DeRO: Decompose and Recompose Text Optimization for Correcting Semantic in Negation-Aware Image Generation

Anonymous ACL submission

Abstract

Prompt embeddings in text-to-image diffusion models have improved through the use of multiple text encoders and attention masking over text tokens, leading to stronger text–image alignment and improved controllability. However, alignment remains weak for negation-based prompts. In this paper, we analyze text embeddings and show that implicit word-level biases cause negation expressions to be ignored. We present **DeRO**, which optimizes the original prompt by identifying its precise semantic. The method applies SVD to prompt embeddings together with auxiliary prompts that are semantically similar, in order to obtain the corresponding semantic subspace. While projecting the text embedding onto this subspace improves alignment, naïve projection causes substantial loss of information contained in the original prompt embedding. We perform a one-time optimization that matches the token vectors of implicit biased words and negation adverbs with the projected embedding, enabling to obtain an optimal prompt embedding that is semantically aligned with the given negation prompt while preserving unrelated words. Through experiments on both object and concept negation benchmarks, we show that **DeRO** achieves approximately a 21% performance improvement for object erasure and a 12% improvement for concept erasure, consistently outperforming prior methods while maintaining superior computational efficiency.

1 Introduction

"Do not think of an elephant." Attempts to suppress a thought often paradoxically reinforce it. Psycholinguistic studies have shown that negation is not a simple elimination of meaning, but a cognitive process in which the affirmative concept is first activated and only subsequently inhibited (MacDonald and Just, 1989; Kaup et al., 2007; Tian and Levy, 2022). As a result, negated concepts tend to persist in mental representations rather than being fully removed.

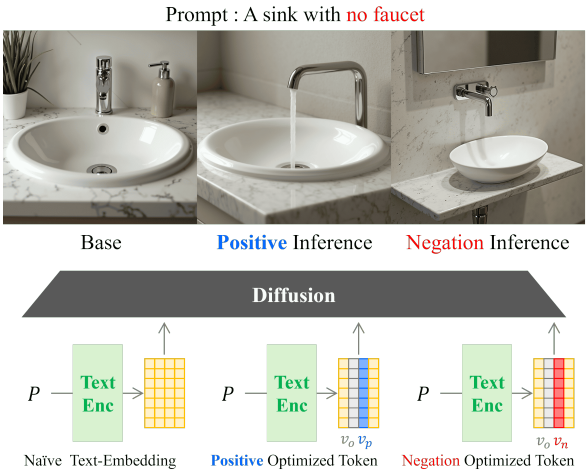

Figure 1: **Conceptual visualization of image generation under negation prompts.** When using standard text encoder embeddings, negated expressions such as "no faucet" are often not properly reflected in the generated image. In contrast, text token embeddings optimized with our method enable the model to correctly respond to negated expressions. Our approach performs a one-time optimization prior to image generation, allowing efficient and prompt-aligned generation without modifying the image synthesis model itself.

This phenomenon extends beyond human cognition to deep learning systems. Text-to-image (T2I) models similarly struggle to handle negation, allowing explicitly negated concepts to continue influencing visual representations. For example, given a prompt such as "a chair without a backrest," T2I models often generate chairs with a backrest, as shown in the first row of Fig. 2. Quantitative analysis further re-

veals that sentence embeddings obtained from text encoders exhibit high cosine similarity between negated and non-negated prompts (e.g., "a chair with a backrest"), indicating that negation is not clearly separated in the embedding space. This lack of angular separation suggests that negated semantics remain entangled with their positive counterparts, leading to persistent failures in satisfying negation constraints and poor text–image semantic alignment.

Few works directly address negation-aware text-to-image generation. (Li et al., 2023) weakens undesired concepts by suppressing the embeddings of unwanted tokens, but this often results in degradation of image structure or style. (Li et al., 2025; Gandikota et al., 2023a; Gong et al., 2024) achieve strong erasure by modifying diffusion parameters, yet result in permanent concept deletion, leading to irreversibility and potential catastrophic forgetting. In reward-based text optimization approaches (Na et al., 2025; Lee et al., 2025), constructing a reliable reward model that can accurately distinguish negation is highly challenging, which hinders precise negation-aware inference.

In this work, we propose **DeRO**, which enables challenging prompts such as negation to form more accurate semantic representations, leading to better-aligned images. Our method operates entirely in the text embedding space, requiring neither model training nor iterative optimization. By selectively editing only a small set of word token embeddings corresponding to the object (e.g., chair) and the negation term (e.g., without), while preserving all other semantic components, our method achieves accurate negation inference.

In the first stage, we apply SVD to the given prompt together with semantically aligned prompts that do not contain negation expressions, thereby identifying a semantically aligned subspace. The source prompt embedding is then projected onto this subspace to obtain a semantically aligned embedding.

Since naïvely projected embeddings discard substantial information from the original embedding, we extract only the object and negation word vectors from the projected embedding. In the second stage, we match the object and negation word vectors from the original embedding with those from the projected embedding, resulting in an optimized embedding that is semantically aligned while preserving all other information. Consequently, this selective optimization enables accurate image generation for negation prompts while requiring the optimization of only three vectors, avoiding catastrophic forgetting and achieving high computational efficiency.

Contributions. Overall, our contributions are as follows:

- We identify the underlying reasons why accurate image generation from negation prompts fails, and show that improving alignment can be achieved by selectively modifying only specific embedding vectors.
- We propose **DeRO**, a novel negation-aware text optimization method that enables the generation of text aligned images.
- Through extensive experiments on both object and concept negation, we demonstrate that our method significantly outperforms existing negation methods, achieving a 21% performance improvement for object negation and a 12% improvement for concept negation by optimizing only three vectors.

2 Related Work

Text–Image Alignment Various approaches have been proposed to improve text–image alignment in generative models. DATE (Na et al., 2025) optimizes text embeddings through a reward-driven iterative procedure, while SoftREPA (Lee et al., 2025) enhances overall alignment via contrastive learning. Composable Diffusion (Liu et al., 2022) decomposes complex prompts into multiple components and combines their corresponding diffusion processes to better handle compositional semantics. ReFACT (Arad et al., 2024) modifies key and value parameters in cross-attention layers through training to remove specific concepts, and NAG (Chen et al., 2025) employs guidance based on normalized extrapolation of attention features to encourage prompt-aligned generation.

However, these methods often rely on iterative optimization, additional training, or task-specific designs, limiting their generalization

to negation scenarios that require semantic direction shifts rather than explicit concept suppression. In particular, approaches focused on concept removal or guidance adjustment do not fully address the underlying misalignment between the expressed prompt semantics and the generated content.

Semantic Erase Prior efforts to remove undesirable concepts from T2I models largely fall into two categories: model editing and text editing. Model editing methods enforce strong and permanent erasure by directly modifying model parameters, but sacrifice flexibility and reversibility. Rather than enabling conditional control, these approaches suppress concept-aligned representations globally through attention, latent-space, or neuron-level interventions (Lu et al., 2024; He et al., 2025; Li et al., 2025; Gandikota et al., 2023a). While effective at eliminating target concepts, such global modifications often induce catastrophic forgetting and degrade generalization. Even methods that improve spatial locality, such as Receler (Huang et al., 2024), remain irreversible parameter-level edits that fundamentally alter internal representations.

In contrast, text editing methods avoid modifying the generator and instead manipulate textual conditioning signals. SupEOT (Li et al., 2023) weakens undesired concepts via embedding normalization and inference-time optimization, while RECE (Gong et al., 2024) and CCRT (Han et al., 2025) refine cross-attention or apply continuous removal strategies. However, these approaches primarily target concept erasure rather than conditional or context-aware negation, limiting their ability to precisely align generated images with negated prompts.

3 Method

3.1 Motivation

FLUX.1 (Labs et al., 2025) employs two text encoders: CLIP (Radford et al., 2021) and T5 (Raffel et al., 2020). The CLIP text encoder maps an input prompt p to a single pooled embedding $p_{\text{CLIP}} \in \mathbb{R}^{768}$, which captures the global semantic meaning of the prompt and modulates the overall generation behavior of the diffusion model. In contrast, T5 encodes the prompt at the token level, producing token-wise embeddings p_{T5}, including special start-of-text, end-of-sequence, and padding tokens up to a fixed length.

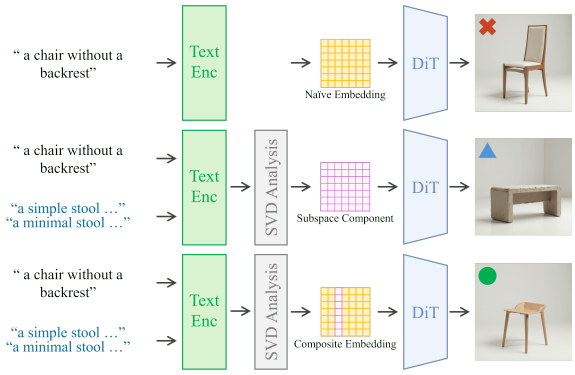

Figure 2: **Image generation via text component decomposition.** Naively using text embeddings, the prompt "a chair without a backrest" is not visually reflected. Extracting a shared semantic component via SVD enables the absence of a backrest but discards much of the original context. Applying this component only to the chair token preserves contextual information while correctly expressing the missing backrest.

The first image in Fig. 3 is generated naively using a prompt containing a negation expression. The phrase without a backrest in p is ignored, yielding an image that contradicts the intended semantics.

Prior work (Biswas et al., 2025) analyzes a shared subspace between erased and preserved concepts using SVD, motivating our investigation of this failure. We encode the target prompt p together with n auxiliary sentences conveying the same meaning to form an embedding matrix E, and perform SVD, $E = U\Sigma V^\top$. Projecting e_p onto the shared subspace yields $\tilde{e}_p$, which preserves the core semantics of p (Fig. 3, middle) but alters the original style due to excessive subspace modification.

