAG	**1.000**	0.100	**0.895**
NASA	0.950	0.150	0.685
VSF Quality	0.900	**0.550**	0.823

Rather than altering the attention values, we explored a simpler and more intuitive approach: flipping the text embedding prior to input into the DiT (Whole Embedding Flip, WEF). This is similar to applying the CFG on text embeddings studied in Nguyen et al. (2024), but keeps the positive and negative tokens separated. Specifically, the negative text embedding is scaled by $-\alpha$, concatenated with the positive prompt embedding in the sequence length dimension, and used as the prompt embedding for the DiT. We did not remove the padding for the negative prompt, as we found out that removing it causes the negative prompts to have no effect at all.

We also tested our approach with no bias, no mask (but still duplication), and no duplication no mask. The trade-off plot is shown in Figure 7 Right. The simpler and more intuitive WEF approach appears to have almost no effect at all. We hypothesize that this is because it is similar to flipping both the key and the value, causing regions most similar to the flipped key (i.e., least similar to the original negative prompt) to be pushed away, rather than pushing away regions most similar to the original negative prompt (i.e., unflipped key). From the figure, we can see that the configurations without masking have a sharp positive score drop as the negative score increases. The WEF has a very limited range of negative scores. Our methods and the one without attention bias have similar results; this could be due to the MLLM not being sensitive enough to see the minor changes in quality.

Ablation study on hyperparameters is shown in the Appendix.

7 Conclusion

In this paper, we introduced VSF, a novel approach for enhancing negative prompt adherence in image and video generation models. Our method involves flipping the sign of attention values and duplicating negative prompts and attention masking, effectively suppressing unwanted content. Experimental results indicate that VSF significantly outperforms previous methods, NAG (Chen et al., 2025a), NASA (Nguyen et al., 2024), and even CFG in terms of negative prompt adherence, with much lower trade-offs in overall quality and positive prompt following. We also showed that VSF can be applied to create more creative (by style avoidance, abstract images, and anti-aesthetics styles) images. VSF also only has one main hyperparameter and one minor hyperparameter, making it easier to tune them in downstream tasks. Future work directions are discussed in the Appendix.

8 Reproducibility Statement

All code, dataset, and fine-tuned models (NegAwareQwen) will be released after publication.

REFERENCES

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-Language Models Do Not Understand Negation, May 2025. URL http://arxiv.org/abs/2501.09425. arXiv:2501.09425.

Anne Arzberger, Stefan Buijsman, Maria Luce Lupetti, Alessandro Bozzon, and Jie Yang. Nothing Comes Without Its World – Practical Challenges of Aligning LLMs to Situated Human Values through RLHF. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:61–73, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31617. URL https://ojs.aaai.org/index.php/AIES/article/view/31617.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, February 2025. URL http://arxiv.org/abs/2502.13923. arXiv:2502.13923.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the Impact of Negative Prompts: When and How Do They Take Effect?, June 2024. URL http://arxiv.org/abs/2406.02965. arXiv:2406.02965.

Black Forest Lab. black-forest-labs/FLUX.1-schnell · Hugging Face, June 2025. URL https://huggingface.co/black-forest-labs/FLUX.1-schnell.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models, May 2023. URL http://arxiv.org/abs/2301.13826. arXiv:2301.13826 [cs].

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Normalized Attention Guidance: Universal Negative Guidance for Diffusion Models, June 2025a. URL http://arxiv.org/abs/2505.21179. arXiv:2505.21179.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. ShareGPT-4o-Image: Aligning Multimodal Models with GPT-4o-Level Image Generation, June 2025b. URL http://arxiv.org/abs/2506.18095. arXiv:2506.18095.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling, January 2025c. URL http://arxiv.org/abs/2501.17811. arXiv:2501.17811.

Tri Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning, July 2023. URL http://arxiv.org/abs/2307.08691. arXiv:2307.08691.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs, May 2023. URL http://arxiv.org/abs/2305.14314. arXiv:2305.14314.

Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion Self-Guidance for Controllable Image Generation, June 2023. URL http://arxiv.org/abs/2306.00986. arXiv:2306.00986 [cs].

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, March 2024. URL http://arxiv.org/abs/2403.03206. arXiv:2403.03206.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing Concepts from Diffusion Models, June 2023. URL http://arxiv.org/abs/2303.07345. arXiv:2303.07345.

Ashish Goswami, Satyam Kumar Modi, Santhosh Rishi Deshineni, Harman Singh, Prathosh A. P, and Parag Singla. GraPE: A Generate-Plan-Edit Framework for Compositional T2I Synthesis, March 2025. URL `http://arxiv.org/abs/2412.06089`. arXiv:2412.06089.

Wenqi Marshall Guo, Yiyang Du, Heidi J. S. Tworek, and Shan Du. Position: The Pitfalls of Over-Alignment: Overly Caution Health-Related Responses From LLMs are Unethical and Dangerous, August 2025. URL `http://arxiv.org/abs/2509.08833`. arXiv:2509.08833.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance, July 2022. URL `http://arxiv.org/abs/2207.12598`. arXiv:2207.12598.

Felix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester, and Luca Ambrogioni. Dynamic Negative Guidance of Diffusion Models, January 2025. URL `http://arxiv.org/abs/2410.14398`. arXiv:2410.14398.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space, June 2025. URL `http://arxiv.org/abs/2506.15742`. arXiv:2506.15742.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, February 2022. URL `http://arxiv.org/abs/2201.12086`. arXiv:2201.12086.

Shanchuan Lin, Anran Wang, and Xiao Yang. SDXL-Lightning: Progressive Adversarial Diffusion Distillation, March 2024. URL `http://arxiv.org/abs/2402.13929`. arXiv:2402.13929.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling, February 2023. URL `http://arxiv.org/abs/2210.02747`. arXiv:2210.02747.

Anne-Sofie Maerten, Li-Wei Chen, Stefanie De Winter, Christophe Bossens, and Johan Wagemans. LAPIS: A novel dataset for personalized image aesthetic assessment, 2025. URL `https://arxiv.org/abs/2504.07670`. Version Number: 1.

Viet Nguyen, Anh Nguyen, Trung Dao, Khoi Nguyen, Cuong Pham, Toan Tran, and Anh Tran. SNOOPI: Supercharged One-step Diffusion Distillation with Proper Guidance, December 2024. URL `http://arxiv.org/abs/2412.02687`. arXiv:2412.02687.

Open AI. Introducing OpenAI o3 and o4-mini. URL `https://openai.com/index/introducing-o3-and-o4-mini/`.

Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. Know ”No” Better: A Data-Driven Approach for Enhancing Negation Awareness in CLIP, March 2025. URL `http://arxiv.org/abs/2501.10913`. arXiv:2501.10913.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL `http://arxiv.org/abs/2103.00020`. arXiv:2103.00020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, September 2023. URL `http://arxiv.org/abs/1910.10683`. arXiv:1910.10683 version: 4.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models, April 2023. URL `http://arxiv.org/abs/2211.05105`. arXiv:2211.05105.

Jukka Seppanen. Kijai/WanVideo_comfy · Hugging Face. URL https://huggingface.co/Kijai/WanVideo_comfy.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn "No" to Say "Yes" Better: Improving Vision-Language Models via Negations, March 2024. URL http://arxiv.org/abs/2403.20312. arXiv:2403.20312.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learning the Power of "No": Foundation Models with Negations. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8002–8012, February 2025. doi: 10.1109/WACV61041.2025.00777. URL https://ieeexplore.ieee.org/document/10943641. ISSN: 2642-9381.

Stability AI. Introducing Stable Diffusion 3.5, 2024. URL https://stability.ai/news/introducing-stable-diffusion-3-5.

Margit Sutrop. Challenges of Aligning Artificial Intelligence with Human Values. *Acta Baltica Historiae et Philosophiae Scientiarum*, 8(2):54–72, December 2020. ISSN 22282009, 22282017. doi: 10.11590/abhps.2020.2.04. URL https://www.ies.ee/bahps/acta-baltica/abhps-8-2/04_Sutrop-2020-2-04.pdf.

