Negation without Confusion: Decomposing Bias and Recomposing Style in Text Embeddings for Negation-aware Image Generation

Sooyeon Park[♠] **Hemraj Singh**[♠] **Kyunghyun Min**[♠] **Youngjae Yu**[◇] **Sung-Bae Cho**[♠]

[♠] Yonsei University [◇] Seoul National University

{dreamyou070, shemraj07, khmin1104, sbcho}@yonsei.ac.kr

youngjaeyu@snu.ac.kr

Abstract

Negation-aware image generation remains challenging for text-to-image diffusion models, which often struggle with negation due to implicit linguistic biases that are deeply entangled within text representations. As a result, bias-aligned semantics frequently dominate over intended negative constraints. To overcome this limitation, we propose a simple training-free method of **De**compose and **R**ec**O**mpose text embedding (DeRO), which constructs a semantic-preserving subspace and a bias subspace using auxiliary prompts and performs semantic disentanglement based on bidirectional SVD. By decomposing entangled representations and selectively recomposing disentangled semantic components, DeRO realigns text representations toward intended negation semantics while restoring residual stylistic features to preserve semantic fidelity and image quality. As a model-agnostic inference-time framework operating solely on text representations, DeRO can be seamlessly applied to diverse diffusion architectures, including both U-Net and DiT based models. Extensive experiments on both object- and concept-level negation demonstrate consistent improvements in negation-aware image generation, achieving stronger text-image alignment, better semantic fidelity, and higher human preference scores.

1 Introduction

Recent Diffusion Transformer (DiT) based text-to-image (T2I) models have achieved remarkable progress in visual quality and semantic alignment capabilities (Labs, 2024; Podell et al., 2024), but they still struggle to reliably handle negation prompts (Li et al., 2024; Chen et al., 2025). For example, even for a prompt of "a bakery without bread," the undesired concept "bread" is often still generated as shown in Fig. 2.

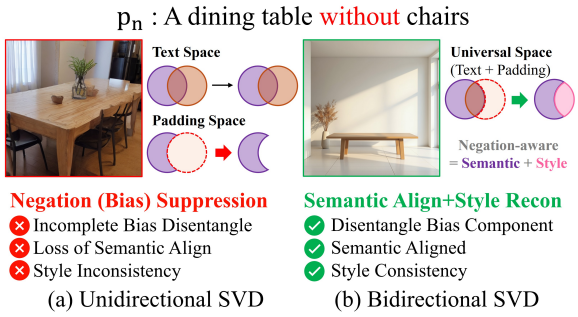


Figure 1: (a) Existing **unidirectional SVD** methods (e.g., SupEOT (Li et al., 2024)) suppress only bias or negation directions within padding representations in order to remove undesired information while preserving the original prompt style. However, due to the strong entanglement among semantic, stylistic, and bias components, residual biased semantics often remain in the text embedding, resulting in incomplete negation or degraded stylistic consistency. (b) In contrast, **bidirectional SVD** operates in a universal embedding, enabling precise removal of bias components while preserving semantic and stylistic information, thereby achieving faithful negation-aware generation.

A negation prompt inherently contains two coupled semantic components: contextual semantics to be *preserved* and target attributes to be *removed*. The difficulty of negation largely depends on the semantic bias between these components. When weakly correlated (e.g., *room* and *bread*), the negated concept is reliably removed; however, strongly coupled pairs, such as *bakery* and *bread*, often persist in generated images. This mainly arises from implicit linguistic biases formed during large-scale training (Geirhos et al., 2020; Park and Cho, 2026), which strongly entangle certain semantics within text representations (Fu et al., 2025; Friedrich et al., 2023). Therefore, negation in T2I models is fundamentally a problem of disentangling preserved and removed semantics rather than simply suppressing target concepts.

Existing negation-aware generation methods

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 50492–50517
October 24-29, 2026 ©2026 Association for Computational Linguistics

mainly focus on suppressing unwanted semantics through attention manipulation (Chen et al., 2025) or direct embedding suppression (Li et al., 2024). However, as illustrated in Fig. 1(a), suppression-only methods struggle under strong semantic coupling (Saha et al., 2025), consequently suffering from semantic collapse, style distortion, or incomplete negation.

To address these limitations, we propose DeRO (**De**compose and **R**ec**O**mpose text embedding), a simple yet effective method that treats negation failure as representation-level bias entanglement. Unlike suppression based methods(Li et al., 2024; Biswas et al., 2025), DeRO constructs biased and semantic auxiliary prompt sets (A_B and A_S) and performs SVD based bidirectional disentanglement. Given "a bakery without bread," the method selectively removes only bread-related semantics entangled within the bakery representation. Through semantic recomposition and residual style restoration before denoising, DeRO suppresses undesired semantics while preserving contextual structure and visual style. This representation-level intervention is supported by theoretical analysis, and fundamentally differs from methods relying on attention manipulation (Chen et al., 2025; Biswas et al., 2025) or iterative sampling optimization (Na et al., 2025).

Because the proposed method requires only a one-time auxiliary prompt construction stage and operates on low-dimensional embedding subspaces, DeRO introduces negligible additional computational overhead during inference. Furthermore, representation-level semantic decomposition naturally generalizes from object-level negation to more abstract concept-level negation scenarios. Without requiring parameter updates or additional fine-tuning, DeRO is model-agnostic and can be seamlessly integrated into diverse diffusion architectures, as demonstrated in both U-Net (Rombach et al., 2022) and DiT based models (Labs, 2024; Podell et al., 2024).