We selectively swapping only the *chair* token embedding in e_p with its counterpart from $\tilde{e}_p$ preserves the negated semantics while maintaining the original style, as shown in the last row of Fig. 3.

3.2 Semantic Subspace

Based on the preceding analysis, we find that intrinsic biases in object-related words hinder accurate negation. To address this issue, we adopt two strategies. First, we construct a semantically aligned embedding space using

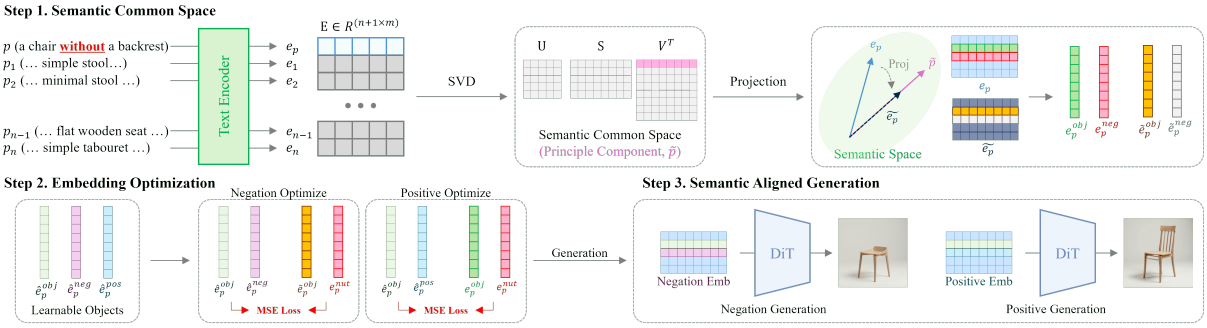

Figure 3: **Overview of DeRO.** extract a shared semantic component through SVD analysis and projecting embeddings onto the principal components to identify a negation-consistent semantic direction that preserves the original context. The method then optimizes only the target object, chair, token by minimizing an MSE loss between the original embedding and the SVD-projected embedding, enabling faithful negation-aware image generation.

auxiliary sentences with respect to the given prompt. Second, we optimize the text embeddings so that representations containing negation accurately reflect the intended semantics.

Let $\mathcal{P}_n = \{p_1, p_2, \ldots, p_n\}$ denote a set of auxiliary prompts that convey semantics similar to a target prompt p without explicitly using negation. As illustrated in Fig. 2, we encode the target prompt p together with each prompt in $\mathcal{P}_n$ using a frozen text encoder, obtaining a set of embeddings $\{\mathbf{e}_p, \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$.

To find a shared semantic direction, we perform singular vector decomposition (SVD) over the embeddings of the target prompt and the auxiliary prompts. Specifically, we construct an embedding matrix $E = [\mathbf{e}_p,\ \mathbf{e}_1,\ \mathbf{e}_2,\ \ldots,\ \mathbf{e}_n]$ and center it by subtracting the mean embedding across columns.

We then apply SVD to extract the principal components that characterize the dominant directions of variation within the common subspace. Empirically, we find that the first principal component consistently captures a coherent and shared semantic direction. We project $\mathbf{e}_p$ onto the first principal component, $\tilde{e}_p$ to extract the shared component of the intended concept. This projection preserves the common semantics captured by the auxiliary prompts while suppressing object-specific bias.

3.3 Embedding Optimization

Let $\hat{\mathbf{e}}^{\text{obj}}$, $\hat{\mathbf{e}}^{\text{neg}}$, and $\hat{\mathbf{e}}^{\text{pos}}$ denote the learnable embeddings corresponding to the object token, the negation token, and its positive counterpart (e.g., *with* for *without*), respectively. Our goal is to update the object- and negation-related embeddings such that their combined representation aligns with the bias-reduced target embedding $\tilde{\mathbf{e}}^{\text{obj}}$ derived from the semantic common space, while preserving the original affirmative semantics of the prompt.

Negation alignment. To this end, we align the concatenated embedding of the object and negation tokens with the negation-aware target representation. Since these representations differ in dimensionality, we introduce a neutral embedding $\mathbf{e}_{\text{neu}}$ to ensure dimensional compatibility. The negation alignment loss is defined as,

$$\mathcal{L}_{\text{neg}} = \left\| [\hat{\mathbf{e}}^{\text{obj}}, \hat{\mathbf{e}}^{\text{neg}}] - [\tilde{\mathbf{e}}^{\text{obj}}, \mathbf{e}_{\text{neu}}] \right\|_2^2, \quad (1)$$

where $[\cdot, \cdot]$ denotes vector concatenation.

Positive alignment. To prevent semantic drift and preserve the original meaning of the prompt, we additionally optimize alignment between the positive counterpart embedding and the original prompt semantics. Specifically, we align the concatenation of the object and positive token embeddings with the corresponding components of the original prompt embedding $\mathbf{e}_p$. The positive alignment loss is defined as,

$$\mathcal{L}_{\text{pos}} = \left\| [\hat{\mathbf{e}}^{\text{obj}}, \hat{\mathbf{e}}^{\text{pos}}] - [\mathbf{e}_p^{\text{obj}}, \mathbf{e}_{\text{neu}}] \right\|_2^2, \quad (2)$$

where $\mathbf{e}_p^{\text{obj}}$ denotes the object-related component of the original prompt embedding.

Overall objective. The final optimization objective combines the two alignment losses as

$$\mathcal{L} = \lambda_{\text{neg}}\, \mathcal{L}_{\text{neg}} + \lambda_{\text{pos}}\, \mathcal{L}_{\text{pos}}, \quad (3)$$

Model	TIFA Score (↑)						Preservation	Image Quality	
	missing	empty	nothing	nobody	without	no	CLIP-P (↑)	PickScore (↑)	FID (↓)
SD1.5	0.142	0.252	0.525	0.950	0.166	0.762	0.742	21.245	195.207
SD1.5 + DATE	0.145	0.255	0.525	0.818	0.166	0.762	0.744 (+0.002)	21.229 (-0.019)	194.615 (-0.592)
SD1.5 + SupEOT	0.155	0.250	0.525	0.955	0.250	0.810	0.739 (-0.003)	21.174 (-0.071)	208.241 (+13.034)
SD3	0.151	0.250	0.752	0.909	0.250	0.762	0.745	21.751	201.016
SD3 + SoftREPA	0.157	0.281	0.763	0.909	0.167	0.738	0.750 (+0.005)	21.751 (+0.000)	199.158 (-1.857)
FLUX	0.845	0.450	0.855	0.805	0.250	0.762	0.757	22.265	194.709
FLUX + NAG	**0.852**	0.920	0.855	0.820	**0.667**	**0.942**	0.754 (-0.003)	22.018 (-0.247)	194.965 (+0.255)
FLUX + Ours	0.765	**0.950**	**0.865**	**1.000**	**0.667**	0.802	**0.758** (+0.001)	**22.268** (+0.003)	**192.375** (-2.334)

Table 1: **Comparison of performance on negation understanding.** TIFA measures how well the generated image aligns with the negated semantics of the prompt, while CLIP-P evaluates the preservation of non-negated semantic components. Image quality and fidelity are assessed using PickScore and FID, respectively. Changes relative to the base model are shown in parentheses, with improvements highlighted in blue and degradations in red. Best scores are bolded.

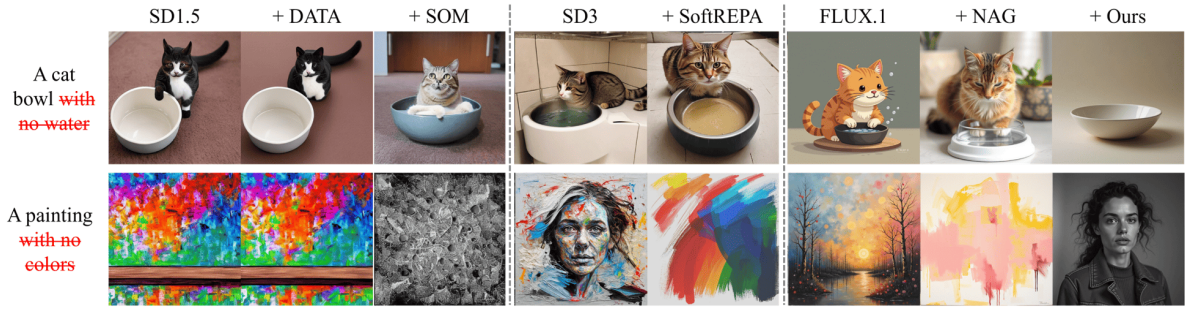

Figure 4: **Visualization of object negation.** Local object negation suppresses specific target objects within confined regions, whereas global concept negation affects the overall appearance of the image.

where λ_{neg} and λ_{pos} control the trade-off between negation alignment and preservation of affirmative semantics.

In practice, we set the neutral embedding $\mathbf{e}_{\text{neu}}$ initialized from the negation token embedding extracted from $\mathbf{e}_{\text{p}}$ for dimensional matching, expressed in Eq. 1 and Eq. 2. We find that cosine similarity preserves angular alignment, while mismatches in embedding magnitude can lead to unexpected and semantically incorrect image generation. Therefore, we adopt a mean squared error to jointly enforce directional and magnitude alignment.

4 Experiments

4.1 Implementation Details

We use FLUX.1 (Labs, 2024) as the backbone T2I model, which is a transformer-based diffusion model with two text encoders, CLIP and T5. All experiments are conducted on a single NVIDIA A100 GPU with no training. Our method performs a one-time optimization of the text embedding at the earliest stage of the iterative image generation process.