Christopher T. H. Teo, Milad Abdollahzadeh, Xinda Ma, and Ngai-man Cheung. FairQueue: Rethinking Prompt Learning for Fair Text-to-Image Generation, October 2024. URL http://arxiv.org/abs/2410.18615. arXiv:2410.18615 [cs].

Alexey Turchin. Ai alignment problem: Human values don't actually exist. 2019. URL https://philarchive.org/rec/TURAAP.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, Steven Liu, William Berman, Yiyi Xu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. URL https://github.com/huggingface/diffusers.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models, April 2025. URL http://arxiv.org/abs/2503.20314. arXiv:2503.20314.

Luozhou Wang, Yijun Li, Zhifei Chen, Jui-Hsien Wang, Zhifei Zhang, He Zhang, Zhe Lin, and Yingcong Chen. TransPixeler: Advancing Text-to-Video Generation with Transparency, January 2025. URL http://arxiv.org/abs/2501.03006. arXiv:2501.03006.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. TIIF-Bench: How Does Your T2I Model Follow Your Instructions?, June 2025. URL http://arxiv.org/abs/2506.02161. arXiv:2506.02161.

Max Woolf. Stable Diffusion 2.0 and the Importance of Negative Prompts for Good Results, November 2022. URL https://minimaxir.com/2022/11/stable-diffusion-negative-prompt/.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image Technical Report, August 2025. URL http://arxiv.org/abs/2508.02324. arXiv:2508.02324.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis, September 2023. URL http://arxiv.org/abs/2306.09341. arXiv:2306.09341.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. BoxDiff: Text-to-Image Synthesis with Training-Free Box-Constrained Diffusion, August 2023. URL http://arxiv.org/abs/2307.10816. arXiv:2307.10816 [cs].

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation, December 2023. URL http://arxiv.org/abs/2304.05977. arXiv:2304.05977.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-Grained Multi-Dimensional Human Preference Learning for Image and Video Generation, March 2025. URL http://arxiv.org/abs/2412.21059. arXiv:2412.21059 [cs].

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, Conghui He, and Weijia Li. Echo-4o: Harnessing the Power of GPT-4o Synthetic Images for Improved Image Generation, August 2025. URL http://arxiv.org/abs/2508.09987. arXiv:2508.09987.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models, January 2025. URL http://arxiv.org/abs/2412.07772. arXiv:2412.07772.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI, June 2024. URL http://arxiv.org/abs/2311.16502. arXiv:2311.16502.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, March 2023. URL http://arxiv.org/abs/2210.01936. arXiv:2210.01936.

Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. ITI-GEN: Inclusive Text-to-Image Generation, September 2023. URL http://arxiv.org/abs/2309.05569. arXiv:2309.05569 [cs].

Appendix

Table of Contents

Appendix

A Negation in Vision Language Models

Much previous work has shown that existing vision language models (VLM) struggle to understand negation (Yuksekgonul et al., 2023; Singh et al., 2024; Alhamoud et al., 2025; Park et al., 2025). In classification tasks, the model cannot correctly understand text with negation in it, e.g. “a dog running” vs “a dog not running” might have very close embeddings, even though they are opposite. In our test using a CLIP-ViT-Base-32, the cosine similarity between “a dog running” and “a dog not running” is 0.9243, where the similarity between “a dog” and “a dog running” is only 0.8710. In Figure 8, we show 4 prompts “a photo of a bike”, “a photo of a bike without wheels”, “a photo of a bike with wheels”, and “a photo of a car with wheels”. We can see that the bike with wheels and the

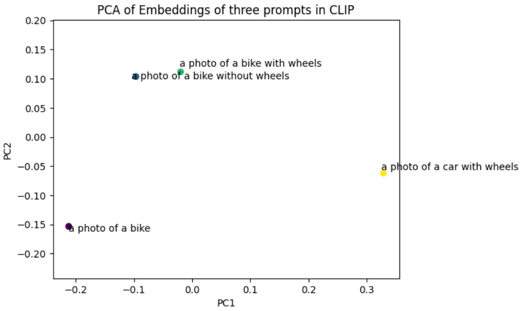


Figure 8: PCA plot of the 3 different prompts with a negation prompt.

bike without wheels have extremely close embeddings. This problem has been introduced into text-to-image generation tasks, making it hard for the model to generate images without certain concepts (examples in Figure 1 of Singh et al. (2025) and Figure 5 of Park et al. (2025)). Thus, classifier-free guidance (CFG) was used to introduce a negative prompt to the image generation process. More details in the next section. Several studies have attempted to tackle this issue by employing alternative training strategies, such as incorporating harder samples in the training data designed for negation tasks (Yuksekgonul et al., 2023; Singh et al., 2024; 2025; Alhamoud et al., 2025; Park et al., 2025). Some of these methods have shown improvements in image generation tasks. For instance, Park et al. (2025) reported gains in Neg Score—measuring whether the model retains the primary subject while correctly omitting the negated object—for both SD-1.4 and SDXL-1.0, by replacing the default CLIP encoder with their NegationCLIP on their dataset, without additional T2I training.

These methods generally require re-training the text encoder (usually a CLIP-like model) with contrastive learning, which poses challenges for models that do not use contrastively pre-trained encoders, such as T5 (Raffel et al., 2023) in Stable Diffusion 3 (Stability AI, 2024; Esser et al., 2024) and Flux (Black Forest Lab, 2025). Moreover, each model using a different text encoder would require a separate, dedicated adaptation. Additionally, even if the text encoder understands the negation, the diffusion model might still fail to avoid certain items because of their strong association.

B Proof That Our Method is the Same as Token-Weighted Subtraction

In this section, we prove that our method Z^{VSF} is equivalent to token-weighted subtraction, denoted Z^{W}.

Proof. We define

$$A^{+} = \exp(\frac{Q(K^{+})^{T}}{\sqrt{d}}), \quad A^{-} = \exp(\frac{Q(K^{-})^{T}}{\sqrt{d}}).$$

Then

$$W = \frac{\sum A^{+}}{\sum A^{+} + \sum A^{-}}.$$

Substituting into the expression for Z^{W}:

$$Z^{W} = W \cdot \sigma(Q(K^{+})^{T})V^{+} - (1 - W) \cdot \alpha \cdot \sigma(Q(K^{-})^{T})V^{-},$$

and using the softmax definitions

$$\sigma(Q(K^{+})^{T}) = \frac{A^{+}}{\sum A^{+}}, \quad \sigma(Q(K^{-})^{T}) = \frac{A^{-}}{\sum A^{-}},$$

we obtain

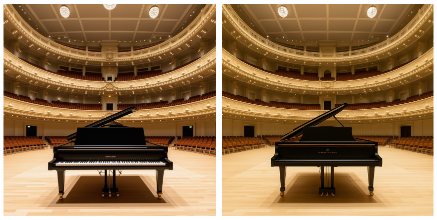


Figure 9: The left image is the original image, and the right image is generated by GPT-4o, where piano keys are removed. When scored using HPSv2, the left image got a score of 0.330 while the right image got a score of 0.319 using prompt of “A grand piano dominates an empty concert hall, a smooth ebony board stretching across the front.” and the left get a score of 0.330 and the right got a score of 0.324 if we mention “no keys” in the promps.

