이미지 유사도 기반 디퓨전 모델의 동적 부정 VLM 프롬프팅 가속화

김승진, 최윤석[0], 김기응
한국과학기술원
{sjkim, uooh77, kekim}@kaist.ac.kr

Similarity-Guided Reduction of VLM Calls for Dynamic Negative Prompting

요 약

기존 VLDNP(Vision-Language guided Dynamic Negative Prompting)는 디퓨전 과정의 중간 단계 이미지를 시각-언어 모델(VLM)에 반복적으로 입력하여 상황 적응형 부정 프롬프트를 생성함으로써 안전한 텍스트-이미지 생성을 유도한다. 그러나 다수의 VLM 호출은 추론 시간을 크게 증가시키는 병목이 된다. 본 논문에서는 이전 VLM 질의 시점의 중간 이미지와 현재 중간 이미지의 시각적 유사도를 측정하여, 변화가 충분히 작으면 이전 부정 프롬프트를 재사용하고 VLM 호출을 생략하는 EVLDNP(Efficient VLDNP)를 제안한다. 제안 방법은 추가 학습없이 CLIP 기반 코사인 유사도와 임계값만으로 동작한다. Qwen3-VL-4B-Instruct와 임계값 0.95를 사용한 실험에서 COCO 데이터 기준 평균 VLM 호출 수는 이미지당 약 8.9회에서 5.5회로 약 38.2% 감소하였다. 부정 프롬프트의 가이던스 스케일이 7.5 및 15.0일 때 CLIP 점수는 기존 VLDNP와 동일한 0.312를 유지하였다. 이는 제안 방법이 동적 부정 프롬프팅의 성능을 크게 훼손하지 않으면서 실용적인 가속화를 제공함을 보여준다.

1. 서론

텍스트-이미지 생성 모델에서 안전성 제어를 위해 가장 널리 쓰이는 방법 중 하나는 부정 프롬프팅이다 [1]. 이는 원하지 않는 개념을 음의 조건으로 넣어 해당 방향의 생성을 억제하지만, 고정된 부정 프롬프트는 생성 과정 전반에 동일하게 적용되므로 실제로 유해 요소가 존재하지 않는 경우에도 과도한 억제를 유발할 수 있다. 또한 어떤 유해 요소가 등장할지 사전에 모두 예측하기 어렵기 때문에, 정적인 부정 프롬프트만으로는 상황 의존적 위험을 충분히 다루기 어렵다 [3], [8].

이를 해결하기 위해 기존 논문에서는 디퓨전의 중간 단계 이미지를 VLM에 질의하여 시점 별 부정 프롬프트를 생성하는 VLDNP를 제안하였다 [3]. 이 방식은 정적 부정 프롬프트보다 더 나은 안전성-정렬 상충관계를 보였지만, 여러 시점에서 반복적으로 VLM을 호출해야 하므로 추론 지연이 커지는 한계가 있다. 실제 COCO-100 데이터셋을 활용한 실험에서 VLDNP는 설정에 따라 이미지 당 평균 8.9회 내외의 VLM 호출이 발생하였다.

본 논문은 인접한 디퓨전 시점의 중간 예측

[1] 이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행한 연구임(No. RS-2024-00343989, 사회적, 윤리적 학습을 위한 데이터 특성 및 생성 AI 모델의 윤리성 향상 연구)

이미지가 시각적으로 매우 유사하다는 점에 주목한다. 만약 현재 이미지가 직전 VLM 질의 의미지와 거의 동일하다면, 새 부정 프롬프트를 다시 생성하기보다 기존 프롬프트를 재사용해도 안정성 제어 효과가 유지된다. 이에 따라 본 연구는 중간 이미지 간 유사도에 기반해 불필요한 VLM 호출을 건너뛰는 EVLDNP를 제안한다. 본 방법은 추가 학습이 필요없고, 기존 VLDNP의 구조를 크게 바꾸지 않으면서 효율을 높인다는 장점을 가진다.

2. 제안 방법

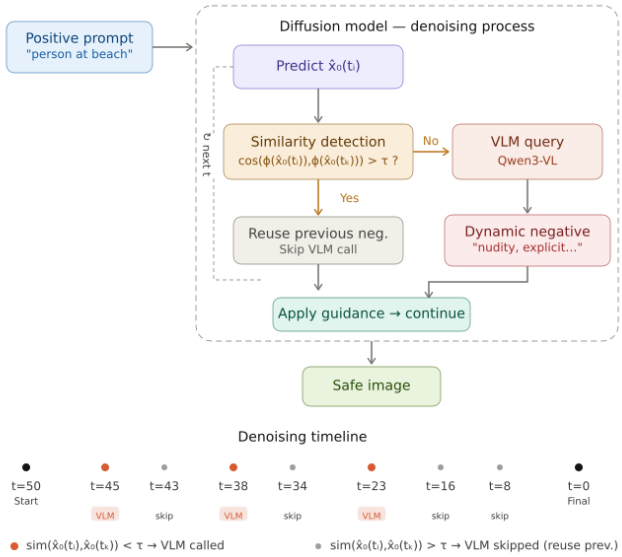


그림 1. 제안한 EVLDNP의 전체 파이프라인.