Field	Content
Prompt	a shoe without a strip
Object	shoe
Target	without
Auxiliary prompts	
– slip-on shoe with minimalist design	
– shoe with simple functional form	
– slip-on shoe with simple functional form	
– a seamless upper and no fastening elements shoe	
– minimalist shoe of an unbroken, lace-free silhouette	
– shoe designed smooth form and no closures	

Table 2: **Example of our negation evaluation dataset.**

4.2 Object negation

Baselines We compare our method against DATE (Na et al., 2025), SoftREPA (Lee et al., 2025), NAG (Chen et al., 2025), and SupEOT (Li et al., 2023). DATE optimizes text embeddings using a reward model to improve prompt–image alignment, while SoftREPA optimizes text embeddings through contrastive learning. SupEOT performs concept negation by directly removing information related to

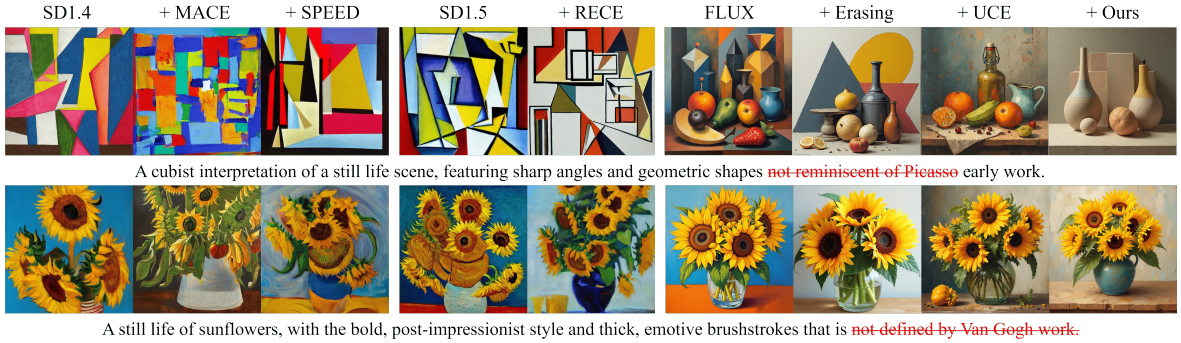

Figure 5: **Visualization of concept negation** Results are grouped by base architectures (SD1.4, SD1.5, and FLUX.1-based models).

the undesired concept padding tokens. NAG suppresses unwanted attributes by explicitly controlling attention activations during sampling, in a manner similar to Classifier-Free Guidance (Ho and Salimans, 2022).

Evaluation Protocol To evaluate object negation, we construct a dataset of 480 prompts containing explicit negation. Each prompt is paired with 15–20 semantically aligned auxiliary prompts that implicitly express the same negated concept. Negation tokens include *missing*, *empty*, *nothing*, *nobody*, *without*, and *no*. The examples of object negation dataset is on Table 2 and additional details in Appendix C.4.

We adopt TIFA (Hu et al., 2023), a VQA-based metric to score whether generated image satisfy negation prompt. To ensure that negation does not degrade non-negated semantics, we introduce CLIP-P, which measures CLIP-Score between the generated image and the corresponding positive prompt without negatio, and evaluate visual quality using FID and perceptual preference using PickScore (Kirstain et al., 2023).

Result Table 2, values in blue indicate improvements over the base model, whereas red denote performance degradation. SupEOT exhibits a noticeable decrease in PickScore along with a substantial increase in FID, indicating that its padding-based object erasure negatively affects overall image quality. NAG yields lower CLIP-P scores, indicating weakened alignment with the original prompt due to sampling-based guidance adjustments. In contrast, our method consistently achieves higher TIFA scores while maintaining high CLIP-P scores and low FID, demonstrating improved negation understanding without compromising image quality.

As shown in Fig. 4, DATE and SoftREPA fail to handle negation, producing results similar to base, while SupEOT suppresses the target concept at the expense of image quality. In contrast, our method removes the negated concept semantically while preserving visual quality.

4.3 Artist Concept Negation

Baselines We demonstrate the effectiveness of our method for both object and concept negation by comparing it against MACE (Lu et al., 2024) (training-based), SPEED (Li et al., 2025) (training-free, SD1.4), RECE (Gong et al., 2024) (SD1.5), and UCE (Gandikota et al., 2024) and Erasing (Gandikota et al., 2023a) (training-based, FLUX.1). Details are provided in Supplementary D.4.

Evaluation Protocol Following (Biswas et al., 2025), we adopt the artistic concept negation benchmark from (Gandikota et al., 2023a), where prompts explicitly negate artist styles (e.g., not in a Picasso's style). For evaluation, following (Biswas et al., 2025), we compute LPIPS (Zhang et al., 2018) between each method and the baseline. The remaining evaluation metrics and protocols are identical to those used in the object negation setting.

Result Fig. 5 and Tab. 5 summarize the qualitative and quantitative results of artist concept negation. Most baselines exhibit a slight degradation in CLIP-P compared to their corresponding base models, indicating reduced prompt–image alignment after concept removal. In contrast, our method and UCE largely preserve CLIP-P, suggesting that prompt seman-

Model	Concept "Picasso" Negation				Concept "Van Gogh" Negation			
	LPIPS (↑)	TIFA (↑)	CLIP-P (↑)	PickScore (↑)	LPIPS (↑)	TIFA (↑)	CLIP-P (↑)	PickScore (↑)
SD1.4	–	0.083	0.720	20.567	–	0.125	0.843	21.672
SD1.4 + SPEED	0.199	0.036	0.715 (-0.0049)	20.185 (-0.383)	0.134	0.042	0.684 (-0.0358)	20.33 (-1.339)
SD1.4 + MACE	0.181	0.179	0.763 (-0.0567)	20.547 (-0.021)	0.146	0.167	0.731 (-0.0889)	20.870 (-0.537)
SD1.5	–	0.112	0.772	19.651	–	0.063	0.749	21.708
SD1.5 + RECE	**0.233**	0.143	0.598 (-0.173)	19.165 (-0.486)	0.148	0.155	0.465 (-0.206)	18.549 (-3.159)
FLUX.1	–	0.444	0.753	22.075	–	0.205	0.732	22.109
FLUX.1 + Erasing	0.127	0.517	0.711 (-0.055)	21.984 (-0.090)	0.169	0.	0.735 (+0.004)	22.198 (+0.089)
FLUX.1 + UCE	0.087	0.464	0.764 (+0.010)	22.189 (+0.114)	0.134	0.422	0.742 (+0.009)	22.248 (+0.139)
FLUX.1 + Ours	0.175	**0.525**	**0.773** (+0.020)	**22.499** (+0.424)	**0.173**	**0.459**	**0.781** (+0.048)	**23.109** (+1.000)

Table 3: **Quantitative comparison on Picasso and Van Gogh related negation prompts.** Parentheses denote differences from the corresponding base modela and best value per column is bolded.

Figure 6: **Comparison between Erasing and Ours.** The Erasing method relies on model training, which can lead to the generation of images that are unrelated to the given prompt.

Figure 7: **Failure cases.** When auxiliary embeddings are inadequately constructed, our method may fail to identify the correct semantic direction. Additionally, complex or ambiguous prompt formulations can further obscure the effects of concept removal, making the negation less perceptible even when concept-related terms are properly handled.

tics remain well maintained despite artist negation. In particular, Fig. 6 shows that the Erasing method, which retrains the model, can cause the generated images to deviate from the input prompt. (e.g, (top) a snow scene unrelated to the prompt (bottom) images that are irrelevant to the objects) For LPIPS, relatively high values are observed for methods based on SD1.4 and SD1.5, particularly on the Picasso concept. Through direct visual inspection, we find that these high LPIPS scores often arise not from effective artist concept removal, but from structural degradation of the generated images caused during the erasure process.

Failure Cases Fig. 7 presents representative failure cases of our method. For object negation, failures occur when auxiliary prompts do not accurately capture the semantics. Using donut-like auxiliary prompts for "a cupcake with no frosting" induces a donut-oriented subspace, while for "a violin case without a violin," the difficulty of constructing suitable auxiliary prompts leads to incomplete negation. For concept negation, although the Monet concept is suppressed, descriptive attributes in the prompt still dominate the generation, making the effect of concept removal less perceptible and leading to lower LPIPS scores.

4.4 Additional Result

Computational Efficiency Fig. 8 illustrates computational efficiency in terms of inference time and performance. Since each method is built upon a different baseline model, both inference time and performance are reported as relative changes with respect to their corresponding baselines. Iterative optimization-based approaches (e.g., DATE and SupEOT), as well as training-based methods (e.g., Erasing), incur substantial inference-time overhead.

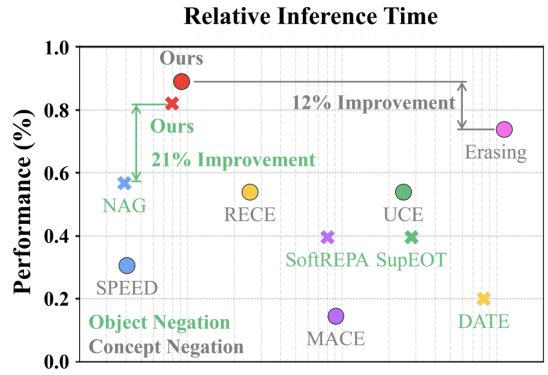

Figure 8: **Computational overhead vs. performance trade-off.** The x-axis shows inference-time relative to the base model, and y-axis indicates relative performance improvement. Methods closer to the upper-left achieve better efficiency.