Canceling the sums:

$$Z^W = \frac{A^+}{\sum A^+ + \sum A^-} V^+ - \frac{\alpha A^-}{\sum A^+ + \sum A^-} V^-.$$

This matches

$$Z^{VSF} = \sigma(Q(K^+ \oplus K^-)^T)(V^+ \oplus -\alpha V^-),$$

since

$$\sigma(Q(K^+ \oplus K^-)^T) = \frac{A^+ \oplus A^-}{\sum A^+ + \sum A^-},$$

and therefore

$$Z^{VSF} = \frac{A^+}{\sum A^+ + \sum A^-} V^+ + \frac{A^-}{\sum A^+ + \sum A^-} (-\alpha V^-) = \frac{A^+}{\sum A^+ + \sum A^-} V^+ - \frac{\alpha A^-}{\sum A^+ + \sum A^-} V^-.$$

Hence, $Z^W = Z^{VSF}$. □

C ATTENTION BIAS AND PADDING REMOVAL

We observe that even when the scaling factor $\alpha = 0$, including the negative prompt in the sequence still sometimes reduces image quality. This could be because the negative prompt “distracts” the image tokens’ attention from the image tokens or positive prompts. To mitigate this effect, we introduce a negative bias $-\beta$ into the attention $\mathbf{I} \to \mathbf{N}^{(1)}$, thereby reducing the influence of the negative prompt.

In most models from Huggingface Diffusers (von Platen et al.), padding tokens in the text input are typically not masked during attention. This is likely because the models have learned to ignore padding, and masking them would add unnecessary overhead (due to some attention implementations like FlashAttention-2 (Dao, 2023) that do not support arbitrary masking). However, when we invert the sign of the padding tokens, it degrades output quality significantly. This could be because, although these tokens carry no semantic meaning, the sign-flipping introduces unseen states into the attention mechanism. To mitigate this, we remove padding tokens from the negative prompt embeddings. For the positive prompt, we retain padding tokens, as they do not introduce novel tokens and can improve generation quality. This aligns with training conditions and may allow the model to use padding positions as registers for auxiliary information.

D DETAILS ABOUT THE METRICS

To evaluate the scores of the generated images, we used LLaMA 4 Maverick, which has a very high image reasoning MMMU (Yue et al., 2024) score, higher than Gemma 3 and even GPT-4o. We avoided using the same model (o3) for both evaluation and generation for cost control and to avoid bias within a model. We did not evaluate the quality of the generated images using models like

Table 4: Reliability Metric For Human And MLLM Evaluation Results

	Negative Score	Positive Score
F1	0.667	0.974
Accuracy	0.850	0.950

ImageReward (Xu et al., 2023) or HPSv2 (Wu et al., 2023) as in NASA or NAG, as current quality or human preference assessment models do not account for negative prompts; traditional methods usually aim for real-world generations (Ye et al., 2025). Removing a key element from the positive prompt (e.g., removing the roof from a house) is likely to reduce perceived quality, since the result deviates from what is considered "normal," even though that is the intended outcome. An example is shown in Figure 9 where removing the keys from the piano results in a much lower quality score, even though other parts of the image look the same. Both ImageReward and HPSv2 are built on top of image-text alignment models (CLIP (Radford et al., 2021) or BLIP (Li et al., 2022)), which will likely lead to a decreased score when the main object is missing a critical part. Thus, we also let the MLLM rate the image quality from 0-1 for each image and told it to ignore the abnormality of following the negative prompt. We did observe that MLLM can make mistakes, especially when there is ambiguity or when the unwanted element is hard to see. However, we believe that in general, under 400 images, the mistakes are minor. To compensate for this, we also did a human evaluation and a more negotiation-aware MLLM evaluation to cross-validate the LLaMA evaluation. Due to model provider stability issues, we used different MLLM providers for different portions of the experiment for the same model under the same config; this could have some limited impact on the stability of metrics.

E Hyper-parameter Tunning

Although NAG (Chen et al., 2025a) also targeted negative concept avoidance, its primary focus was on its effects on improving generation quality (using words like "blurry" or "low quality" as a negative prompt). We believe the hyperparameters reported in their work were tuned with an emphasis on quality rather than negation handling. Therefore, we re-tuned their hyperparameters moderately and manually on guidance scale (ϕ), blending factor (α), and normalization factor (τ). We will report experimental results on both original NAG (noted as NAG) and the improved hyperparameter version (noted as NAG Strong). The final hyperparameters used are $\phi = 11, \alpha = 0.5, \tau = 5$ for NAG Strong and $\phi = 4, \alpha = 0.125, \tau = 2.5$ for original NAG. This pushes the NAG to the edge of acceptable visual quality.

Similarly, for our VSF, we used two set of hyperparameters, VSF Quality ($\alpha = 3.3, \beta = 0.2$) maintained high quality and positive prompt alignment, while VSF Strong ($\alpha = 3.8, \beta = 0.2$) pushes it to the limit, reaching higher negative prompt alignment in trade of positive and quality score.

F Human Validation

To verify the results of the MLLM evaluation, we selected 10 prompts (with 2 seeds each) for VSF, NASA, NAG, and NAG Strong and manually labeled them with positive, negative, and quality scores. The human-labeled results are presented in Table 3. We validated MLLM performance by computing the binary F1 score and accuracy between MLLM outputs and human annotations. Cohen's kappa was not applied due to the highly imbalanced class distribution. The reliability metrics are summarized in Table 4. We observed that quality ratings from MLLM and human labels were uncorrelated in high-quality regions. To investigate this further, we evaluated quality scores over a broader set of conditions. With a large sample size, we found that correlation emerges in a wider range: when scores are close to 1, small fluctuations carry little meaning, but substantially lower scores (e.g., < 0.9) may indicate degraded quality. The correlation is shown in Figure 10. This supports the observation in Figure 12, where MLLM tends to overestimate quality. From the scatter plot and regression, we can see that the MLLM score is usually higher than the human score, and although they are not linearly correlated, they are monotonically correlated.

Listing 1: Pseudocode implementation of the Value Sign Flip (VSF) attention process

```python
# prep for embeddings
pos_embeds = get_embed(prompt)
neg_embeds = get_embed(neg_prompt, padding=False)
pos_len, neg_len, img_len = pos_embeds.shape[1], neg_embeds.shape[1], IMG_LEN

# concat positive and negative prompts (N0)
prompt_embeds = torch.cat([pos_embeds, neg_embeds], dim=1)

# prep for attention mask and bias (N1 never acts as query)
total_len = img_len + prompt_embeds.shape[1]
attn_mask = torch.zeros((1, total_len, total_len + neg_len))

# block P and N0 from attending to N1
attn_mask[:,-(pos_len+neg_len):,-neg_len:] = -torch.inf

# block image and P from attending to N0
attn_mask[:,:-neg_len,-(2*neg_len):-neg_len] = -torch.inf

# block N0 and N1 from attending to P
attn_mask[:,-neg_len:,img_len:img_len+pos_len] = -torch.inf

# bias image->N1 connections
attn_mask[:,:img_len,-neg_len:] -= offset


class VSFAttnProcessor(AttnProcessor):
    def __init__(self, attn_mask, neg_prompt_length):
        self.attn_mask = attn_mask
        self.neg_prompt_length = neg_prompt_length

    def forward(self, hidden_states, encoder_hidden_states, attention_mask):
        # get qkv projection for image tokens
        q = self.get_q(hidden_states)
        k = self.get_k(encoder_hidden_states)
        v = self.get_v(encoder_hidden_states)

        # get qkv projection for encoder tokens

        q_enc = self.get_q_encoder(encoder_hidden_states)
        k_enc = self.get_k_encoder(encoder_hidden_states)
        v_enc = self.get_v_encoder(encoder_hidden_states)

        query = torch.cat([q, q_enc], dim=2)

        # append P, N0 (in k_enc and v_enc) and N1 (the last portion of k_enc and v_enc) at the end
        k = torch.cat([k, k_enc, k_enc[:,:,-self.neg_prompt_length:]], dim=2)
        v = torch.cat([v, v_enc, v_enc[:,:,-self.neg_prompt_length:]], dim=2)

        # sign-flip values of N1
        v[:,:,-self.neg_prompt_length:] *= -scale

        hidden_states = F.scaled_dot_product_attention(
            query, k, v,
            dropout_p=0.0, is_causal=False,
            attn_mask=self.attn_mask.to(query.dtype)
        )
        hidden_states = hidden_states.transpose(1, 2).reshape(batch_size, -1, attn.heads * head_dim)
        return self.out_proj(hidden_states)


for block in model.transformer.blocks:
    block.attn1.processor = VSFAttnProcessor(attn_mask, neg_len)

# diffusion process continues
```

