Dynamic VLM-Guided Negative Prompting for Diffusion Models

Hoyeon Chang*, **Seungjin Kim***, **Yoonseok Choi***
KAIST
retapurayo, sjkim, uooh77@kaist.ac.kr

Abstract

We propose a novel approach for dynamic negative prompting in diffusion models that leverages Vision-Language Models (VLMs) to adaptively generate negative prompts during the denoising process. Unlike traditional Negative Prompting methods that use fixed negative prompts, our method generates intermediate image predictions at specific denoising steps and queries a VLM to produce contextually appropriate negative prompts. We evaluate our approach on various benchmark datasets and demonstrate the trade-offs between negative guidance strength and text-image alignment.

1 Introduction

A prevalent method for content filtering in T2I is negative prompting, which guides the model away from specified concepts through Classifier-Free Guidance (CFG) [Ho, 2022]. However, this technique suffers from key limitations. First, it can disrupt the image generation process even when the unwanted content is not present, leading to over-correction and semantic drift [Ban et al., 2024, Chang et al., 2024, Koulischer et al., 2024]. Second, negative prompts are typically predefined, yet it is difficult to anticipate all potential unwanted elements that might arise from a given positive prompt [Yoon et al., 2024]. This can lead to either unnecessary or inefficient filtering.

In this work, we propose a novel solution: Vision-Language guided Dynamic Negative Prompting (VL-DNP). By leveraging the advanced image and language understanding of modern open-source Vision-Language Models (VLMs), our method acts as a dynamic negative prompt generator. VL-DNP detects the emergence of unwanted content during the denoising process and generates targeted negative prompts in real-time to erase it. Importantly, our framework can be easily integrated into any pretrained CFG-based diffusion model without requiring joint training or model modifications.

2 Related Work

Negative prompting and Negative Guidance Negative prompting was proposed to filter out unwanted content. In this work, we write positive conditions we like to align as $c+$, and negative conditions we want to filter as $c-$.

$$s_\theta(x_t, t, c+, c-) = \nabla_{x_t} \log p(x_t|c+) + \omega \left(\nabla_{x_t} \log p(x_t|c+) - \nabla_{x_t} \log p(x_t|c-)\right). \quad (1)$$

In Eq. (1), conventional negative prompting is considered a particular example that blends positive and negative guidance. The corresponding score function augmented with both positive and negative guidance is defined as follows:

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

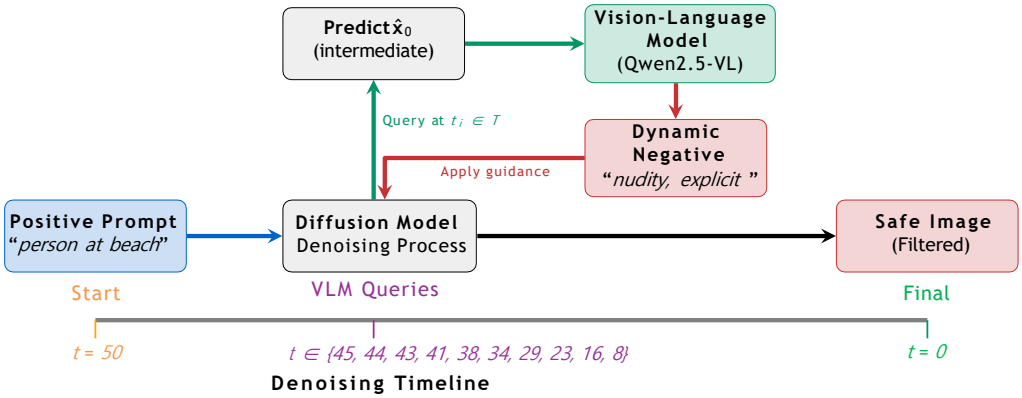

Figure 1: **VL-DNP inference pipeline.** The positive text prompt is fed to a pretrained diffusion model. At a small set of timesteps $t_i \in \mathcal{T}$ we predict the clean image $\hat{x}_0$, query a lightweight vision–language model (VLM) and obtain a *dynamic negative prompt*. The prompt is fed back as classifier-free guidance, steering the remaining denoising steps away from any unsafe content detected in the intermediate image.