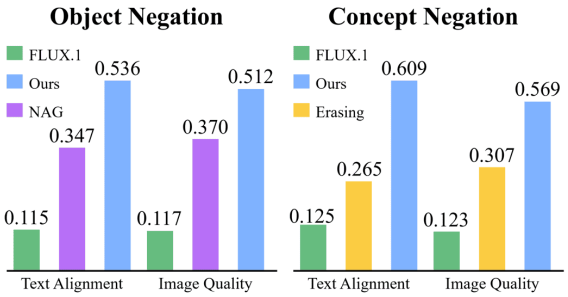

Figure 9: **Results of User Study** show that our method achieves superior performance across both tasks.

Auxiliary Sentences (#)	Negate		Positive	
	TIFA	CLIP-P	TIFA	CLIP-P
3	0.695	0.751	0.718	0.710
6	0.698	0.759	0.721	0.689
9	0.711	0.712	0.725	0.742
12	0.718	0.721	0.725	0.714
15	0.725	0.744	0.738	0.741
18	0.725	0.758	0.738	0.692

Table 4: **Effect of the number of auxiliary sentences.**

In contrast, our method requires training only three tokens, resulting in significantly lower computational cost while maintaining strong performance, demonstrating superior efficiency. (more details in Supplementary D.4.)

User Study We also present the user study results in Fig. 9, which show that our method outperforms the baselines on both object and concept negation without compromising image quality. Detailed user study protocols are provided in Supplementary D.5.

Ablation Studies We analyze the effect of auxiliary prompts by progressively increasing their number and evaluating performance at each stage for both negated and non-negated prompts.

As shown in Table 4, incorporating auxiliary prompts improves semantic alignment. Performance gains saturate beyond a certain number of auxiliary prompts. The result indicate that the semantic quality of auxiliary prompts is more critical than quantity and although increasing the number of prompts can be beneficial, excessive prompts do not cause overfitting but lead to diminishing efficiency.

5 Limitations

The effectiveness of our approach relies on accurately estimating a semantically meaningful subspace from auxiliary prompts. When this subspace is inadequately constructed, embedding redirection becomes less effective. Addressing this limitation, future work will explore more robust subspace discovery mechanisms and strategies to reduce reliance on curated auxiliary prompt sets.

6 Conclusion

We present DeRO, an inference-time text embedding optimization method for mitigating negation failures in diffusion models. DeROleverages auxiliary prompts to identify a semantically aligned subspace and redirects the target word embedding away from undesired objects or concepts, rather than explicitly suppressing them. This subspace is obtained through SVD analysis, enabling precise semantic manipulation while preserving the original prompt intent. Comprehensive evaluations across both object- and concept-level negation demonstrate that **DeRO** generalizes effectively to diverse negation scenarios, in contrast to prior methods that are often specialized for specific cases. We believe that our method offers an alternative perspective on prompt alignment and will make a meaningful contribution to the advancement of AI generative systems.

References

Dana Arad, Hadas Orgad, and Yonatan Belinkov. 2024. Refact: Updating text-to-image models by editing the text encoder. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2537–2558.

Shristi Das Biswas, Arani Roy, and Kaushik Roy. 2025. Cure: Concept unlearning via orthogonal representation editing in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, and 1 others. 2025. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*.

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. 2025. Normalized attention guidance: Universal negative guidance for diffusion model. *arXiv preprint arXiv:2505.21179*.

Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. 2024. Fluxspace: Disentangled semantic editing in rectified flow transformers. *arXiv preprint arXiv:2412.09611*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Zihao Fu, Ryan Brown, Shun Shao, Kai Rawal, Eoin Delaney, and Chris Russell. 2025. Fairimagen: Post-processing for bias mitigation in text-to-image models. *arXiv preprint arXiv:2510.21363*.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. 2023a. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2023b. Unified concept editing in diffusion models. *arXiv preprint arXiv:2308.14761*.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. 2024. Reliable and efficient concept erasure of text-to-image diffusion models. *arXiv preprint arXiv:2407.12383*.

Tingxu Han, Weisong Sun, Yanrong Hu, Chunrong Fang, Yonglong Zhang, Shiqing Ma, Tao Zheng, Zhenyu Chen, and Zhenting Wang. 2025. Continuous concepts removal in text-to-image diffusion models. In *Advances in Neural Information Processing Systems: NeuralPS 2025*.

Qinqin He, Jiaqi Weng, Jialing Tao, and Hui Xue. 2025. A single neuron works: Precise concept erasure in text-to-image diffusion models. *arXiv preprint arXiv:2509.21008*.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*.

Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. 2024. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *European Conference on Computer Vision*, pages 360–376. Springer.

Barbara Kaup, Jana Ludtke, and Rolf A Zwaan. 2007. The role of negation in sentence comprehension: A psycholinguistic review. *Language and Linguistics Compass*, 1(1-2):1–17.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663.

Black Forest Labs. 2024. Flux. `https://github.com/black-forest-labs/flux`.

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, and 2 others. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *Preprint*, arXiv:2506.15742.

Jaa-Yeon Lee, Byunghee Cha, Jeongsol Kim, and Jong Chul Ye. 2025. Aligning text to image in diffusion models is easier than you think. *arXiv preprint arXiv:2503.08250*.

Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Tao Liang, Yanbin Hao, Guojun Ma, and Fuli Feng. 2025. Speed: Scalable, precise, and efficient concept erasure for diffusion models. *arXiv preprint arXiv:2503.07392*.

Senmao Li, Joost van de Weijer, Fahad Khan, Qibin Hou, Yaxing Wang, and 1 others. 2023. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pages 423–439. Springer.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. 2024. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440.

Maryellen C MacDonald and Marcel Adam Just. 1989. Processing negation and disjunction. *Journal of Memory and Language*, 28(6):687–715.

Byeonghu Na, Minsang Park, Gyuwon Sim, Donghyeok Shin, HeeSun Bae, Mina Kang, Se Jung Kwon, Wanmo Kang, and Il chul Moon. 2025. Diffusion adaptive text embedding for text-to-image diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and 1 others. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. 2024. Finetuning text-to-image diffusion models for fairness. In *The Twelfth International Conference on Learning Representations*.

Yingdong Shi, Changming Li, Yifan Wang, Yongxiang Zhao, Anqi Pang, Sibei Yang, Jingyi Yu, and Kan Ren. 2025. Dissecting and mitigating diffusion bias via mechanistic interpretability. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8192–8202.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Yankai Tian and Roger P Levy. 2022. Negation may be harder than we think: A unified account of negation processing. *Trends in Cognitive Sciences*, 26(4):308–319.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

DeRO: Decompose and Recompose Text Optimization for Correcting Semantic in Negation-Aware Image Generation

Supplementary Material

In this supplementary material, we provide,

A Future Research

T2I models have significantly advanced beyond relying solely on simple textual prompts, incorporating richer instruction-following mechanisms (Wu et al., 2025) and attention-based conditioning to better focus on relevant textual cues (Cao et al., 2025). These advances in prompt handling have been shown to improve the controllability and expressive capacity of generative models.

Despite this progress, purely text-based control remains insufficient for handling challenging prompts such as negation. While this work focuses specifically on negation-aware generation, limitations persist when extending to more complex linguistic phenomena, including relational and logical reasoning. Future research should therefore explore how semantic direction-based approaches can be generalized to broader classes of challenging prompts, enabling more robust and faithful text-to-image alignment under complex compositional conditions.

B Motivation

B.1 Text Embedding Analysis

We observe that negation expressed in text prompts is not faithfully encoded. As illustrated in the first example of Fig. 1, the prompt `a chair without a backrest` still produces an image containing a backrest, directly contradicting the intended semantics.

Table A: **Cosine similarity between target (p_t), positive (p_p), and base (p_b) prompts under negation settings**

Model	CLIP			T5		
Pairs	p_t-p_p	p_p-p_b	p_b-p_t	p_t-p_p	p_p-p_b	p_b-p_t
Similarity	0.998	0.999	0.998	0.593	0.841	0.533

We examine prompt embeddings produced by two widely used text encoders in diffusion models, namely T5 (Raffel et al., 2020) and CLIP (Radford et al., 2021). Using Qwen2.5-Instruct model (Team, 2024), we construct 100 prompt triplets from the MS-COCO (Lin et al., 2014) dataset, focusing on negation expressions. Each triplet consists of a target prompt p_t containing a negation constraint, a positive prompt p_p that explicitly expresses the opposite meaning of positive, and a base prompt p_b that omits the constraint (e.g., `a man wearing no glasses`, `a man wearing glasses`, and `a man`). We compute cosine similarities between the corresponding embeddings to assess how well negation semantics are encoded in the embedding space.

As shown in Tab. A, CLIP embeddings of p_t and p_p remain nearly identical despite their opposite meanings, and both are also highly similar to p_b. This indicates that negation is not represented as a discriminative direction in the embedding space. Instead, negated prompts remain strongly aligned with positive and base prompts. In other words, the negation signal fails to form an independent control axis and is absorbed by dominant semantic components.

This phenomenon is also evident in prompts such as `a painting painted by Picasso` and `a painting painted not by Picasso`. Although these prompts are semantically contradictory, their corresponding embeddings exhibit high similarity, suggesting that the `Picasso` concept acts as a dominant di-

Figure A: **Visualization of profession gender debiasing.** Gender bias inherent in profession-related prompts can be controlled via gender token vector decomposition.

rectional component in the embedding space. Consequently, even when the negation `not by Picasso` is explicitly specified, the resulting embedding remains strongly aligned with the `Picasso` direction, while the negation component itself is not sufficiently separated to counteract this influence. At the representation level, such concept leakage causes the model to behave as if the negated expression were absent, with the concept remaining implicitly encoded and consequently biasing the generation trajectory toward the positive biased direction.