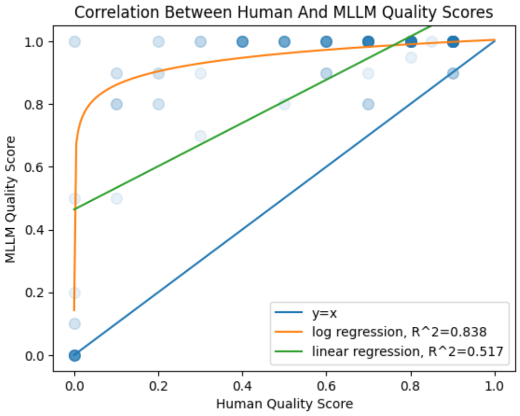


Figure 10: Correlation Between the Human-rated quality score and MLLM-rated quality score

Table 5: Comparsion of Flux Schnell VSF and Original Schnell (Black Forest Lab, 2025)

Method	Positive Score	Negative Score	Quality Score
Flux Schnell	1.00	0.22	0.99
Flux Schnell VSF	0.97	0.41	0.99

G Adapting to Other DiT Models

In this paper, we primarily use SD-3.5 (Stability AI, 2024) due to simplicity and elegant architecture. However, our method can theoretically be adapted to any transformer-based diffusion or flow-matching model. To demonstrate this adaptability, we implemented our method on Wan 2.1 with CausVid LoRA (Yin et al., 2025; Seppanen) and Flux Schnell (Black Forest Lab, 2025).

For Wan 2.1, which uses cross-attention between image and text, duplication and masking are unnecessary and not used. Because our approach does not perform extrapolation and solely provides negative guidance, it cannot enhance overall quality significantly or replace CFG sampling in non-disstilled models, making it incompatible with the original Wan 2.1 model. Instead, we utilize CausVid (Yin et al., 2025), which enables Wan to function effectively without classifier-free guidance in few-step settings. Specifically, we used a LoRA distilled from the original CausVid that can be directly applied on top of Wan 2.1 (Seppanen). For qualitative results from Wan, please see the appendix.

We also tested our method on Flux Schnell (Black Forest Lab, 2025). However, due to the model likely being trained to associate items with their associated items that often appear together strongly, we need to make some modifications. Before the negative prompt was fed into the model, we did a CFG-like extrapolation on the negative prompt, with a mean padding embedding as a null condition. This follows the implementation of the Compel package. Noted as:

$$p^- = p^- + \lambda \cdot (p^- - p^\emptyset), \tag{10}$$

we used $\lambda = 8$ in this case. Quantitative results are shown in Table 5. We can observe that even without any negative guidance, the Flux Schnell model can slightly better avoid the items solely based on the positive prompts (since the positive prompts implied the item is missing using terms like “empty”), but with the help of VSF, it further increases the negative score without compromising the positive and quality score.

H Computational Cost

Since our method does not require two passes through the entire model (as in CFG) or the attention module (as in NAG or NASA), and only slightly increases the sequence length ($< 0.2\%$), its theoretical computational cost is significantly lower, close to that of a single pass. However, due to implementation limitations (specifically, FlashAttention-2’s lack of support for arbitrary attention

Table 6: The computation cost of each model. Time is measured in total runtime per sample, and VRAM is the peak RAM during the 25 samples generation. Since VSF Wan does not require a mask, and it is only used for bias, we also tested it without the bias. The SD3.5 model used is SD-3.5-Large-Turb,o and the Wan model used is Wan-2.1-T2V-1.3B.

	Wan		SD3.5	
	Time	VRAM	Time	VRAM
Baseline	23.10s	22.05GB	2.14s	28.49GB
NASA	-	-	2.89s	28.50GB
NAG	25.58s	22.06GB	2.98s	28.50GB
CFG (Theroatical)	46.20s	-	4.28s	-
VSF	22.70s	23.05GB	3.00s	28.53GB
VSF (No mask/bias)	22.70s	22.05GB	-	-

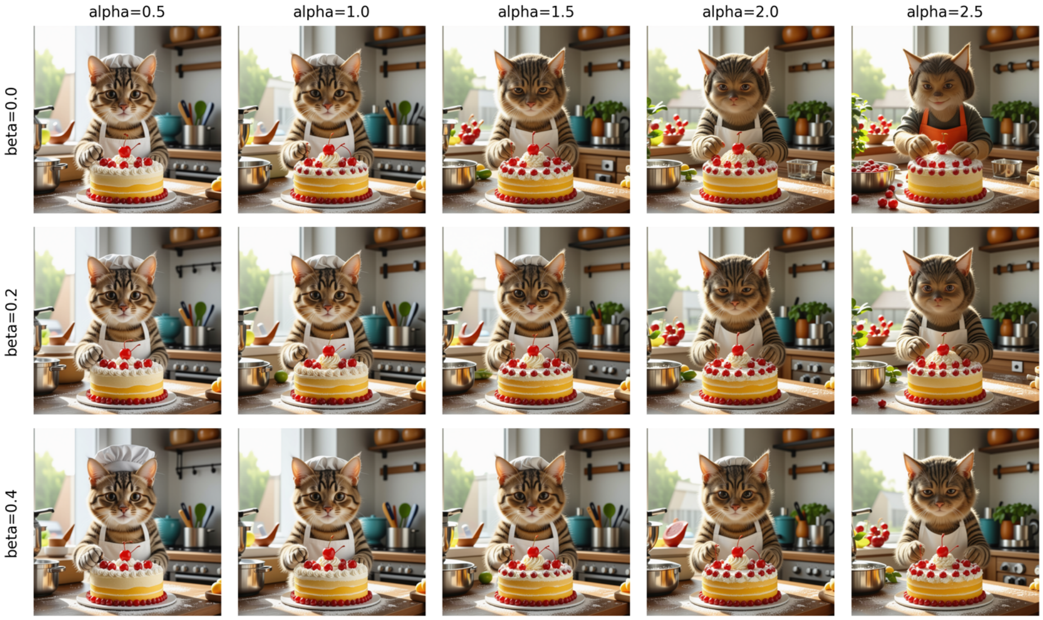


Figure 11: Effects of guidance scale (α) and attention bias (β) in image generation. Positive prompt is "a cat making a cake in the kitchen, the cat is wearing a chef's apron..." and negative prompt is "chef hat."

masking), the actual runtime of our method is higher than the original single-pass MM-DiT models, and similar to NAG or NASA, but still lower than CFG.

To accurately measure the computational cost, we evaluate the runtime of 25 identical prompts under four settings: no guidance, NAG, NASA, and our proposed guidance, VSF, and then report the average runtime and peak memory usage for each setting. We also reported the theoretical CFG time as double the one without guidance. To avoid GPU thermal throttling affecting the results, we pause for at least 5 minutes between each set of tests. The tests are done on NVIDIA A100 40GB on Google Colab, as this is the most accessible option for high-end GPUs for users. Stable-Diffusion-3.5-Large-Turbo is generated in 8 steps for 1024x1024 resolution, Wan is generated in 8 steps with 480x832 resolution, and 81 frames. The results are shown in Table 6.

From the table, VSF requires marginally more time and memory than NAG in SD3.5, while they are both significantly faster than theoretical CFG time, which would be twice the baseline. In Wan, VSF outperforms NAG and is even slightly better than the baseline (likely due to nature variation or noise) in terms of compute time, though it consumes 1GB more memory, likely due to the attention bias being stored. Since this bias is optional, we tested VSF Wan's performance with it removed, which results in an improvement in VRAM usage such that it uses the same amount of VRAM as baseline and NAG, and no change in runtime.