The main contributions of this paper are as follows:

- A simple yet effective training-free negation-aware generation method is proposed from the perspective of representation-level semantic disentanglement, which is model-agnostic and applicable to diverse diffusion architectures, including both U-Net and DiT based models.
- An auxiliary prompt-driven bidirectional SVD strategy is introduced to disentangle bias-related semantics from text representations, while preserving visual style through semantic recomposition and residual restoration, supported by formal theoretical analysis.
- Extensive experiments on object- and concept-level negation benchmarks demonstrate consistent improvements over existing baselines, achieving up to 27.4% improvement in TIFA score and 2.7% gain in PickScore, while consistently improving negation understanding in both conventional U-Net and recent DiT based diffusion architectures.

2 Related Work

Text-Image Alignment. T2I research has primarily focused on improving text-image alignment through techniques such as reward optimization (Na et al., 2025), representation editing (Arad et al., 2024), and attention strength modulation (Park et al., 2026). While recent approaches such as DATE (Na et al., 2025) and Soft-REPA (Lee et al., 2025) further improve alignment, they primarily target general text-image correspondence rather than the representation-level semantic entanglement inherent in negation prompts. More recently, negation-specific methods have emerged. SupEOT (Li et al., 2024) performs explicit object suppression through U-Net-based interventions, while NAG (Chen et al., 2025) employs attention extrapolation in transformer-based architectures. However, these methods rely on architecture-specific mechanisms, limiting their applicability in heterogeneous diffusion backbones.

Semantic Erasure. Concept erasure in diffusion models typically involves model editing (Lu et al., 2024; Gandikota et al., 2023; Biswas et al., 2025) or text editing (Li et al., 2024; Han et al., 2025). While model editing often irreversibly alters model parameters and may affect generalization, text editing primarily performs global concept suppression but struggles with conditional negation, which requires context-aware selective removal. Although recent methods attempt to mitigate unintended side effects by localizing the erasure, such as leveraging graph-based spaces (Bui et al., 2025) or null-space editing (Li et al., 2026), they still fundamentally rely on direct suppression of semantic directions.

3 Proposed Method

3.1 Motivation

Recent diffusion models (Podell et al., 2024; Labs, 2024) have substantially improved text understanding and text-image alignment capabilities. Furthermore, recent architectures increasingly leverage multiple text encoders, such as CLIP (Radford et al., 2021) and T5 (Raffel et al., 2020), to jointly capture global semantic context and fine-grained textual details. Despite these improvements, negation prompts remain challenging because removing a target concept requires suppressing not only explicit textual information but also implicit semantic priors learned from large-scale training data.

Fig. 2(a) illustrates this phenomenon using the prompt "a bakery with no bread." Although the prompt explicitly specifies the absence of bread, the generated image still contains bread-related content. This failure suggests that the concept of a bakery is strongly associated with bread through implicit data bias, causing bias-aligned semantics to dominate over the intended negation constraint(Kang et al., 2025; Orgad et al., 2023). In contrast, as shown in Fig. 2(b), prompts with weaker semantic associations naturally preserve the desired semantics while successfully removing the target concept.

A straightforward solution is to project text embeddings into a semantic-preserving subspace using SVD based disentanglement. However, as shown in Fig. 2(c), uni-directional projection often removes not only bias-related representations but also contextual semantics and visual style information. While the undesirable concept is partially suppressed, important characteristics such as scene identity and visual appearance can also be lost.

To address this limitation, we propose a bidirectional SVD formulation that explicitly separates bias-aligned semantics from semantic-preserving representations while additionally restoring residual stylistic information. As shown in Fig. 2(d), the proposed formulation removes implicit bias while preserving both semantic consistency and visual style, leading to faithful negation-aware generation.

3.2 Negation-aware Text Embedding

Problem Setup. Fig. 3 illustrates the overall pipeline. A key challenge in negation-aware generation is that negated semantics cannot be represented as an independent embedding, since target semantics are strongly entangled with contextual and stylistic semantics within pretrained text representations. For example, in "bakery with no bread," bread-related semantics remain deeply coupled with bakery-related contextual semantics due to pretrained co-occurrence priors (more details are in Appendix B). Consequently, direct suppression often removes useful contextual information or leaves residual target semantics in generated images.

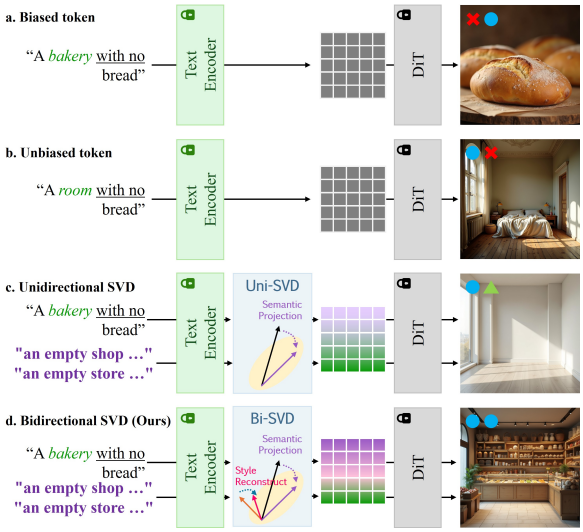


Figure 2: Comparison of semantic disentanglement in negation. (a) Strongly coupled: negation fails, (b) Weakly coupled: style lost, (c) Uni-SVD: style degraded, and (d) Bi-SVD (Ours): negation succeeds, style preserved. For each result, the left mark indicates whether the negation constraint is satisfied, and the right mark indicates whether the original contextual style (e.g., bakery interior) is preserved.