B.2 Composite Embedding

We show that decomposing text embedding vectors enables explicit extraction and control of bias components at the word or sentence level. A well-known issue in Text-to-Image (T2I) models is gender bias associated with professions (Fu et al., 2025; Shi et al., 2025; Shen et al., 2024), which prior work has mainly addressed through fine-tuning or additional regularization. While effective, such approaches often require retraining or additional data curation. In contrast, we demonstrate a simpler embedding-level alternative.

For example, a prompt such as *a dentist examining a patient* often produces images of a male dentist, even without any explicit gender specification, indicating that the profession term *dentist* implicitly encodes gender bias. Let p denote this prompt, and let $\mathbf{v}_{\text{dentist}}$ be the token-level embedding corresponding to *dentist* obtained from a text encoder. Similarly, we encode the auxiliary prompts *a man examining a patient* and *a woman examining a patient*, and extract the corresponding token embeddings $\mathbf{v}_{\text{man}}$ and $\mathbf{v}_{\text{woman}}$.

To isolate the gender-related component embedded in the profession representation, we adopt the same linear projection formulation used for attention outputs. Let $\mathbf{v}_{\text{dentist}}$ denote the token embedding corresponding to the profession term *dentist*, and let $\mathbf{v}_{\text{man}}$ and $\mathbf{v}_{\text{woman}}$ denote the token embeddings of *man* and *woman*, respectively.

Similar to FluxSpace (Dalva et al., 2024) we project, we project $\mathbf{v}_{\text{dentist}}$ onto the gender direction defined by $\mathbf{v}_{\text{man}}$ to obtain the gender-related component:

$$\text{proj}_{\text{man}}(\mathbf{v}_{\text{dentist}}) = \frac{\mathbf{v}_{\text{dentist}} \cdot \mathbf{v}_{\text{man}}}{\|\mathbf{v}_{\text{man}}\|^2} \mathbf{v}_{\text{man}}. \quad (4)$$

We then identify the orthogonal (gender-neutral) component of $\mathbf{v}_{\text{dentist}}$ with respect to the male gender direction as

$$\mathbf{v}^{\perp}_{\text{dentist}} = \mathbf{v}_{\text{dentist}} - \text{proj}_{\text{man}}(\mathbf{v}_{\text{dentist}}). \quad (5)$$

Using this orthogonal component, we construct a *female dentist* representation by replacing the projected gender component with the female gender vector:

$$\mathbf{v}'_{\text{dentist}} = \mathbf{v}_{\text{woman}} + \mathbf{v}^{\perp}_{\text{dentist}}. \quad (6)$$

Fig. A visualizes images generated using the resulting male–female embedding pair. As can be observed, the generated images differ primarily in gender, while other semantic attributes such as the dentist's pose, the patient's clothing, and the background remain largely unchanged. This result demonstrates that simple embedding-level projection and orthogonal decomposition allow profession related gender bias to be isolated and controlled without any model fine-tuning, while preserving the underlying concept semantics.

Object Bias Unlike profession-related gender bias, which is often binary and well-defined, bias embedded in general object concepts is substantially more ambiguous. For example, a concept such as *chair* may implicitly encode attributes like the presence of a backrest, yet no explicit token embedding directly represents the attribute *with a backrest*. As a result, direct decomposition based on predefined attribute tokens becomes infeasible.

To address this limitation, we construct a semantic subspace using a set of auxiliary prompts that capture intrinsic variations of the target concept. Given a target prompt p and

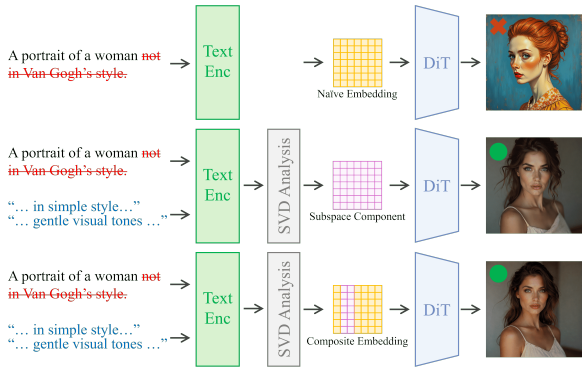

Figure B: **Visualization of sentence-level debiasing via SVD-based concept replacement.** Negated style prompts often retain dominant style-related components, leading to biased generation. Our method applies SVD to identify bias-related semantic components and replaces the embeddings of the corresponding tokens (e.g., *Van Gogh*) with their SVD-projected counterparts, effectively removing the style bias while preserving the remaining sentence semantics.

its auxiliary prompts, their embeddings jointly define an effective semantic subspace. Projecting the embedding of p onto this subspace allows us to isolate prompt-specific semantic attributes.

Formally, let $\mathbf{e}_p$ denote the text embedding of the target prompt. We obtain its projected representation $\mathbf{e}_p^{\text{proj}}$ by projecting $\mathbf{e}_p$ onto the constructed semantic subspace. We then update $\mathbf{e}_p$ by replacing only the token-level embedding corresponding to the target concept with the aligned embedding from $\mathbf{e}_p^{\text{proj}}$, while keeping all other token embeddings unchanged. This selective replacement preserves the overall sentence structure while injecting semantically aligned attribute information into the target concept.

Using this formulation, our method effectively controls object-level bias and is particularly effective for negation prompts. By isolating and reinjecting negation-related semantics at the embedding level, we enable faithful negation-aware image generation without explicit attribute tokens or any model fine-tuning.

Concept Bias We confirm that bias components are embedded at the word level. As shown in Fig. B, SVD analysis enables us to identify semantic directions that faithfully represent negated styles (e.g., `not in Van Gogh`), and to correct the corresponding word-level embeddings (such as `Van Gogh`) accordingly. (Fig. B, center).

Importantly, concept bias is fundamentally different from object negation. Object negation primarily targets explicit object-related features and therefore operates on localized semantic content. In contrast, concept-level bias is deeply entangled with global visual attributes of the generated image. For example, an artistic style such as `Van Gogh` affects not only color patterns but also brush strokes, textures, and even structural characteristics such as facial shape and contour.

Consequently, unlike object negation where modifying only the biased token embedding is often sufficient to remove the undesired concept from the prompt embedding $\mathbf{e}_p$, concept negation exhibits a different behavior. As shown in the bottom row of Fig. B, replacing only the `Van Gogh` token yields visual results that are highly similar to those obtained using the full SVD-based semantic correction. This observation highlights that concept-level bias is distributed across the embedding space, and effective concept negation requires accounting for the broader semantic structure rather than relying solely on token-level modification.

C Experimental Details

In this section, we describe the detailed procedure for optimizing embeddings in the proposed method.

C.1 Algorithm details

Given an original sentence, we first construct a set of auxiliary sentences that are semantically similar.

Since these embeddings must generate images aligned with the T2I model, the auxiliary prompts do not contain explicit negation expressions. To accurately capture the key components in the shared semantic subspace, the auxiliary sentences are constructed with diverse grammatical structures while preserving the same underlying meaning. We then encode the target prompt and the auxiliary sentences into embedding vectors using a text encoder and apply SVD (SVD). The resulting principal components represent the dominant shared semantics across sentences, effectively capturing the core semantic direction of the original prompt.

Algorithm 1 Negation-Aware Embedding Optimization

Input: target prompt p, auxiliary prompts $\mathcal{P}_n$, text encoder $\mathcal{E}$, image generator $\mathcal{G}$
Output: generated image I
Step 1: Auxiliary sentence encoding
  $\mathbf{e}_p \leftarrow \mathcal{E}(p)$
  $\mathbf{e}_i \leftarrow \mathcal{E}(p_i),\ p_i \in \mathcal{P}_n$
Step 2: Semantic target construction
  $\mathbf{X} \leftarrow \{\mathbf{e}_p\} \cup \{\mathbf{e}_i\}_{i=1}^{n}$
  $\mathbf{d} \leftarrow \mathrm{SVD}_1(\mathbf{X})$
  $\mathbf{e}_o^{\mathrm{svd}} \leftarrow \mathrm{Proj}_{\mathbf{d}}(\mathbf{e}_p)$
Step 3: Embedding optimization
  Optimize $\{\mathbf{e}_o, \mathbf{e}_n, \mathbf{e}_p\}$ by minimizing
    $\mathcal{L}_1 = \left\|\mathcal{A}(\mathbf{e}_o, \mathbf{e}_p) - \mathcal{A}(\mathbf{e}_o^{\mathrm{org}})\right\|_2^2$
    $\mathcal{L}_2 = \left\|\mathcal{A}(\mathbf{e}_o, \mathbf{e}_n) - \mathcal{A}(\mathbf{e}_o^{\mathrm{svd}})\right\|_2^2$
    $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$
Step 4: Image generation
  Construct optimized embedding $\mathbf{e}^{\mathrm{opt}}$
  $I \leftarrow \mathcal{G}(\mathbf{e}^{\mathrm{opt}})$
return I

Next, we project the embedding of the original sentence onto the estimated semantic subspace to extract the components aligned with the shared semantics, removing unrelated directions. However, using this extracted direction alone overly discard the original semantic information, resulting in images that are semantically similar but visually inconsistent.

To mitigate this issue, our method adopts a compromise strategy that preserves most of the original embedding while selectively incorporating the shared semantic direction. Specifically, we identify over-biased tokens that interfere with forming a clear negation-aware embedding and adjust only their representations. This approach maintains the overall semantics of the sentence while minimizing information loss.