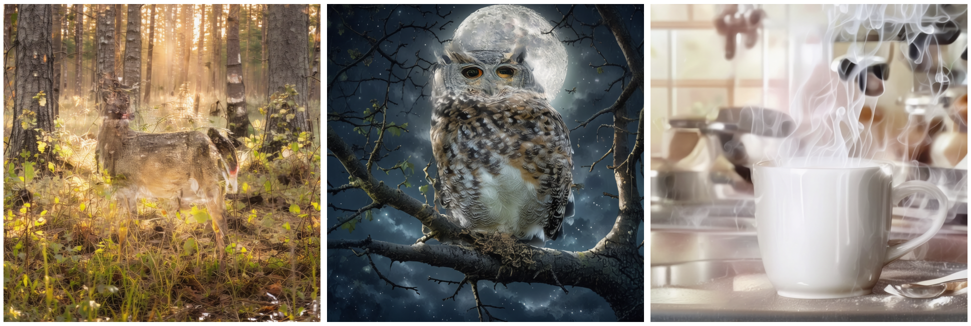


Figure 12: An example of a completely distorted image gets a relatively high quality score. The left one has a score of 70, the middle one has a score of 90, and the right one is a slightly distorted image, but still rated for 100.

I External Baselines

In addition to other guidance methods applied to SD-3.5-turbo, we evaluated several external baselines. The first baseline employs a generate-then-edit approach, loosely inspired by Generate-Plan-Edit (GraPE) (Goswami et al., 2025), but omitting the planning stage as our goal is straightforward (removing unwanted elements). Specifically, we first generated images using SD-3.5-Large-Turbo without a negative prompt, and subsequently edited out the unwanted elements automatically using Flux Knoest (Labs et al., 2025), an image editing model, using prompt `Remove [negative prompt]`.

The second baseline utilizes GPT-4o's native image generation capability. GPT-4o has demonstrated strong prompt-following performance(Wei et al., 2025), including in handling negation tasks. As GPT-4o lacks explicit negative prompt functionality, we formatted prompts as `[Positive prompt], but with no [negative prompt]`. Since our focus is on evaluating negation rather than image quality, we adopted the "low" generation setting. Besides GPT-4o, we also added the newly released Nano Banana from Google. It is also a language model-based image generation model and has received a good reputation in the image generation community.

The third baseline we included is Janus-4o (Chen et al., 2025b), a model distilled from GPT-4o onto the Janus-Pro base architecture (Chen et al., 2025c). Given GPT-4o's strong prompt-following performance, we anticipated competitive results from Janus-4o. We provided negative prompts directly as negations within the positive prompts, same as GPT-4o.

Finally, we tested Qwen-Image (Wu et al., 2025) using two configurations: one employing separate positive-negative prompt pairs using CFG (labeled as Qwen-Image NP), and another embedding negative prompts as negations within the positive prompt itself (labeled as Qwen-Image Negation), while still using CFG with an empty negative prompt. Qwen is run under DFloat-11.

All measure time is measured on Google Colab 40GB A100 GPU, and for Qwen-Image and Generate+Edit, model CPU offloading is enabled.

The results are presented in Table 1 in the main text. The table indicates that VSF Strong achieves the second-best negative score, only behind GPT-4o, while also demonstrating a significantly faster runtime compared to all other methods, outperforming even the generate-then-edit pipeline. The GPT-4o distilled model, Janus-4o, has an unexpectedly low negative score, which could be because they did not have enough negation-included prompts in the distillation data. The VSF Quality had a lower negative score compared to Generate+Edit, while having a much higher positive and quality score, and shorter runtime.

J SD3.5-Large-Turbo Qualitative Results

Selected qualitative results are shown in Figure 13. The positive prompt is condensed for spacing. For the glasses without lens images, both NAG and NAG Strong generated classes clearly have a lens. For the VSF-generated image, we can see the lens is missing, even though the frames are

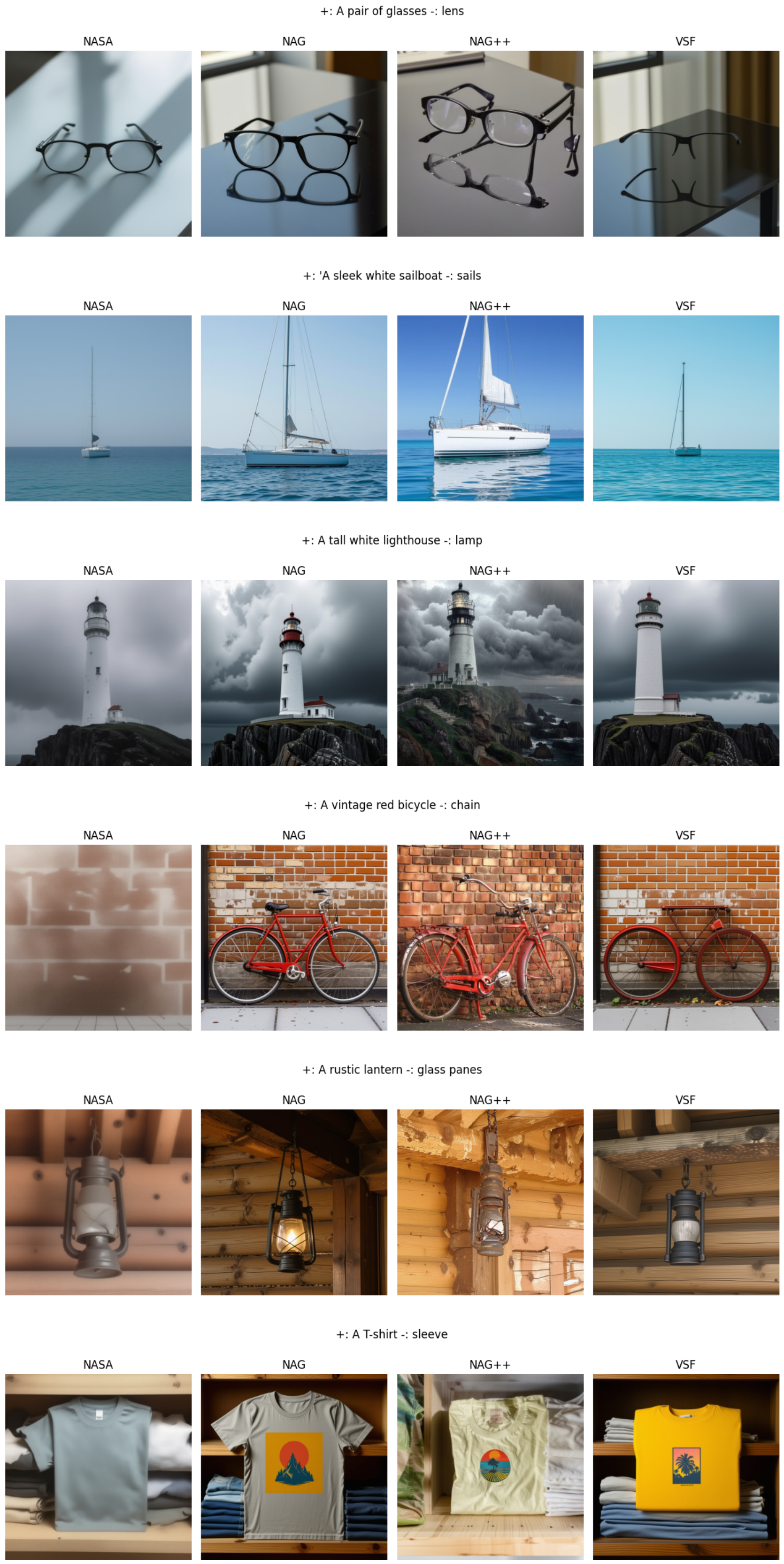

+: A pair of glasses -: lens
NASA
NAG
NAG++
VSF
+: 'A sleek white sailboat -: sails
NASA
NAG
NAG++
VSF
+: A tall white lighthouse -: lamp
NASA
NAG
NAG++
VSF
+: A vintage red bicycle -: chain
NASA
NAG
NAG++
VSF
+: A rustic lantern -: glass panes
NASA
NAG
NAG++
VSF
+: A T-shirt -: sleeve
NASA
NAG
NAG++
VSF

Figure 13: Selected Results for Comparison. Positive promtes are condensed for spacing.

floating. However, this issue was also presented in NAG Strong's image, even though it still has glasses. For a sailboat without sails, all other methods generated smaller but still existing sails, while VSF successfully avoided sails. In the third image of a lighthouse without a lamp, both VSF and NASA have no visible lamp, while the images from NAG and NAG Strong have a clear lit lamp. In the image of a bicycle without a chain, NASA generated a blurry image without bikes at all, while NAG generated a normal image, and NAG Strong generated a slightly distorted image of a bike with no seat yet the chain is still present. VSF successfully generated a bike without a chain, even though it also removed the seat. For the prompt of a lantern with no glass panes, NASA generated a lamp with frosted panes, NASA++ generated a classic glass pane, and NASA++ generated broken frosted panes. VSF, in this case, generated a pane that is clearly not glass. In the last example of a T-shirt, NASA generated a blurry image with still one sleeve visible, and NAG generated the T-shirt with both sleeves visible. NAG Strong and VSF both avoided the sleeve, even though NAG Strong has some artifacts.