To address this issue, given a negation prompt p_n, two auxiliary prompt sets are constructed to estimate disentangled semantic structures. The **semantic auxiliary prompt set** A_S captures information that should be preserved under negation constraints, whereas the **biased auxiliary prompt set** A_B models dominant semantics aligned with the negated target attribute. Each prompt set is transformed into an embedding matrix, from which a semantic-preserving subspace $\mathcal{S}$ and a bias subspace $\mathcal{B}$ are estimated via SVD. The original text embedding x_n, obtained from the negation prompt p_n, is then projected onto $\mathcal{S}$ while suppressing bias-aligned components within $\mathcal{B}$. Residual reconstruction further restores contextual and stylistic information, and the recomposed embed-

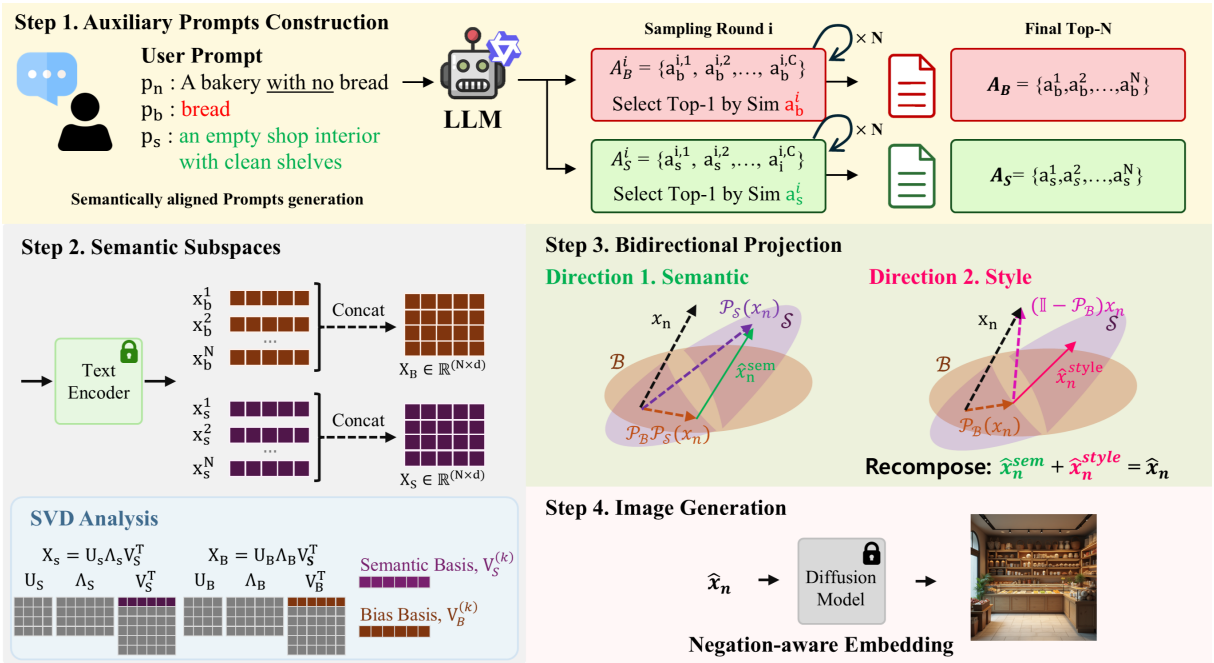


Figure 3: Overview of DeRO. (Step 1) Biased and semantic auxiliary prompts are iteratively generated via an LLM and ranked by semantic similarity. (Step 2) Text embeddings of auxiliary prompts are encoded and decomposed via SVD to estimate the semantic basis $V_S^{(k)}$ and bias basis $V_B^{(k)}$. (Step 3) Bidirectional projection constructs $\hat{x}_n = \hat{x}_n^{\text{sem}} + \hat{x}_n^{\text{style}}$ via two directions. Direction 1 removes bias after semantic alignment, and Direction 2 restores style after bias removal. (Step 4) The recomposed negation-aware embedding $\hat{x}_n$ is applied for image generation.

ding $\hat{x}_n$ is injected into a frozen diffusion model for negation-aware image generation.

Field	Content
Prompt	a shoe without laces
Biased prompt	shoelaces
Semantic prompt	a slip-on shoe without stripes
Biased auxiliary prompts	
– a sneaker with clearly tied shoelaces	
– a lace-up shoe with threaded eyelets	
Semantic auxiliary prompts	
– a slip-on shoe with a clean seamless upper	
– a minimalist loafer with no fastening elements	

Table 1: Examples of auxiliary prompt.