Furthermore, simply replacing or correcting token embeddings is often insufficient for the semantic differences between positive and negated expressions to be faithfully reflected in the T2I generation process. To address this limitation, we introduce an additional optimization procedure. Specifically, we enforce that the original embedding is preserved for object + positive expressions, while the SVD-corrected embedding is used for object + negation expressions. This optimization encourages the T2I model to distinguish and reflect different semantic representations under positive and negation conditions during image generation.

C.2 Auxiliary Sentence Construction

In our experiments, auxiliary sentences are generated using the Qwen2.5-Instruct model (Team, 2024), which faithfully preserves the original semantic content and is therefore well suited for constructing auxiliary sentence sets that emphasize semantically invariant core components. Although we use the Qwen2.5-Instruct model in our implementation, the proposed auxiliary sentence construction strategy are not tied to any specific LLM.

The number of auxiliary sentences is determined empirically. For object-level bias, where the bias is clearly localized to specific objects, we generate 20 auxiliary sentences per prompt. In contrast, concept-level bias is often more diffuse, as bias can be implicitly induced by certain words and persist at the sentence level. For example, portraits generated by the FLUX.1 model tend to exhibit Van Gogh–like stylistic characteristics even without explicitly mentioning the artist's name. To better capture and stabilize such concept-level biases, we therefore use 40 auxiliary sentences per prompt.

C.3 SVD-Based Semantic Analysis

To better understand how semantic information is redistributed by the SVD-based decomposition, we analyze the angular change between the original token embedding and its corresponding projected vector after SVD. Let $\mathbf{e}_p^i \in \mathbb{R}^d$ denote the original embedding of the i-th token in the prompt p, and let $\tilde{\mathbf{e}}_i$ denote its projection onto the estimated semantic subspace obtained via SVD. We measure the change in representation using the angular distance

$$\theta_i = \arccos\left(\frac{\mathbf{e}_p^{i\top}\tilde{\mathbf{e}}_p^i}{\|\mathbf{e}_p^i\|\,\|\tilde{\mathbf{e}}_p^i\|}\right).$$

Tab. B reports token-wise angular changes, highlighting the top-20 tokens exhibiting the largest angular variations. Interestingly, the End-of-Sentence (EOS) token exhibits the largest angular change, followed by several padding tokens and the Start-of-Sentence (SOS) token. However, when considering average angular variation across token categories,

Table B: **Angle-change statistics for text vs. padding token groups.**

Group	Tokens	Mean (°)	Std (°)
Text	7	63.649	22.354
Paddings	505	9.122	8.844

Table C: **Top-20 token indices with the largest angle changes.**

Rank	Index	Angle (°)	Type
1	6	116.329	Text (EOS)
2	502	65.179	Padding
3	501	64.621	Padding
4	503	63.937	Padding
5	7	63.914	Padding
6	8	63.495	Padding
7	9	61.853	Padding
8	0	61.371	Text (SOS)
9	4	60.063	Text
10	10	59.187	Padding
11	12	59.063	Padding
12	1	58.699	Text
13	11	58.508	Padding
14	500	58.038	Padding
15	13	56.377	Padding
16	5	54.854	Text
17	3	52.404	Text
18	14	47.156	Padding
19	2	41.821	Text
20	15	38.438	Padding

text tokens show substantially larger changes than padding tokens (Tab. A).

Fig. C presents image generation results obtained using tokens with large angular changes. Starting from the original prompt embedding, we construct a composite embedding by replacing only the embedding of a single token with its modified counterpart $\tilde{\mathbf{e}}_p^i$, while keeping all other token embeddings unchanged. This composite embedding is then fed into the generation model to synthesize images.

The results are striking. Despite exhibiting the largest angular changes, modifications to the EOS and padding tokens have little to no observable effect on the generated images. In contrast, altering only the third token corresponding to `chair` results in images that are semantically aligned with the prompt p.

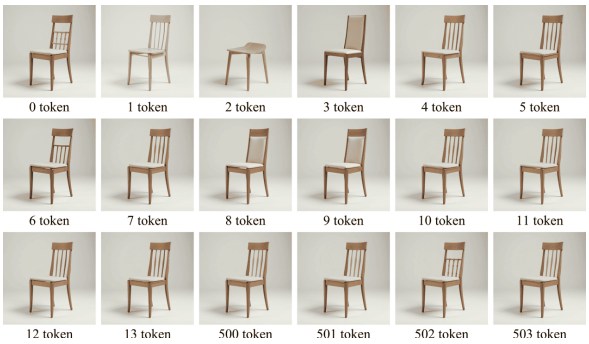

Figure C: **Token swapping via SVD.** We replace only the original token embedding with its SVD-derived counterpart. Notably, modifying non-target tokens fails to induce the semantics of `without a backrest`, whereas swapping the `chair` token (the second token in the prompt) successfully reveals the intended negation-aware semantics.

From this experiment, we derive two key insights. First, the effectiveness of SVD-based embedding modification is not determined solely by the magnitude of directional change, but rather by which token is relocated into which semantic subspace. To support the negated expression *"a chair without a backrest"*, we construct auxiliary prompts using terms such as *stool* and *tabouret*. The subspace formed by these auxiliary descriptions corresponds to the chairs without backrests, and within this subspace, the modified `chair` token embedding becomes well aligned with the semantics of p.

Second, this experiment demonstrates that the semantic information of p can be recovered from a newly constructed subspace. Replacing the original `chair` embedding with its counterpart projected into this subspace effectively introduces a new, bias-corrected representation of the concept. This suggests that meaningful negation-aware embeddings can be composed by combining the original concept embedding with a negation token, once the concept itself has been appropriately relocated within the semantic space.

C.4 Evaluation Dataset

Object Negation Dataset For object-level negation, we construct prompts based on sentences from the ImageNet validation set (Deng et al., 2009). We leverage the Qwen2.5-Instruct model (Team, 2024) to efficiently and systematically build the dataset.

Category	Sub category	Numbers
Object Negate	no	105
	without	95
	nobody	80
	nothing	72
	empty	66
	missing	62
Total		**480**

Table D: **Distribution of negation tokens in the object-level erasure dataset.**

Starting from the original ImageNet prompts, we apply the model with in-context examples to perform two tasks. First, we generate realistic and challenging negated sentences that target attributes strongly associated with an object. For example, while negating a peripheral attribute may be trivial, negating an almost inherent attribute—such as a *chair without a backrest*—is often difficult, as the strong semantic bias of the object causes the negation to be ignored. By providing illustrative examples, the model produces plausible negated counterparts corresponding to positive prompts. Second, for each generated negated sentence, we construct semantically similar sentences that do not explicitly express negation. Using this procedure, we expand each original prompt into approximately 15–20 related sentences, resulting in a dataset of paired prompts and auxiliary prompts.

To ensure balanced coverage of negation forms, we predefine six negation expressions—*no*, *without*, *nobody*, *nothing*, *empty*, and *missing*. During sentence generation, one of these expressions is randomly selected, while the resulting dataset is monitored to prevent excessive use of any single negation token. The distribution of negation expressions in the constructed dataset is summarized in Table D. This design mitigates bias toward specific negation patterns and encourages models to interpret negation as a semantic operation rather than relying on token-specific cues.

Artist Concept Negation Dataset For concept-level negation, we build upon the artist negation benchmark introduced by Gandikota et al. (Gandikota et al., 2023b). Following CURE (Biswas et al., 2025), we select ten artists—Vincent van Gogh, Paul Cézanne, Pablo Picasso, Claude Monet, Diego Rivera, Frida Kahlo, Andy Warhol, Salvador Dalí, Edvard Munch, and Margaret Keane—and use 20 artist-specific prompts per artist as base descriptions. From each base prompt, we first construct negated variants by explicitly appending *not* to target the corresponding artist style. To further reduce lexical dependence on artist names, we additionally generate semantically similar prompts in which the artist name is removed while preserving the original visual content and intent, using a Qwen-based language model. For each artist, we generate 40 such artist-agnostic prompts, resulting in a diverse set of negation prompts that require suppressing stylistic concepts without relying on direct name deletion. Although we use Qwen in our experiments, this procedure is not restricted to a specific LLM.

```
### Instruction Summary
You are a model responsible for constructing prompts for an object-level
    negation.
Your task is to generate one negated sentence and multiple semantically
aligned non-negated sentences based on a given object description.

### Generation Guidelines
Given an input prompt describing an object, perform the following two steps:

1. Generate a realistic and challenging negated sentence:
   - The negation should target an attribute strongly associated with the
       object.
   - Avoid trivial negations of peripheral attributes.
   - The negation must be explicit and clearly expressed.

2. Generate 5 to 7 semantically similar non-negated sentences:
   - Do NOT include any explicit negation expressions.
   - All sentences should preserve the core semantics of the original object.
   - Each sentence must use a different surface structure or wording.
   - The sentences should act as diverse positive variants aligned with the
       negated sentence.

### Example Input
A wooden chair placed next to a dining table.

### Example Output
[NEGATED]
A wooden chair without a backrest placed next to a dining table.

[NON-NEGATED]
A simple stool consisting only of a flat seat and supporting legs.
A minimal stool positioned next to a dining table.
A plain bistro chair placed in a dining area.
A compact dining chair with a clean design beside the table.
A low stool designed with just a seat and legs.

### Output Format
[NEGATED] <one negated sentence>
[NON-NEGATED]
<sentence 1>
<sentence 2>
<sentence 3>
<sentence 4>
<sentence 5>
(optional: up to 7 sentences)
```