$$s_{\theta,\text{mixed}}(x_t, t, c+, c-) = \nabla_{x_t} \log p(x_t) + \omega_{\text{pos}} \Big(\nabla_{x_t} \log p(x_t|c+) - \nabla_{x_t} \log p(x_t) \Big) - \omega_{\text{neg}} \Big(\nabla_{x_t} \log p(x_t|c-) - \nabla_{x_t} \log p(x_t) \Big). \tag{2}$$

Using Eq. (2) for sampling implies sampling from $\tilde{p}(x_0) \propto p(x_0) \frac{p(c+|x_0)^{\omega_{\text{pos}}}}{p(c-|x_0)^{\omega_{\text{neg}}}}$.

The importance of negative guidance at each intermediate generation step might differ. Building upon this intuition, recent work [Koulischer et al., 2024] attempts to determine the weight of the negative guidance scale dynamically based on the estimated $p_t(c-|x_t)$. However, their approach still relies on a pre-defined negative prompt specified prior to inference.

Vision-Language Models in Generation VLMs have demonstrated remarkable capabilities in understanding visual content and generating descriptive text [Radford et al., 2021, Li et al., 2022, 2023, Liu et al., 2023]. In this work, we harness the image recognition and language generation capabilities of VLMs to provide dynamically adapting negative prompts during diffusion model inference.

3 Methodology

The framework operates through the following steps: (1) Perform standard CFG denoising steps using the positive prompt. (2) At predefined timesteps, predict the denoised image $\hat{x}_0$ from current latent x_t. (3) Query a VLM to generate negative prompts based on the predicted image $\hat{x}_0$. (4) Apply negative guidance using the VLM-generated prompt for subsequent denoising steps.

Mathematical Formulation Based on the negative prompting formulation in Eq. (2), we introduce temporal adaptivity to the negative conditioning. Let Θ denote a Vision-Language Model, and $\mathcal{T} = t_1, t_2, \ldots, t_k \subseteq [0, T]$ be a predefined set of timesteps where VLM queries occur. At each query timestep $t_i \in \mathcal{T}$, we predict the denoised image denoted by $\hat{x}_0^{(i)}$ using the current estimate as follows:

$$\hat{x}_0^{(i)} = \frac{x_{t_i} + (1 - \bar{\alpha}_{t_i}) \cdot s_{\theta,\text{cfg}}(x_{t_i}, t_i, c+)}{\sqrt{\bar{\alpha}_{t_i}}}, \tag{3}$$

where $\bar{\alpha}_{t_i} = \prod_{s=1}^{t_i}(1 - \beta_s)$, and β_s is the forward process variance at timestep s [Ho et al., 2020]. Classifier-free guidance (CFG) is defined as:

$$s_{\theta,\text{cfg}}(x_t, t, c+) \;=\; \nabla_{x_t} \log p(x_t) \;+\; \omega \left(\nabla_{x_t} \log p(x_t|c+) \;-\; \nabla_{x_t} \log p(x_t)\right).$$

The VLM then generates an adaptive negative prompt:

$$c-_{t_i} = \Theta(\hat{x}_0^{(i)}, \mathcal{D}), \tag{4}$$

where $\mathcal{D}$ represents optional few-shot demonstration examples provided to the VLM. Although $\hat{x}_0^{(i)}$ is generally blurred during the initial denoising stage, we find that the VLM can detect objects even at early stages.

VLM Integration The VLM receives two inputs: (1) the intermediate image prediction $\hat{x}_0^{(i)}$, (2) demonstration examples that guide the model toward generating appropriate negative content descriptors. Our prompting strategy instructs the VLM to identify potentially inappropriate or unwanted visual elements in the predicted image and generate concise negative prompts to suppress such content.

Dynamic Negative Guidance For timesteps $t_i < t < \min(t_{i+1}, T)$, we apply the augmented score function:

$$\begin{aligned}\tilde{s}_\theta(x_t, t, c+, c-_{t_i}) = \nabla_{x_t} \log p(x_t|c+) + \omega_{\text{pos}}\Big(\nabla_{x_t} \log p(x_t|c+) - \nabla_{x_t} \log p(x_t)\Big) \\ - \omega_{\text{neg}}\Big(\nabla_{x_t} \log p(x_t|c-_{t_i}) - \nabla_{x_t} \log p(x_t)\Big)\end{aligned} \tag{5}$$

The key advantage of this formulation is that $c-_{t_i}$ adapts to the evolving image content, allowing for more precise and contextual content filtering compared to static negative prompting approaches.