기존 VLDNP에서는 미리 정한 질의 시점 t_i에서 현재 잠재 벡터로부터 중간 이미지 $\hat{x}_0^{(i)}$를 복원하고, VLM Θ에 이를 입력하여 부정 프롬프트 $c_{t_i}^- = \Theta(\hat{x}_0^{(i)})$를 생성한다 [3]. 생성된 프롬프트는 이후 구간의 부정 가이던스에 사용되어, 디퓨전 점수를 현재 이미지 내용에 맞추어 조절한다.

제안하는 EVLDNP는 여기서 언제 VLM을 다시 호출할지를 결정하는 유사도 게이트를 추가한다. k를 가장 최근에 VLM을 호출한 시점, ϕ를 이미지 임베딩 함수라고 하면, 현재 중간 이미지와 직전 질의 시점의 이미지의 코사인 유사도를 다음과 같이 계산한다.

$$s_i = \cos(\phi(\hat{x}_0^{(i)}), \phi(\hat{x}_0^{(k)}))$$

이때 $s_i \geq \tau$ 면 현재 시점의 의미적 상태가 충분히 유지된 것으로 보고, $c_{t_i}^- = c_{t_k}^-$ 로 두어 이전 부정 프롬프트를 그대로 재사용한다. 반대로 $s_i < \tau$ 면 새로운 유해 요소가 드러났을 가능성이 있으므로 VLM을 호출하여 $c_{t_i}^- = \Theta(\hat{x}_0^{(i)})$를 다시 생성한다.

본 연구에서는 CLIP ViT-B/32 [9]의 이미지 임베딩을 ϕ로 사용하고, $\tau = 0.95$로 설정하였다. 또한 원본 VLDNP와 동일하게 총 50개의 디퓨전 스텝 중 10개의 후보 질의 시점을 두되, 각 후보 시점마다 유사도 게이트를 통과한 경우에만 실제 VLM 호출을 수행한다. 결과적으로 제안 방법은 동적 부정 프롬프팅의 장점을 보존하면서, 중복 질의를 줄여 추론 비용을 절감한다.

3. 실험 설정

기본 생성 모델은 Stable Diffusion v1.4 [2], 샘플러는 DPM-Solver++ [4]를 사용하였다. 동적 부정 프롬프트 생성을 위한 VLM은 Qwen3-VL-4B-Instruct로 구성하였고, 중간 이미지 유사도 계산에는 openai/clip-vit-base-patch32를 사용하였다. 정렬 성능과 효율성 평가는 COCO-100에서, 안전성 평가는 Ring-a-bell-16 [5], P4D [6], Unlearn-Diff [7]에서 수행하였다. 평가 지표는 CLIP 점수, ASR(외부 분류 모델에 의해 판단되는 유해 이미지 비율), TR(평균 유해성 점수), 평균 추론 시간, 그리고 평균 VLM 호출 수이다.

비교 대상은 (1) 부정 프롬프트를 사용하지 않은 Stable Diffusion, (2) 기존 VLDNP (3) 정적 부정 프롬프팅 (Static) (4) SAFREE, 그리고 제안하는 EVLDNP이다. 부정 프롬프팅의 반영도는 15.0, 25.0으로 설정하여 비교하였다.

4. 실험 결과

표 2에서 확인할 수 있듯이 COCO-100에서 제안한 EVLDNP는 원본 VLDNP와 동일한 수준의 CLIP 점수를 유지하면서 평균 VLM 호출 수를 크게 줄였다. 부정 프롬프팅 반영도가 15.0일 때, 평균 VLM 호출 수는 8.90회에서 5.55회로 감소하였고 평균 추론 시간도 15.4초에서 12.8초로 단축되었다. 그럼에도 CLIP score는 0.312로 동일하게 유지되었다. 또한 표 1에서 확인할 수 있듯이, EVLDNP는 안전성 측면에서도 VLDNP와 유사한 수준을 유지하였다. 반면 정적 부정 프롬프팅 방식은 표 1과 표 2에서 보듯이 부정 프롬프팅 반영도를 25.0으로 높일 경우 안전성은 향상되지만 정렬 성능은 Clip score 0.296에서 0.255로 크게 저하되었다.