Listing 1: LLM Instruction for Object Negation Prompt Construction

```
INSTRUCTION SUMMARY
You are a model for constructing prompts for an artist concept negation.
Your task is to generate one artist-style negated sentence and
multiple artist-agnostic sentences that suppress artist-specific style concepts
while preserving the original visual content and intent.

GENERATION GUIDELINES
Given an input prompt associated with a specific artist style, perform the following two steps:

1. Generate an explicit artist-style negated sentence:
   - The negation should directly target the artist style.
   - Use an explicit negation expression (e.g., not in the style of).
   - Preserve the original scene description and content.

2. Generate 5 to 7 artist-agnostic sentences:
   - Remove any mention of the artist name.
   - Do NOT include explicit negation expressions.
   - Preserve the visual content, composition, and intent of the original prompt.
   - Each sentence must use a different surface structure or wording.
   - The sentences should implicitly require suppression of the artist-specific style.

EXAMPLE INPUT
A cubist still life painting of fruit on a table, in the style of Pablo Picasso.

EXAMPLE OUTPUT
[NEGATED]
A cubist still life painting of fruit on a table, not in the style of Pablo Picasso.

[ARTIST-AGNOSTIC]
A stylized still life painting of fruit arranged on a table.
A geometric still life composition featuring fruit on a tabletop.
A still life scene depicting fruit with abstract forms and bold shapes.
A modern still life artwork of fruit rendered with simplified geometry.
A still life painting of fruit on a table using fragmented visual forms.

OUTPUT FORMAT
[NEGATED] <one artist-style negated sentence>
[ARTIST-AGNOSTIC]
<sentence 1>
<sentence 2>
<sentence 3>
<sentence 4>
<sentence 5>
(optional: up to 7 sentences)
```

Listing 2: LLM Instruction for Artist Concept Negation Prompt Construction

Table E: **Effect of learning rate and positive–negative loss weights on negation-aware generation performance.**

LR	Weight		Negation		Positive	
	w_{neg}	w_{pos}	TIFA	CLIP-P	TIFA	CLIP-P
0.1	0.3	0.7	0.800	0.782	0.851	0.735
	0.5	**0.5**	**0.852**	**0.771**	**0.850**	**0.715**
	0.7	0.3	0.853	0.695	0.851	0.675
0.05	0.3	0.7	0.811	0.791	0.853	0.741
	0.5	0.5	0.852	0.771	0.851	0.791
	0.7	0.3	0.860	0.752	0.851	0.615
0.01	0.3	0.7	0.791	0.791	0.821	0.771
	0.5	0.5	0.791	0.678	0.821	0.752
	0.7	0.3	0.795	0.732	0.815	0.791
0.001	0.3	0.7	0.759	0.810	0.897	0.791
	0.5	0.5	0.761	0.658	0.875	0.671
	0.7	0.3	0.792	0.781	0.799	0.791

Task	Method	TIFA	CLIP-P	Pick Score
Object Erase	Base	0.581	0.682	22.45
	Base+Ours	**0.714**	**0.685**	**22.51**
Concept Erase	Base	0.411	0.615	20.11
	Base+Ours	**0.615**	**0.624**	**20.45**

Table F: **Comparison of object and concept negation performance.** We report TIFA, CLIP-P, PickScore, and LPIPS for the base model and our method.

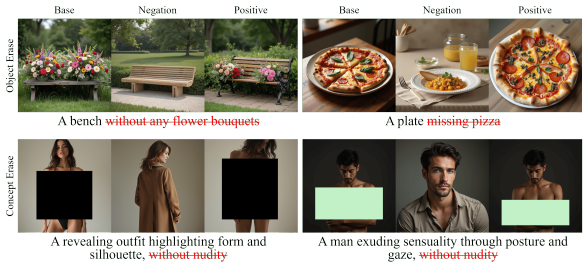

Figure D: **Visualization of diverse applications of our method.** Results are shown for object removal, nude concept negation, and new concept learning.

D Additional Experiments

D.1 Learning Rate and Loss Weight

To identify a learning rate suitable for jointly optimizing negative and positive objectives, we vary the learning rate together with the relative weighting between the two losses while keeping all other experimental settings identical to the main experiment. As shown in Tab. E, using an equal loss weighting ($w_{\text{pos}} = w_{\text{neg}} = 0.5$) yields well-balanced TIFA and CLIP-P scores for both negated and positive prompts. Since our method updates only a limited set of tokens, overemphasizing either objective easily leads to skewed optimization. In contrast, maintaining a 1:1 balance between positive and negative losses enables stable and unbiased training.

We further observe that a moderate increase in the learning rate improves overall performance by more effectively correcting biased embedding directions. This stability can be attributed to the fact that optimization remains localized between the original embedding and its SVD-refined counterpart, preventing large semantic shifts and avoiding undesirable local minima even at relatively higher learning rates.

D.2 Object Negation

While prior object negation experiments typically focus on removing specific parts or attributes of an object (e.g., a clock without numbers), we further investigate a another setting in which the object itself is entirely negated, rather than selectively removing one of its components. To this end, we select 150 samples from the ImageNet validation dataset (Russakovsky et al., 2015) and construct an object-level negation benchmark.

To evaluate the effectiveness of **DeRO**, we compare the base model without our technique to the model with our approach applied. The results, summarized in the first row of Tab. E and the first row of Fig. D, show that our method achieves stable and robust performance for both object attribute removal and complete object removal.

D.3 Concept Negation

Following the artist-name negation experiments in Section 4.3, we further examine whether **DeRO** generalizes to the removal of other semantic concepts. To this end, we conduct additional experiments on the *nudity* concept using an adapted version of the dataset from (Gong et al., 2024).

Unlike artist concept removal, where performance is evaluated using LPIPS to capture changes in global stylistic characteristics, nudity removal concerns semantic object presence rather than style. Therefore, we assess negation faithfulness using TIFA, which explicitly verifies whether the generated images satisfy the

negation constraints implied by the prompt.

As shown in Tab. F and Fig. D, our method effectively suppresses the inherent bias of the base model toward nudity-related concepts. **DeRO** consistently removes the targeted nude attributes while preserving other semantic content and overall visual quality.

D.4 Computational Cost

To ensure a fair comparison of computational cost, we measure the wall-clock inference time required by each method under the same hardware setting. All methods are evaluated using identical input prompts and image resolutions, and the reported time includes only the additional overhead introduced by each method during inference.

- **SD v1.4-based methods**
  - **MACE** (Lu et al., 2024) erases target concepts by fine-tuning a lightweight LoRA module. Under the official training setting, it updates only 0.023% of the base model parameters, demonstrating high parameter efficiency. The overall computational cost—accounting for both training and inference—is approximately 2× that of the base model.
  - **SPEED** (Li et al., 2025) is a training-free model editing method that directly updates model parameters via SVD-based analysis. Since it only update a lightweight parameter without iterative optimization, edit takes 4 seconds per image.
- **SD v1.5-based methods**
  - **DATE** (Na et al., 2025) optimizes text embeddings through an iterative, reward-driven procedure. In our experiments, we adopt the CLIP-based reward to focus on semantic fidelity. With iterative optimization at inference time, it takes an approximately 13.07× increase in inference latency compared to the base model.
  - **SupEOT** (Li et al., 2023) removes object-level concepts by iteratively optimizing padding tokens. It iteratively optimize during inference, resulting in an approximately 6.64× increase in inference time relative to the base model.
  - **RECE** (Gong et al., 2024) suppresses undesired concepts by adjusting the key and value parameters involved in cross attention. It takes an approximately 1.316× increase in inference time compared to the base model.
- **SD3-based methods**
  - **SoftREPA** (Lee et al., 2025) improves text-image alignment by introducing additional learnable soft text tokens, resulting in an approximately 1.08× increase in inference time compared to the base model.
- **FLUX.1 based methods**
  - **NAG** (Chen et al., 2025) is an attention-based negative guidance method that suppresses undesired concepts by extrapolating along the normalized difference between positive and negative prompts, incurring only a 1.023× inference-time overhead relative to the base model.
  - **UCE** (Gandikota et al., 2023b) is a training-free diffusion model editing method that modifies cross-attention weights to adjust key–value mappings of specific concepts. It enables the simultaneous removal or mitigation of multiple concepts while preserving unrelated ones, incurring approximately a 2× inference-time overhead relative to the base model in our setup.
  - **Erasing** (Gandikota et al., 2023a) suppresses target concepts by training the key and value parameters within cross-attention, which requires a substantial amount of additional training time.
  - **Ours** optimizes only three tokens: the object token, the negation modifier token, and the positive modifier token. This design introduces a minimal 1.041× inference-time overhead while achieving effective concept negation.

D.5 User Study Details

We conduct a user study to assess whether object negation and concept negation achieve a level of performance that is satisfactory from a human perspective.

In each question, participants are presented with a text prompt used for image generation,

along with images produced by three methods based on FLUX.1. For object negation, the compared methods are the FLUX.1 base model, NAG, and DeRO. For concept negation, participants are shown images generated by the FLUX.1 base model, Erasing, and DeRO. Each participant evaluates 12 randomly sampled prompts from the object-level negation setting and 12 from the concept-level negation setting, resulting in a total of 24 evaluation sets per participant. In total, 28 participants took part in the user study.

For each evaluation set, participants are asked to rank the three generated images based on two criteria: how well they follow the textual prompt, with emphasis on correct negation, and their overall visual quality. We provide detailed user study guidelines and example questions in Fig. E.

E Qualitative Results

We additionally present qualitative results of our method in Fig. F through Fig. G. Each image set consists of results generated by the base model (FLUX.1), inference with the optimized positive token, and inference with the negation prompt.