K QUALITATIVE RESULTS FOR WAN

In Figure 14, we showed 3 examples generated from Wan-2.1-14B. In the first example, we successfully removed the stars in the background while keeping other elements intact. On the right side, in the absence of stars, the moon lander is generated to fill the space. In the second video, we successfully generated a windowsill without a curtain. In the last video, the generated video from VSF contains no trees on the left, and instead, it fills it with a hill. There are still some bushes on the right side, which do not violate the negative prompt of "trees. All videos have the same high quality as the original one.

L FAILURE CASES

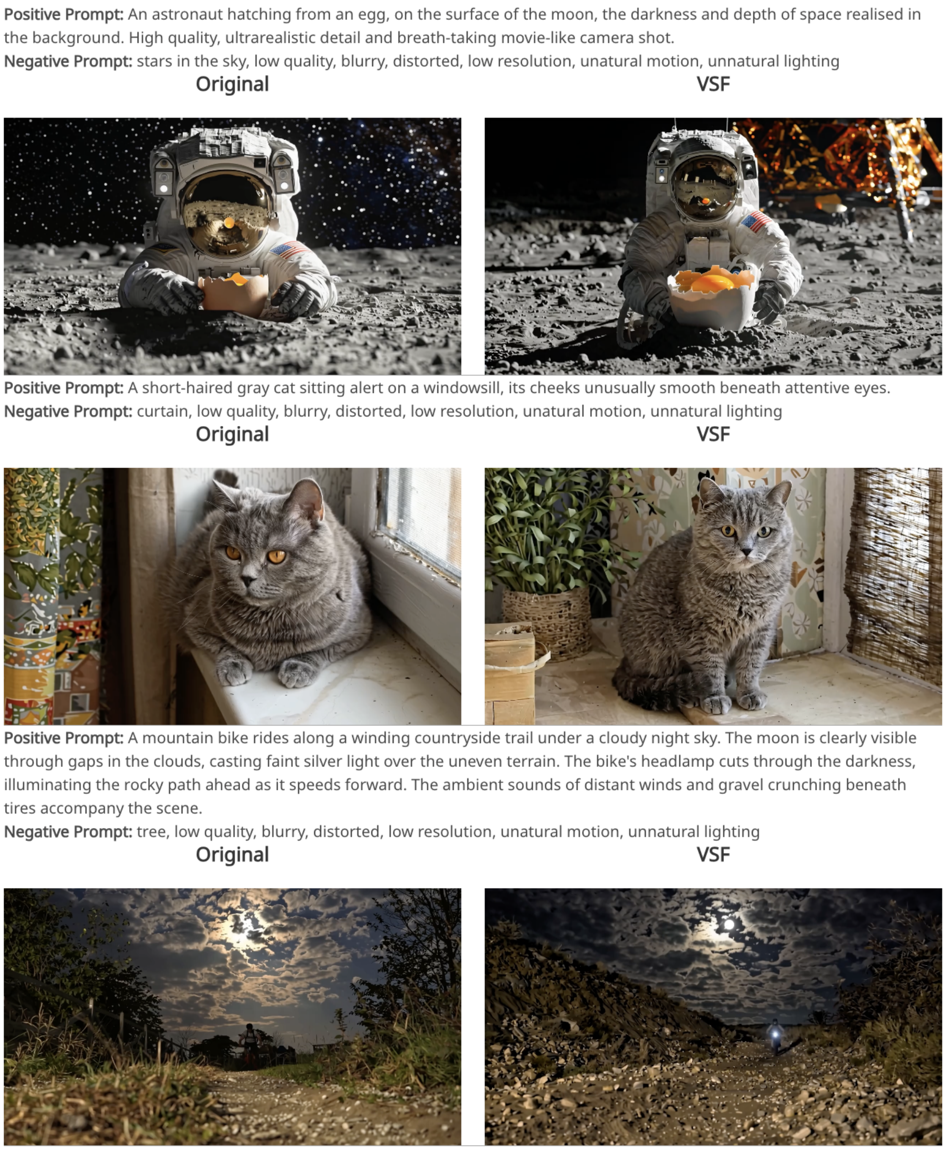
Like any method, our method is not perfect, especially in a challenge dataset like NegGenBench. In Figure 15, we showed 4 failed cases. In the case where we want to generate a keyboard without a spacebar, the generated object is technically a key-board (an array of keys) and has no space bar, but it is not what people imagine when they think about "keyboard with no spacebar." The second failed case is another image from glasses with no lens; the generated image has no lens, but the frame is twisted in an unnatural way. In the third example, where a house with no roof is needed, VSF completely missed the negative prompt, possibly due to the strong association between roof and house. In the last example of a cat with no whiskers, the generated image technically has no whiskers and looks like a cat, but it looks more like a cat statue instead of a living cat.

M NON-OBJECT NEGATIVE GUIDANCE

In this paper, we focused on removing a critical component in the image. To further validate our negative guidance method in other areas, we also tested it on style avoidance. In Figure 5 (in main text), we show four examples, each of which is generated using the same seed. We can see that when prompted with famous artwork (e.g.,"A painting of Starry Night from the 1890s" or "Mona Lisa oil painting") but with a negative prompt of the artist's name style, the generated image avoided any elements related to the style (including the town in the Starry Night) but kept the semantic meaning of the positive prompt. When prompted to give an old photo but not monochrome, the generated image is more like an old-style color photo, follows both non-monochrome and also not very bright (as old photos, even in color, are less vibrant). We find these examples interesting and think they can be used for machine unlearning, using a similar method as in (Gandikota et al., 2023).

N AN EXPERIMENT ON ANTI-AESTHETICS ARTS

Current image generation models are typically finetuned to align with so-called "human preference." However, we argue that there is no universal standard for human preference, and it cannot be defined solely by developers, who inevitably bring their own interests and assumptions. Aligning models exclusively with such values risks introducing bias and potentially marginalizing minority perspectives and interests (Arzberger et al., 2024; Turchin, 2019; Sutrop, 2020; Guo et al., 2025).



Positive Prompt: An astronaut hatching from an egg, on the surface of the moon, the darkness and depth of space realised in the background. High quality, ultrarealistic detail and breath-taking movie-like camera shot.
Negative Prompt: stars in the sky, low quality, blurry, distorted, low resolution, unatural motion, unnatural lighting

Positive Prompt: A short-haired gray cat sitting alert on a windowsill, its cheeks unusually smooth beneath attentive eyes.
Negative Prompt: curtain, low quality, blurry, distorted, low resolution, unatural motion, unnatural lighting

Positive Prompt: A mountain bike rides along a winding countryside trail under a cloudy night sky. The moon is clearly visible through gaps in the clouds, casting faint silver light over the uneven terrain. The bike's headlamp cuts through the darkness, illuminating the rocky path ahead as it speeds forward. The ambient sounds of distant winds and gravel crunching beneath tires accompany the scene.
Negative Prompt: tree, low quality, blurry, distorted, low resolution, unatural motion, unnatural lighting

Figure 14: Qualitative Results for Wan

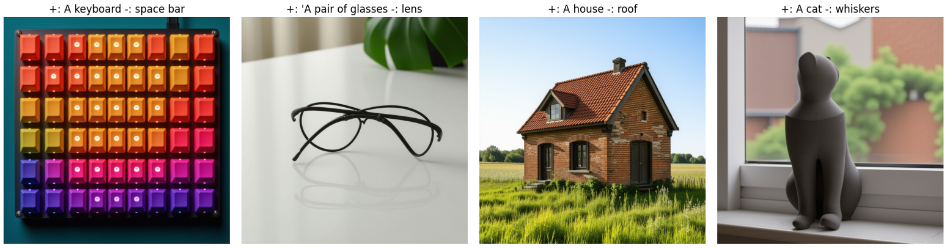


Figure 15: Failed examples, positive prompts are condensed for spacing.