Auxiliary Prompt Construction. In object-level negation, it is difficult to explicitly define a semantic state where the target attribute is completely removed. For example, in "a bakery with no bread", bread-related semantics remain strongly entangled with the bakery representation, making direct construction of a bread-free embedding highly challenging (See § B.1). Moreover, semantically similar directions may still preserve residual bias even after partial suppression (Yin et al., 2019; Huang et al., 2025). Therefore, instead of relying solely on a single negation prompt, we construct auxiliary prompt sets to estimate stable bias-reduced semantic structures.

Specifically, we construct a biased auxiliary prompt set $A_B = \{a_b^i\}_{i=1}^N$ and a semantic auxiliary prompt set $A_S = \{a_s^i\}_{i=1}^N$ to refine the representation of the negation prompt p_n. To guide this process, we define the biased prompt p_b and the semantic prompt p_s. While p_b captures the target attribute to be removed (e.g., bread) along with its deeply entangled bias components, p_s preserves the desired contextual meaning without the target (e.g., an empty bakery interior). To ensure diversity and stability, the auxiliary prompts are iteratively generated using an LLM (e.g., Qwen2.5-Instruct) conditioned on p_b and p_s, respectively. In each of the N rounds, the LLM generates C candidates, and the one maximizing semantic similarity to its corresponding target concept is selected. Detailed LLM instructions and specific examples of these prompts can be found in Appendix D.1 and Table 1.

Semantic Subspaces. Each prompt in A_B and A_S, along with the original negation prompt p_n, is transformed into the embedding space using a frozen text encoder, yielding the biased embed-

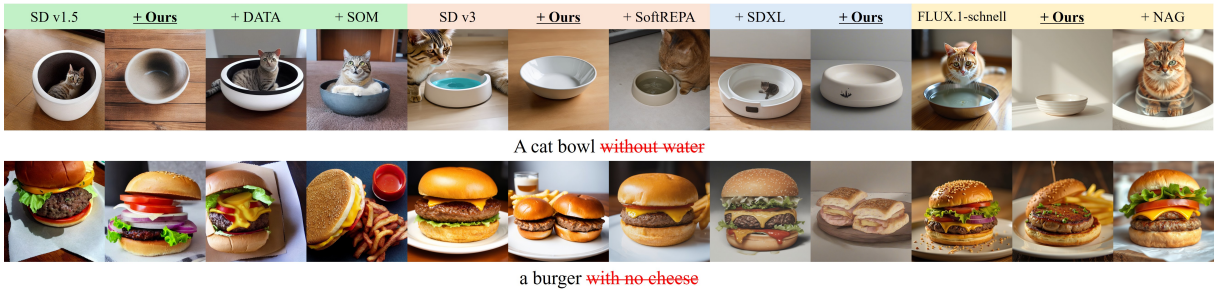


Figure 4: Comparison of object negation in different diffusion backbones. Existing methods often fail to completely remove the negated concept or degrade visual style, whereas the proposed method achieves faithful negation while preserving semantic structure and visual fidelity. Red strike-through text indicates the negated concepts.

ding matrix $X_B \in \mathbb{R}^{N\times d}$ and the semantic embedding matrix $X_S \in \mathbb{R}^{N\times d}$, where d denotes the embedding dimension. To extract dominant semantic directions, we apply SVD:

$$X_S = U_S \Sigma_S V_S^\top, \qquad X_B = U_B \Sigma_B V_B^\top, \quad (1)$$

where the column vectors of V_S and V_B represent orthogonal bases of the semantic-preserving subspace and bias subspace, respectively, while Σ_S and Σ_B contain singular values indicating the importance of each semantic direction. Here, k denotes the number of dominant semantic components used to construct the subspace. In practice, we empirically found that using only the first singular component ($k = 1$), corresponding to the largest singular value, consistently produced the best performance across different settings. Consequently, the semantic-preserving subspace $\mathcal{S}$ and bias subspace $\mathcal{B}$ are defined by $V_S^{(k)}$ and $V_B^{(k)}$, respectively. Here, $\mathcal{S}$ represents the semantic direction that preserves the desired semantic structure after removing the target attribute, whereas $\mathcal{B}$ captures the implicit semantic bias direction strongly entangled with the target attribute.

Bidirectional Projection. Decomposing text embeddings requires careful structural manipulation, as implicit biases and contextual features are tightly coupled. Consequently, a naive one-directional projection inevitably creates a significant trade-off between bias mitigation and contextual fidelity. To overcome this limitation, we propose a bidirectional projection mechanism that simultaneously enhances semantic alignment while suppressing residual bias components.

Specifically, the original negation embedding x_n is first projected onto the semantic-preserving subspace $\mathcal{S}$ to extract the desired semantic representation, and the residual bias component is subsequently removed to obtain the bias-reduced semantic representation:

$$\hat{x}_n^{\text{sem}} = (\mathbb{I} - \mathcal{P}_{\mathcal{B}})\mathcal{P}_{\mathcal{S}}(x_n), \quad (2)$$

where $\mathbb{I}$ denotes the identity operator, and $\mathcal{P}_{\mathcal{S}}$ and $\mathcal{P}_{\mathcal{B}}$ represent the orthogonal projection operators onto $\mathcal{S}$ and $\mathcal{B}$ (see Appendix D.2 for details).