4 Experimental Setup

Datasets To evaluate the general performance of image generation, we use 100 prompts randomly sampled from COCO-30K [Lin et al., 2014]. We compute a CLIP and FID score for the COCO-100 prompt set. To evaluate the filtering of unsafe images, we use Ring-a-Bell-16 [Tsai et al., 2023], P4D [Chin et al., 2024], and Unlean-diff [Zhang et al., 2024] datasets, which consist of adversarial prompts designed to test content filtering in T2I tasks.

Implementation Details We use Stable Diffusion v1.4 [Rombach et al., 2022] as the base diffusion model and Qwen2.5-VL-7B-Instruct [Bai et al., 2025] as the Vision-Language Model for dynamic negative prompt generation. The denoising process employs DPM-Solver++ [Lu et al., 2022] with 50 inference steps, where VLM queries are performed at timesteps $45, 44, 43, 41, 38, 34, 29, 23, 16, 8$ out of the total 50 steps. We evaluate our approach using CLIP-based metrics for text-image alignment, complemented by NudeNet [notAI tech, 2019] Attack Success Rate and Toxic Rate measurements to assess content filtering effectiveness.

Baseline Comparisons We compare our approach against baseline configurations: fixed negative prompting with various negative guidance scales, token embedding projection method SAFREE [Yoon et al., 2024] and our dynamic VLM-guided approach across different negative guidance scale settings.

5 Results

Table 1 collects the headline numbers. We report Attack-Success Rate (ASR) and Toxic Rate (TR) to quantify safety, and CLIP score together with FID to quantify fidelity.

In every dataset, raising the negative-guidance scale lowers ASR/TR but tends to raise FID. With static negative prompting, increasing the scale from $\omega_{\text{neg}} = 7.5$ to 20 drags CLIP from 0.313 to 0.277 and drives FID from 107 to 153. The same sweep under our VLM-guided schedule leaves CLIP essentially

Table 1: Safety (**ASR**, **TR**) vs. alignment/quality (**CLIP**, **FID**↓). "VL–DNP" = our dynamic VL-guided negative prompting, "Neg" = conventional static negative prompting. Lower is better for ASR, TR, and FID; higher is better for CLIP. InfT denotes the average time required to generate a single image.

	Ring-a-Bell-16		P4D		Unlearn-Diff		COCO-100		InfT
Method	ASR ↓	TR ↓	ASR	TR	ASR	TR	CLIP ↑	FID ↓	sec ↓
SD v1.4 (no neg)	0.958	0.961	0.960	0.935	0.697	0.734	0.312	–	8.0
VL-DNP (Ours)									
$\omega_{\text{neg}} = 7.5$	0.495	0.521	0.497	0.547	0.310	0.365	0.312	8.0	29.673
$\omega_{\text{neg}} = 15.0$	0.084	0.147	0.225	0.277	0.099	0.171	0.311	12.9	–
$\omega_{\text{neg}} = 20.0$	0.011	0.081	0.113	0.163	0.085	0.139	0.311	15.3	–
$\omega_{\text{neg}} = 25.0$	0.032	0.068	0.086	0.134	0.077	0.130	0.311	15.1	–
Static negative prompting									
$\omega_{\text{neg}} = 7.5$	0.200	0.295	0.298	0.373	0.204	0.252	0.313	107.3	11.967
$\omega_{\text{neg}} = 15.0$	0.000	0.028	0.053	0.078	0.049	0.083	0.296	136.1	–
$\omega_{\text{neg}} = 20.0$	0.025	0.024	0.000	0.030	0.007	0.027	0.277	152.5	–
SAFREE	0.453	0.531	0.377	0.445	0.225	0.267	0.315	104.3	9.01