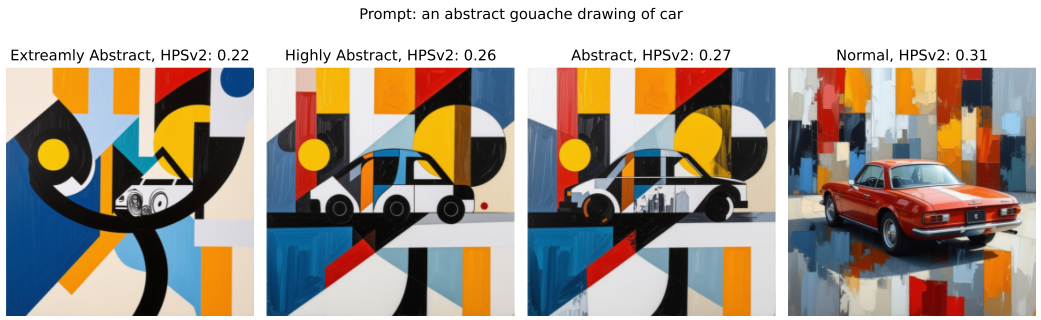

Prompt: an abstract gouache drawing of car
Extreamly Abstract, HPSv2: 0.22
Highly Abstract, HPSv2: 0.26
Abstract, HPSv2: 0.27
Normal, HPSv2: 0.31

Figure 16: Image with abstract style receives a lower score in HPSv2 (reflecting human preference; traditionally, models aim for higher scores).

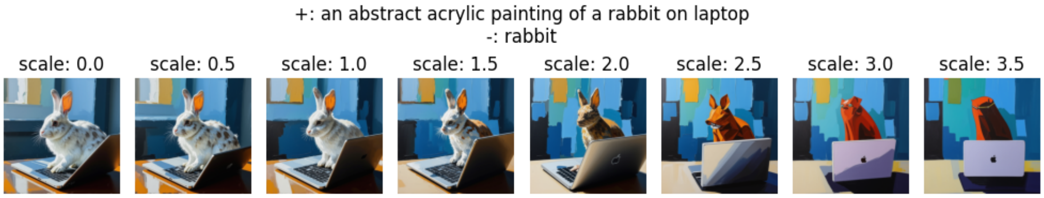
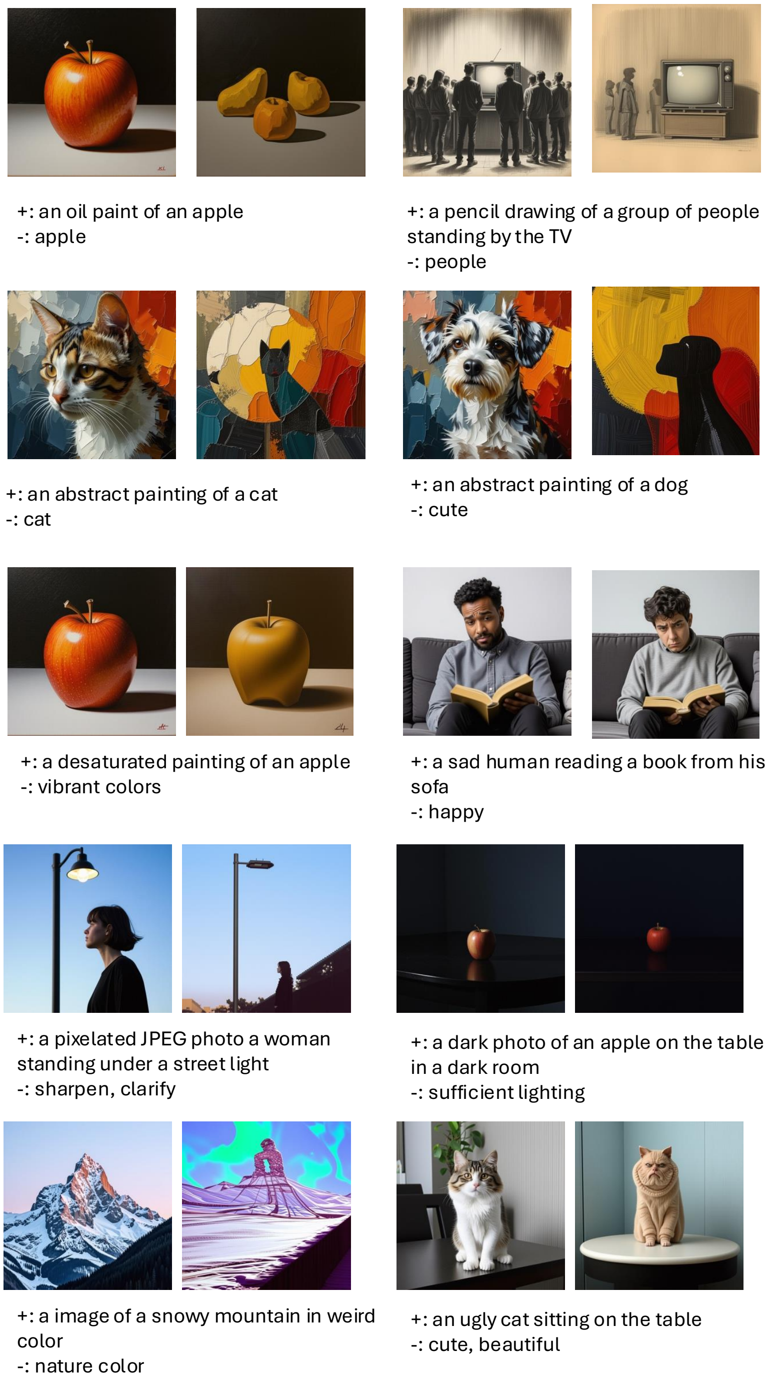


Figure 17: Abstraction of the image as scale increases.

In the context of image generation, this alignment may lead to homogenization of style or taste, producing only broadly pleasing outputs for the general population. Such uniformity can suppress niche demands for degraded, low-quality, or unconventional aesthetics. To counteract this, one possible approach is the use of negative guidance to steer outputs away from mainstream preferences. In this experiment, we tested how VSF can address this issue. We ran our VSF in settings where $\alpha = 0$, which shows on the left, and $\alpha \in [0, 4]$, which shows on the right side. The image with $\alpha = 0$ might not be the same as the one without guidance, but should be an image without negative guidance. The first test used prompts containing the same object in both positive and negative form, with the goal of producing abstract art. This works by semi-canceling the main object, making it appear in an abstract form. Abstract styles are often disadvantaged in alignment settings, since reward models typically favor realistic or figurative outputs. VisionReward (Xu et al., 2025) encodes this bias through its scoring metric, and LAPIS (Maerten et al., 2025) reports that abstract paintings generally receive lower preference scores. Figure 16 shows that an abstract image gets a much lower score compared with a figurative one. As shown in the first two rows of Figure 18, the apple, people, and cat appear in abstract form, demonstrating a clear shift away from the default figurative tendency of the aligned models when VSF is applied. For the last image of a dog, we used "cute" as a negative prompt, which usually describes realistic objects, and we achieved a very abstract and artistic image. Figure 17 shows how abstract the image gets as the scale increases. More examples are shown in Figure 6.