However, excessive semantic alignment may inadvertently remove stylistic information. To compensate for this effect, we utilize the residual representation after removing the bias component from the original embedding and project it onto the semantic-preserving subspace to restore contextual and style-related information:

$$\hat{x}_n^{\text{style}} = \mathcal{P}_{\mathcal{S}}(\mathbb{I} - \mathcal{P}_{\mathcal{B}})(x_n). \quad (3)$$

Finally, the semantic-aligned representation and the restored residual representation are combined to construct the final negation-aware embedding:

$$\hat{x}_n = \hat{x}_n^{\text{sem}} + \hat{x}_n^{\text{style}}. \quad (4)$$

The recomposed negation-aware embedding $\hat{x}_n$ is then injected into the frozen diffusion model, enabling negation-aware image generation while preserving the original semantic structure and visual style.

3.3 Theoretical Analysis

In this section, we theoretically analyze the underlying mechanisms of Uni-directional SVD (Uni-SVD) and Bidirectional SVD (Bi-SVD), confirming how DeRO resolves the fundamental trade-off between bias suppression and style preservation. The detailed formal statement and proofs are provided in Appendix C. We assume that a text embedding x_n is decomposed into entangled semantic, bias, and stylistic components within a shared embedding space: $x_n = x_n^{\text{sem}} + x_n^{\text{bias}} + x_n^{\text{style}}$.

Model	TIFA Score ↑							Image Quality	
	missing	empty	nothing	nobody	without	no	mean	PickScore ↑	FID ↓
SD v1.5	0.142	0.252	0.525	0.950	0.166	0.722	0.460	21.245	195.207
SD v1.5 + DATE	0.145	0.255	0.525	0.818	0.166	0.722	0.439 (-0.021)	21.229 (-0.016)	194.615 (-0.592)
SD v1.5 + SupEOT	0.155	0.250	0.525	0.955	0.250	0.770	0.484 (+0.024)	21.174 (-0.071)	208.241 (+13.034)
SD v1.5 + Ours	0.358	0.455	0.526	0.962	0.458	0.812	**0.595** (+0.135)	21.815 (+0.570)	193.091 (-2.116)
SD v3	0.151	0.250	0.752	0.909	0.250	0.762	0.512	21.751	201.016
SD v3 + SoftREPA	0.157	0.281	0.763	0.909	0.167	0.738	0.503 (-0.009)	21.751 (+0.000)	199.158 (-1.858)
SD v3 + Ours	0.362	0.486	0.774	0.916	0.372	0.748	**0.610** (+0.098)	22.134 (+0.383)	197.582 (-3.434)
SDXL	0.412	0.438	0.421	0.747	0.405	0.432	0.476	20.806	202.049
SDXL + Ours	0.461	0.491	0.452	0.782	0.447	0.475	**0.518** (+0.042)	20.890 (+0.084)	201.934 (-0.115)
FLUX.1	0.845	0.450	0.855	0.805	0.250	0.762	0.661	22.265	194.709
FLUX.1 + NAG	0.852	0.820	0.855	0.820	0.667	0.942	0.826 (+0.165)	22.018 (-0.247)	194.965 (+0.256)
FLUX.1 + Ours	0.765	0.951	0.868	1.000	0.667	0.802	**0.842** (+0.181)	22.268 (+0.003)	192.375 (-2.334)

Table 2: Quantitative comparison of object negation understanding and image quality in different generative models. Blue and red values indicate improvements and degradations relative to the corresponding base model, respectively, while **bold** values indicate the best performance in each comparison.

Model	Concept "Picasso" Negation					Concept "Van Gogh" Negation				
	LPIPS↑	TIFA↑	CLIP w/ C↓	CLIP w/o C↑	PickScore↑	LPIPS↑	TIFA↑	CLIP w/ C↓	CLIP w/o C↑	PickScore↑
SD v1.4	–	0.083	0.507	0.521	18.861	–	0.125	0.743	0.843	20.330
SD v1.4 + SPEED	**0.199**	0.036	0.495 (-0.012)	0.515 (-0.006)	19.479 (+0.618)	0.134	0.042	0.584 (-0.159)	0.684 (-0.159)	20.330 (+0.000)
SD v1.4 + MACE	0.181	0.175	0.463 (-0.044)	0.563 (+0.042)	19.841 (+0.980)	0.139	0.158	0.654 (-0.089)	0.754 (-0.089)	20.870 (+0.540)
SD v1.4 + Ours	0.181	**0.179**	**0.341** (-0.165)	**0.612** (+0.091)	**20.330** (+1.469)	**0.146**	**0.167**	**0.339** (-0.404)	**0.854** (+0.011)	**21.420** (+1.090)
SD v1.5	–	0.112	0.772	0.772	19.651	–	0.063	0.749	0.749	21.708
SD v1.5 + RECE	0.202	0.139	0.598 (-0.174)	0.598 (-0.174)	19.165 (-0.486)	0.143	0.103	0.465 (-0.284)	0.465 (-0.284)	18.549 (-3.159)
SD v1.5 + Ours	**0.233**	**0.143**	**0.346** (-0.426)	**0.834** (+0.156)	**20.367** (+0.716)	**0.148**	**0.155**	**0.344** (-0.405)	**0.864** (+0.124)	**21.736** (+0.028)
SDXL	–	0.428	0.594	0.415	20.298	–	0.489	0.607	0.607	19.428
SDXL + Ours	**0.216**	**0.462**	**0.515** (-0.079)	**0.494** (+0.079)	**21.264** (+0.966)	**0.201**	**0.527**	**0.721** (+0.114)	**0.721** (+0.114)	**20.471** (+1.043)
FLUX.1	–	0.444	0.473	0.553	22.075	–	0.205	0.732	0.732	22.109
FLUX.1 + Erasing	0.127	0.517	0.464 (-0.009)	0.511 (-0.042)	21.984 (-0.091)	0.169	0.112	0.735 (+0.003)	0.735 (+0.003)	22.198 (+0.089)
FLUX.1 + UCE	0.087	0.464	0.453 (-0.020)	0.564 (+0.011)	22.189 (+0.114)	0.134	0.422	0.742 (+0.010)	0.742 (+0.010)	22.248 (+0.139)
FLUX.1 + Ours	**0.175**	**0.525**	**0.411** (-0.062)	**0.573** (+0.020)	**22.499** (+0.424)	**0.173**	**0.459**	**0.581** (-0.151)	**0.781** (+0.049)	**23.109** (+1.000)