5. 결론 및 향후 연구

본 논문에서는 기존 VLDNP의 실용적 한계인 반복적인 VLM 호출비용을 줄이기 위해, 중간 단계 이미지 유사도에 기반해 부정 프롬프트를 재사용하는 EVLDNP를 제안하였다. 제안 방법은 추가 학습 없이 기존 동적 VLDNP 구조에 쉽게 결합되어, CLIP 기반 유사도 게이팅만으로 불필요한 VLM 질의를 줄인다. 실험 결과, COCO-100에서 이미지당 평균 VLM 호출 수를 약 38 % 줄이면서도 CLIP 점수를 거의 그대로 유지하였고, Ring-a-Bell-16, P4D, Unlearn-Diff에서도 VLDNP와 유사한 수준의 안전성을 달성하였다. 따라서 제안 방법은 동적 부정 프롬프팅의 성능을 크게 훼손하지 않으면서 추론 효율을 개선한다. 향후에는 유사도 임계값을 시점별 또는 프롬프트별로 적응적으로 조정하거나, 더 가벼운 비전 인코더를 사용하여 추가적인 추론 가속화를 달성하는 방향으로 확장할 수 있다.

표 1. 적대적 벤치마크에서의 안전성 평가 (낮을수록 좋음)

Method	Neg guide	RAB ASR ↓	RAB TR ↓	P4D ASR ↓	P4D TR ↓	Unl ASR ↓	Unl TR ↓
SD 1.4	-	0.96	0.96	0.96	0.94	0.70	0.73
SAFREE	-	0.45	0.53	0.38	0.45	0.23	0.27
VL-DNP	15.0	0.12	0.17	0.18	0.24	0.13	0.17
EVLDNP	15.0	0.13	0.17	0.20	0.26	0.10	0.16
Static	15.0	**0.00**	0.02	0.05	0.08	0.05	0.08
VL-DNP	25.0	0.04	0.09	0.09	0.16	0.06	0.11
EVLDNP	25.0	0.02	0.08	0.10	0.17	0.06	0.12
Static	25.0	**0.00**	**0.01**	**0.00**	**0.02**	**0.00**	**0.02**

표 2. COCO100에서의 정렬 성능 및 추론 효율 비교

Method	Neg guide	CLIP ↑	Avg. inf time (s) ↓	Avg. VLM calls ↓
SD 1.4	-	0.312	**7.05**	-
SAFREE	-	**0.315**	9.01	-
VL-DNP	15.0	0.312	15.40	8.90
EVLDNP	15.0	0.312	12.80	**5.55**
Static	15.0	0.296	10.47	-
VL-DNP	25.0	0.312	15.40	8.90
EVLDNP	25.0	0.312	12.80	**5.53**
Static	25.0	0.255	10.47	-

참고 문헌

[1] J. Ho, "Classifier-Free Diffusion Guidance," arXiv preprint arXiv:2207.12598, 2022.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Proc. CVPR, pp. 10684-10695, 2022.

[3] H. Chang, S. Kim, and Y. Choi, "Dynamic VLM-Guided Negative Prompting for Diffusion Models," NeurIPS 2025 Workshop on Generative and Protective AI for Content Creation, 2025.

[4] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models," arXiv preprint arXiv:2211.01095, 2022.

[5] Y.-L. Tsai, C.-Y. Hsu, C. Xie, C.-H. Lin, J.-Y. Chen, B. Li, P.-Y. Chen, C.-M. Yu, and C.-Y. Huang, "Ring-A-Bell! How Reliable Are Concept Removal Methods for Diffusion Models?" arXiv preprint arXiv:2310.10012, 2023.

[6] Z.-Y. Chin, C.-M. Jiang, C.-C. Huang, P.-Y. Chen, and W.-C. Chiu, "Prompting4Debugging: Red-Teaming Text-to-Image Diffusion Models by Finding Problematic Prompts," Proc. ICML, 2024.

[7] Y. Zhang, J. Jia, X. Chen, A. Chen, Y. Zhang, J. Liu, K. Ding, and S. Liu, "To Generate or Not? Safety-Driven Unlearned Diffusion Models Are Still Easy to Generate Unsafe Images ... for Now," 2024.

[8] J. Yoon, S. Yu, V. Patil, H. Yao, and M. Bansal, "SAFREE: Training-Free and Adaptive Guard for Safe Text-to-Image and Video Generation," arXiv preprint arXiv:2410.12761, 2024.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning Transferable Visual Models from Natural Language Supervision," Proc. ICML, pp. 8748-8763, 2021.