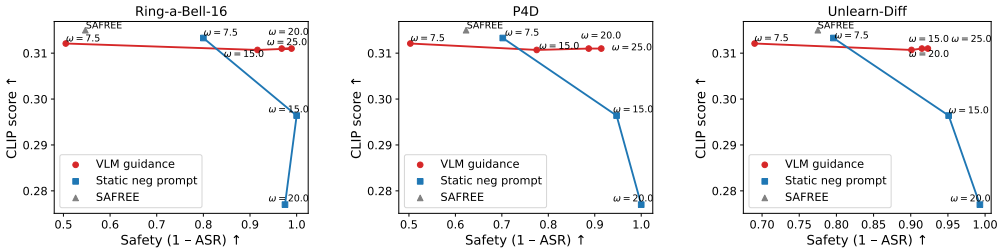

Figure 2: **Safety–alignment Pareto plots.** Circles = dynamic **VLM-guided** prompts; squares = **static** prompts; triangles = **SAFREE** baseline. Axes follow "larger = better": Safety $(1-\text{ASR})$ on x, CLIP on y Markers are labelled by guidance scale ω.

constant ($0.312 \rightarrow 0.311$) and keeps FID in the 8–15 range while still cutting ASR, especially at $\omega_{\text{neg}} = 20$ and 25. SAFREE sits at the opposite extreme of the trade-off. It preserves the highest CLIP score in the table (0.315 on COCO-100) but does so by accepting far poorer safety.

These relations are visualised in Fig. 2. Every static point at $\omega_{\text{neg}} \in \{7.5, 15, 20\}$ is dominated by at least one VLM-guided point, and SAFREE is likewise dominated: for any CLIP it attains, VL-DNP matches or beats it on ASR and usually on FID as well. The dynamic strategy therefore pushes the entire safety–alignment frontier outward.

6 Discussion

Is VLM guidance safer? At a fixed ω the static method can achieve lower raw ASR (e.g. 0.053 vs. 0.225 at $\omega = 15$ on P4D), but only at the cost of a much larger CLIP loss (–5.4 % vs. –0.5 % relative to no-neg). From a multi-objective viewpoint, VLM guidance gives a better *safety–fidelity* trade-off.

Why does dynamic prompting help? *(i)* The VLM may propose narrow, concept-specific negatives ("Male breast", "Buttocks") instead of generic "nsfw", reducing collateral suppression. *(ii)* Prompts evolve: once an artefact disappears the VLM drops the irrelevant negative and targets new risks, avoiding over-suppression.

Limitations & future work Although VLM guidance dominates the static baseline in Pareto terms, its absolute ASR still hinges on the guidance scale. Jointly scheduling *strength* and *content* of

guidance, while trimming the runtime overhead via lightweight vision encoders, remains a promising direction for future work. While the proposed dynamic prompting introduces additional latency due to VLM queries, this overhead may be mitigated by caching intermediate predictions or querying the VLM less frequently (e.g., every k steps). Future work may also leverage lightweight VLMs or distillation-based approaches to enable real-time deployment.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II200940,Foundations of Safe Reinforcement Learning and Its Applications to Natural Language Processing).

References

Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022. URL `https://api.semanticscholar.org/CorpusID:249145348`.

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? *CoRR*, abs/2406.02965, 2024. doi: 10.48550/ARXIV.2406.02965. URL `https://doi.org/10.48550/arXiv.2406.02965`.

Jinho Chang, Hyungjin Chung, and Jong Chul Ye. Contrastive CFG: improving CFG in diffusion models by contrasting positive and negative concepts. *CoRR*, abs/2411.17077, 2024. doi: 10.48550/ARXIV.2411.17077. URL `https://doi.org/10.48550/arXiv.2411.17077`.

Felix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester, and Luca Ambrogioni. Dynamic negative guidance of diffusion models: Towards immediate content removal. In *Neurips Safe Generative AI Workshop 2024*, 2024.

Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. SAFREE: training-free and adaptive guard for safe text-to-image and video generation. *CoRR*, abs/2410.12761, 2024. doi: 10.48550/ARXIV.2410.12761. URL `https://doi.org/10.48550/arXiv.2410.12761`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023.

Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=VyGo1S5A6d.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images ... for now. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVII*, volume 15115 of *Lecture Notes in Computer Science*, pages 385–403. Springer, 2024. doi: 10.1007/978-3-031-72998-0_22. URL https://doi.org/10.1007/978-3-031-72998-0_22.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

notAI tech. Nudenet: Neural nets for nudity classification, detection and selective censoring. 2019.