Table 3: Quantitative comparison of concept negation understanding and image quality in different generative models.

Theorem 1 (Uni-SVD: Style Degradation)**.** Uni-SVD mitigates bias semantics through the operator $\mathcal{T}_{\text{uni}} = \mathbb{I} - \mathcal{P}_{\mathcal{B}}$. However, because stylistic and biased components are inherently entangled (i.e., the stylistic representation shares significant overlap with the bias subspace $\mathcal{B}$), $\mathcal{T}_{\text{uni}}$ inadvertently diminishes these stylistic features alongside the target bias, yielding $\hat{x}_n^{\text{style}} = 0$. Consequently, applying a one-directional projection often leads to a noticeable degradation in visual style when bias suppression is enforced.

Theorem 2 (Bi-SVD: Style Recovery)**.** In contrast, the bidirectional operator $\mathcal{T}_{\text{bi}} = (\mathbb{I} - \mathcal{P}_{\mathcal{B}})\mathcal{P}_{\mathcal{S}} + \mathcal{P}_{\mathcal{S}}(\mathbb{I} - \mathcal{P}_{\mathcal{B}})$ overcomes this limitation through cross-subspace projection. Specifically, by additionally constructing $\hat{x}_n^{\text{style}}$ via cross-subspace projection, Bi-SVD explicitly rescues the stylistic content that Uni-SVD destroys.

4 Experiments

4.1 Object Negation

DeRO is applied to diffusion backbones including the U-Net based Stable Diffusion v1.5 (SD v1.5) (Rombach et al., 2022) and Transformer based Rectified Flow models such as Stable Diffusion v3 (SD v3) (Esser et al., 2024), SDXL (Podell et al., 2024), and FLUX.1 (Labs, 2024). Performance is evaluated using TIFA (Hu et al., 2023), a VQA based metric for assessing text-image semantic alignment, together with FID and PickScore (Kirstain et al., 2023) to evaluate visual quality and the effectiveness of style preservation.

We construct an evaluation set of 480 prompts starting from sentences in the ImageNet validation set. Using Qwen2.5-Instruct, we transform the original sentences into object-negation prompts containing explicit negation expressions such as 'without' and 'no.' These prompts are used to eval-

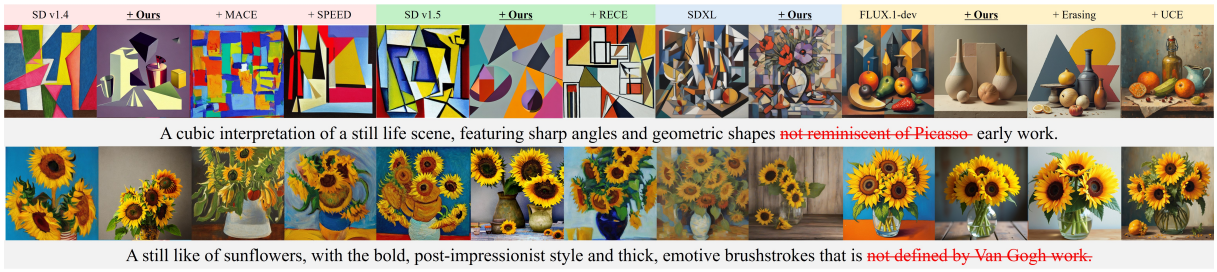


Figure 5: Qualitative comparisons of concept-level negation in different diffusion backbones.

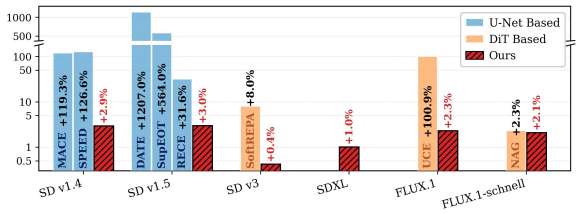


Figure 6: Inference-time overhead comparison in a log scale.

Table 4: Result of the user study.

Method	Object		Concept	
	Align.	Vis.	Align.	Vis.
Base	0.115	0.117	0.125	0.123
Compare	0.347	0.370	0.265	0.307
Ours	**0.536**	**0.512**	**0.609**	**0.569**

uate whether each method can selectively exclude the negated object while preserving the remaining semantics of the prompt.