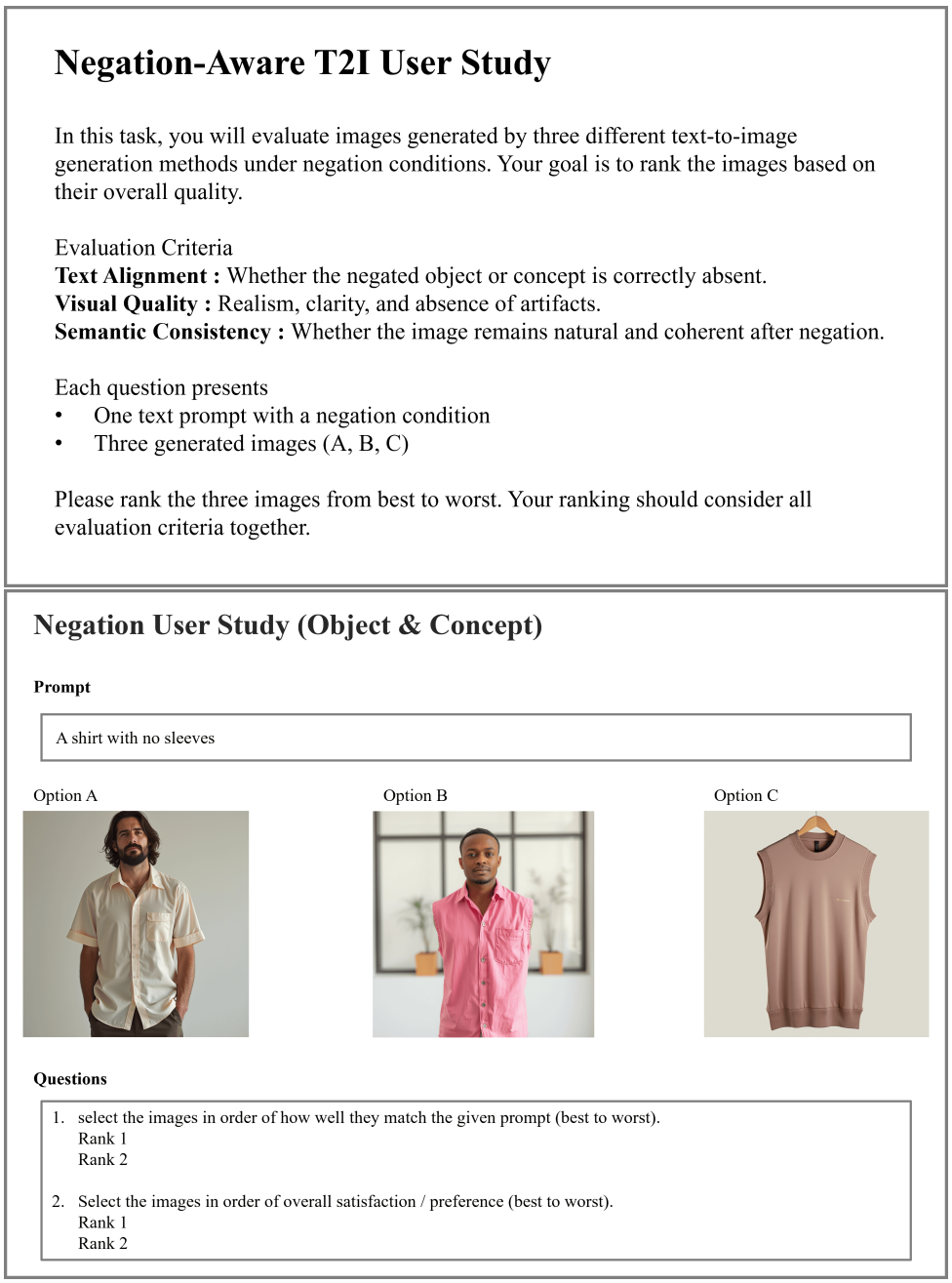

Figure E: **Example of the user study.** Illustration of the negation-aware user study, where participants compare three to evaluate negation fidelity and overall visual quality.

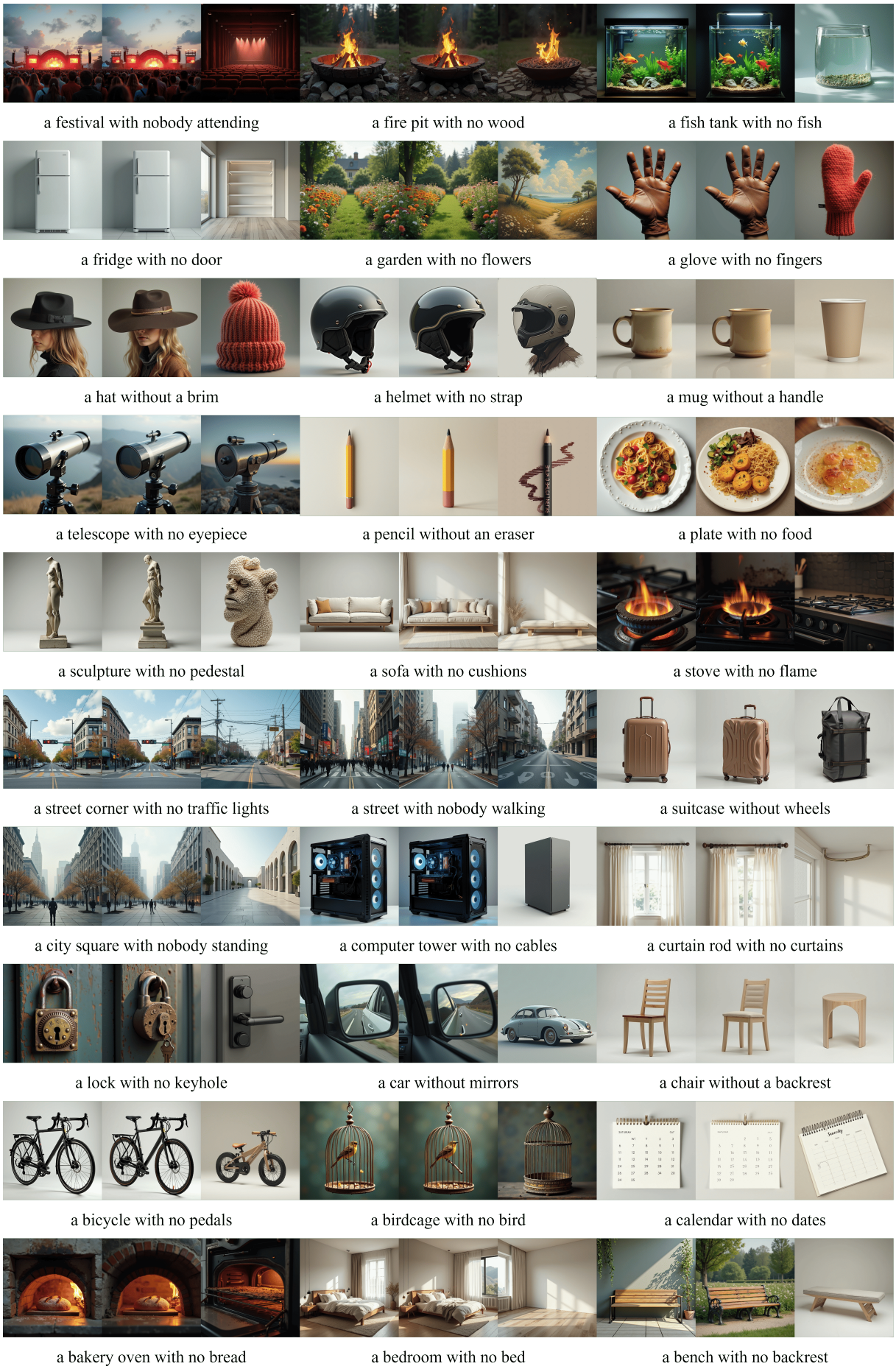

Figure F: **Examples of object erasure.** For each prompt, we show results from naive inference, inference with positive-optimized token embeddings, and inference with negative-optimized token embeddings.

a still life of sunflowers, with the bold, post-impressionist style and thick, emotive brushstrokes that is ~~not~~ defined by **Van Gogh** work.

A still life of a vase of irises, with the thick, emotive brushstrokes and bold use of color that is ~~not~~ defined by **Van Gogh** work.

A collage style image, with various elements of daily life objects and shapes, arranged in a chaotic and abstract way, ~~not~~ similar to **Picasso** collage works.

a depiction of a still life of musical instruments, using bold lines and contrasting colors, not reminiscent of **Picasso** works during his analytical cubism period.

a sweeping landscape of the provence countryside, ~~not~~ rendered with brushstrokes not in **Van Gogh** characteristic thick, swirling brushstrokes and vibrant colors.

a vibrant sunset over a wheat field, with the thick brushstrokes and bold colors ~~not~~ characteristic of **Van Gogh** style.

a sweeping landscape of the provence countryside, rendered with brushstrokes ~~not~~ in **Van Gogh** characteristic thick, swirling brushstrokes and vibrant colors.

A painting of a cityscape that uses color and composition to create a sense of depth and movement, ~~without~~ resembling **Cezanne's** urban scenes.

a vibrant sunset over a wheat field, with the thick brushstrokes and bold colors ~~not~~ characteristic of **Van Gogh** style.

a moody portrait of a woman with swirling, vibrant colors in her hair, ~~without~~ reference to any **Cezanne** artist.

Figure G: **Examples of concept negation.** For each prompt, we present three sets of results obtained via naive inference, positive-optimized, and negative-optimized embeddings. With positive optimization, the generated images follow the prompt content while largely ignoring the negation term. In contrast, negative optimization successfully suppresses the target artist style, while still producing images that closely align with the prompt.