In the second test, the goal was to diverge from styles that are generally appreciated. The positive prompt specified the desired style, while the negative prompt contained descriptions of commonly preferred styles. Importantly, the positive prompt clearly described the intended output, so a faithful model should follow it rather than default to generalized human preference. The tested cases included desaturated color, sad emotion, pixelated art, insufficient lighting, unnatural colors, and a non-beautiful cat. Results show that the baseline model struggled to maintain these characteristics, often reverting to conventionally "beautiful" imagery, whereas VSF successfully produced outputs aligned with the specified unconventional styles.



+: an oil paint of an apple
-: apple

+: a pencil drawing of a group of people standing by the TV
-: people

+: an abstract painting of a cat
-: cat

+: an abstract painting of a dog
-: cute

+: a desaturated painting of an apple
-: vibrant colors

+: a sad human reading a book from his sofa
-: happy

+: a pixelated JPEG photo a woman standing under a street light
-: sharpen, clarify

+: a dark photo of an apple on the table in a dark room
-: sufficient lighting

+: a image of a snowy mountain in weird color
-: nature color

+: an ugly cat sitting on the table
-: cute, beautiful

Figure 18: A test for anti-aesthetics. The left image is generated with $\alpha = 0$, and the right image is generated with $\alpha > 0$. These tests aim to move away from universally pleasing styles and demonstrate the ability to capture more diverse aesthetic preferences.

Table 7: Negation-Aware LLM Evaluation on Testing Set

	Parameters	r (↑)	Acc (↑)	F1 (↑)
Llama Maverick	400B	0.05	0.83	0.59
Llama Maverick CoT	400B	0.03	0.77	0.50
Qwen-2.5-VL 32B	32B	0.28	0.80	0.36
Qwen-2.5-VL 32B CoT	32B	0.31	0.88	0.70
NegAwareQwen-7B	7B	**0.37**	0.86	0.65
NegAwareQwen-32B	32B	0.34	**0.90**	**0.76**

O NEGATION-AWARE MLLM

Upon visual inspection, we observed that GPT-4o effectively avoids many negative prompts, though occasionally the ambiguity within negative prompts (e.g., the term “door” referring either to the door panel or the entrance) and the vision ability of MLLM itself (i.e., hallucination) leads to lower negative scores. It is possible that the negative scores for other methods might also be underestimated. We acknowledge this as a limitation associated with using an MLLM as the evaluator. Thus, we provided a better negotiation understanding of MLLM and used this MLLM to evaluate different guidance methods.

To enable future research on evaluating negation prompting in generative models, a reliable and fast (LLaMA Maverick is too big) evaluation model is necessary. Previous CLIP-based studies concentrated on simple negations (e.g., ”a cat that is not on the grass”) rather than more complex cases such as those in NegGenBench. To address this, we finetuned a multimodal large language model, Qwen-2.5-VL-7B, called NegAwareQwen for improved understanding.

We created 100 additional prompts using GPT-5, each paired with a positive and a negative pair and 2 questions. For each prompt, we generated two images with the three models (NAG, VSF Quality, and NASA) and selected 722 for manual scoring on two dimensions: adherence to the negative prompt and overall quality. We did not assess the positive prompt evaluation as those are simpler, and almost all images in the dataset have a perfect positive prompt following. Negative adherence was rated on a three-level scale: 0 (ignored), 0.5 (partial), and 1 (fully followed). When generating the samples, we slightly randomly adjust the hyperparameter in a small range to create more diverse data (For VSF, $\alpha = 3.3 \pm 1, \beta = 0.2$; for NASA, $\alpha = 0.15 \pm 0.05$; for NAG, $\phi = 8 \pm 4, \alpha = 0.5 \pm 0.2, \tau = 4 \pm 2$). Since this makes image generation models generate sub-optimal images, the rating results of each model’s images are not used for direct comparison. The dataset and the model will be opened after publication.

Note that here Llama showed a very weak r-score for quality; this is because all the images are evaluated using relatively high-quality images (unlike in the ablation study, where many images are lower quality). We did not compare the 7B untrained model because it often failed to output the structure data needed.

The model was finetuned using prompts from the dataset of all 3 models. We trained the model using QLoRA (Dettmers et al., 2023) with rank of 8 and $r = 8$, applied to query, key, value projections in both the vision encoder and language model with dropout of 0.1. Model is trained using $lr = 5 \times 10^{-5}$ (with warm up and decay), WeightDecay=0.1, BatchSize=16, Epoch=5. The dataset is split into train-val with a 90-10 ratio based on the prompt level splitting and aiming for balanced scores in each split. We treat 0.5 as False and calculate negative scores as a binary metric. Results are presented in Table 7 with comparison with the same model without finetuning and LLaMA-Maverick.

P EVALUATION USING NEGAWAREQWEN

We re-evaluated VSF Quality, NAG, NAG Strong, NASA, GPT-4o, and Nano Banana images generated using the prompts and seeds in the main text of the paper using our finetuned NegAwareQwen; the results are shown in Table 8. We did not round the 0.5 score. We used a positive score from the original Table 2 as our finetuned version was trained on largely positive compliance samples. The results match our observation with human validation and MLLM evaluation, that our method gets

Table 8: Evaluation of Different Methods using Our NegAwareQwen-32B

	Positive Score (↑)	Negative Score (↑)	Quality Score (↑)
VSF Quality	0.980	0.330	**0.814**
VSF Strong	0.870	**0.415**	**0.814**
NASA	0.950	0.224	0.727
NAG Strong	0.950	0.168	0.795
NAG	**1.000**	0.147	0.812
GPT-4o	**0.978**	**0.619**	0.812
Nano Banana	0.985	0.406	**0.817**

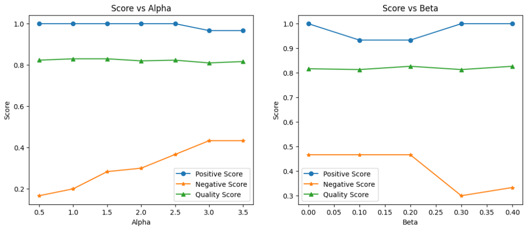


Figure 19: Relationships between metrics and α and β

the highest negative score while having better or comparable positive and quality scores with other open source guidance methods.

Q Ablation Study On α and β

To study the effects of α and β and hyperparameter sensitivity, we studied the effects of the two hyperparameters. We used 30 randomly selected prompts from the dataset and tested the effects of α and β on the positive, negative, and quality scores. When testing the effects of α, we set β to 0, and when studying the effects of β, we set α to 3.5. All generations use the same seed. The images are evaluated using our NegAwareQwen-32B running for negative and quality score and used LLaMA for the positive score. Qualitative results of the effects of α and β are shown in Figure 11 and quantitative results are shown in Figure 19. We can see that as α increases, negative scores increase while positive scores decrease. When β increases, the negative score decreases while the positive score increases, which could be noise. In both cases, the quality scores only change slightly.

R Future Work

Future work may involve applying it to non-diffusion models (like Janus-4o (Chen et al., 2025b)) or models with complex text encoders (like Qwen-Image (Wu et al., 2025)), improving robustness through normalization and blending techniques similar to those employed by NAG, and optimizing computational efficiency by using a better attention implementation. Additionally, we observed some inaccuracies in MLLM judgment due to ambiguities or minimal differences in visual differences. Conducting a larger-scale human evaluation study would help mitigate inaccuracies observed in MLLM-based assessments. Investigating the attention maps and diffusion trajectories of our model could further elucidate the underlying mechanisms of VSF. Decoupling the attention, such that it calculates the positive and negative attention separately and then uses the ratio to extrapolate the output, might yield better quality in exchange for runtime. Or, adding a scaling factor to the positive prompt for better control.

S Use of LLMs in The Paper

In this work, we used LLM for paper-related work collection and consumption. It is also used to polish the paper language or provide feedback for writing/figures. It is also used to brainstorm before and during the project.