Performance comparisons are conducted against representative baseline methods in different diffusion backbones, including the Stable Diffusion v1.5 base model and its variants DATE (Na et al., 2025) and SupEOT (Li et al., 2024), the SD v3 base model and its variant SoftREPA (Lee et al., 2025), as well as SDXL and FLUX.1 with its variant NAG (Chen et al., 2025). The quantitative results are summarized in Table 2. Blue values indicate improvements relative to the corresponding base model, while Red values represent performance degradation.

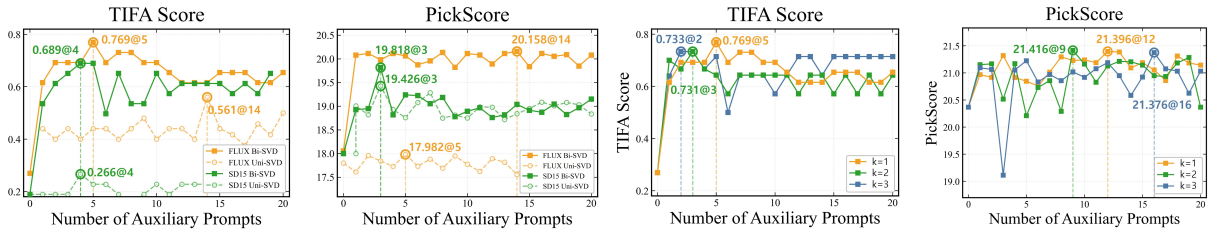
Table 2 presents quantitative comparisons of object negation understanding and image quality. Existing suppression based methods such as SupEOT and NAG improve negation understanding over baseline models, but often struggle to maintain image quality consistently. In contrast, the proposed method effectively improves both negation alignment and generation quality across different diffusion backbones. Under the SD v1.5 setting, DeRO achieves the highest mean TIFA score of 0.595 while improving PickScore from 21.245 to 21.815 and reducing FID from 195.207 to 193.091. Similar trends are consistently observed in SD v3, SDXL, and FLUX.1, demonstrating that auxiliary prompts effectively enhance negation understanding while preserving visual quality.

As shown in Fig. 4, existing methods tend to follow explicit sentence expressions. In contrast, negative attributes are effectively removed, producing semantically cleaner results, such as an empty cat bowl without residual water and a burger without cheese.

4.2 Concept Negation

For concept negation, we compare training-based methods, including MACE (Lu et al., 2024) built on Stable Diffusion v1.4 (SD v1.4), UCE (Gandikota et al., 2024), and Erasing (Gandikota et al., 2023), with training-free methods such as SPEED (Li et al., 2026) and RECE (Gong et al., 2024). Experiments are conducted on the artist negation benchmark (Gandikota et al., 2024); following CURE (Biswas et al., 2025), we select 10 representative artists with 20 prompts per artist, where concept negation targets artist-specific semantic characteristics rather than contextual object co-occurrences. Detailed experimental settings and additional results are provided in Appendix E.3.

We use LPIPS (Zhang et al., 2018) to measure the perceptual difference between images generated by the original model and those generated after concept erasure, where a larger distance indicates stronger modification of the target concept and thus more effective concept removal. TIFA and PickScore follow the same protocol as object negation. For CLIP evaluation, we additionally measure similarity with and without concept terms (CLIP w/ C and CLIP w/o C), since concept



(a) Ablation on Bi-SVD (b) Ablation on SVD Component

Figure 7: Ablation studies on Bi-SVD and semantic basis extraction. (a) Comparison between Uni-SVD and Bi-SVD on FLUX.1 and SD v1.5. (b) Performance analysis with different numbers of SVD components on FLUX.1.

terms may artificially increase similarity when the target concept remains in generated images. Fig. 5 and Table 3 show the results on concept negation. Existing methods often degrade semantic alignment and visual quality. In particular, SPEED and RECE significantly deteriorate both CLIP and PickScore. In contrast, the proposed method consistently preserves semantic alignment and image fidelity. Under the SDXL setting, CLIP improves from 0.415 to 0.494 for Picasso negation and from 0.607 to 0.721 for Van Gogh negation, with consistent PickScore improvements.

Computational Cost. Fig. 6 compares inference time and performance. Since DeRO only performs lightweight SVD analysis on a small set of auxiliary prompts, it introduces negligible computational overhead while maintaining stable performance gains. Further analysis is in Appendix G.

User Study. As shown in Table 4, DeRO achieves the highest human preference scores for both object- and concept-level negation (*Compare* denotes NAG for object negation and Erasing for concept negation). The evaluation involved 32 participants, with the sample size determined through a power analysis ($d = 0.8$, power $= 0.8$, $\alpha = 0.05$). Detailed evaluation protocols, score calculation, and statistical analyses are provided in Appendix H.

4.3 Ablation Studies

Effect of Bi-SVD. Fig. 7(a) compares Uni-SVD and Bi-SVD under varying numbers of auxiliary prompts. Both improve over the base model, indicating that semantic bias disentanglement is beneficial for negation-aware generation. However, Bi-SVD shows more stable semantic alignment and visual quality, reaching optimal performance with fewer auxiliary prompts, whereas Uni-SVD improves more slowly. This can be attributed to the bidirectional projection mechanism of Bi-SVD, which further removes residual biased semantic components after alignment to the semantic-preserving subspace (Eq. 2), while Uni-SVD relies on one-directional suppression that may insufficiently disentangle bias or remove useful semantic information. Similar trends are observed in PickScore. In SD v1.5, incomplete suppression in Uni-SVD occasionally preserves stylistic information, leading to relatively higher PickScore values, whereas excessive suppression in FLUX.1 sometimes degrades visual fidelity. In contrast, Bi-SVD maintains stable PickScore performance while improving semantic alignment.

Table 5: Ablation of individual branches in Bi-SVD in different diffusion backbones.

Base	SVD Method	Score	
		TIFA	PickScore
SD1.5	Uni-SVD	0.510	20.940
	Semantic Branch	0.425	20.900
	Style Branch	0.100	19.504
	Bi-SVD (Ours)	**0.648**	**21.183**
FLUX	Uni-SVD	0.516	21.273
	Semantic Branch	0.431	21.668
	Style Branch	0.416	21.535
	Bi-SVD (Ours)	**0.556**	**22.750**

To further examine the contribution of each branch, Table 5 compares Uni-SVD, the semantic branch only, the style branch only, and the full Bi-SVD. Neither individual branch achieves the performance of the full Bi-SVD across both backbones. In particular, Bi-SVD achieves the highest TIFA and PickScore on both FLUX.1 and SD v1.5, demonstrating that the two branches are complementary for semantic disentanglement and style preservation.

Effect of SVD Component. Fig. 7(b) presents the performance differences under varying numbers of SVD components (k). Larger values of

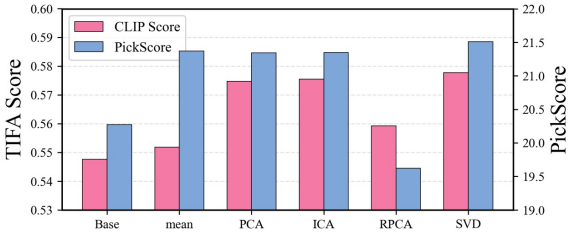


Figure 8: Comparison of different extraction methods for semantic basis decomposition. SVD provides the most stable semantic disentanglement while preserving visual quality, whereas mean aggregation shows limited separation and RPCA exhibits unstable behavior.

k initially show higher TIFA scores, suggesting stronger negation behavior by capturing additional semantic variations associated with the target concept. However, as the number of auxiliary prompts increases, larger k values exhibit bigger performance fluctuations and gradual degradation. This can be attributed to additional noisy components being introduced into the semantic basis together with useful semantic information. In contrast, $k = 1$ consistently achieves more stable performance and the best overall results, as the first singular component mainly captures the dominant semantic direction while minimizing unnecessary noise. Similar trends are also observed in PickScore, where $k = 1$ maintains more stable visual quality and stylistic consistency. Additional quantitative results are provided in Appendix E.5.

Extraction Method. Fig. 8 compares different extraction methods for semantic basis decomposition. Mean aggregation shows limited effectiveness due to insufficient disentanglement of biased semantics, while RPCA tends to produce unstable representations and degraded visual quality. In contrast, SVD more reliably estimates shared semantic structures across auxiliary prompts, leading to more consistent disentanglement behavior and better preservation of image quality. Detailed experimental results and discussion are provided in Appendix E.4.

Robustness of Auxiliary Prompts. We evaluate the robustness of auxiliary prompt construction across different LLMs (Qwen and DeepSeek), prompt types (Noisy and Manual), and selection strategies (Random and Top-1). As shown in Table 6, Manual prompts with Top-1 selection consistently achieve the best performance for both LLMs, reaching TIFA scores of 0.650 and 0.711

Table 6: Robustness of auxiliary prompts across different LLMs, prompt types, and selection strategies.

LLM	Prompt	Selection	TIFA	PickScore
Qwen	Noisy	Random	0.526	21.626
		Top-1	0.550	21.740
	Manual	Random	0.579	21.825
		Top-1 (Ours)	**0.650**	**22.088**
DeepSeek	Noisy	Random	0.503	21.626
		Top-1	0.612	21.750
	Manual	Random	0.604	21.974
		Top-1 (Ours)	**0.711**	**22.148**

with Qwen and DeepSeek (Liu et al., 2024), respectively. Top-1 selection also improves over Random selection in all prompt settings, demonstrating that selecting semantically relevant auxiliary prompts reduces sensitivity to prompt variation. Moreover, the consistent trends across Qwen and DeepSeek indicate that DeRO is not subject to a specific LLM.

5 Concluding Remarks

In this paper, we have proposed a training-free negation-aware generation method that disentangles entangled text representations through a bidirectional SVD-based decomposition-recomposition process between semantic-preserving and bias-suppressing subspaces. The proposed method not only suppresses undesirable biased semantics but also actively guides the representation toward the desired semantic direction while preserving the original semantic context and visual style of the input prompt. However, the proposed method still has limitations. Since semantic subspaces are estimated using auxiliary prompts, performance remains sensitive to prompt quality. Extending the method to more complex linguistic phenomena, such as relational and logical reasoning, remains an important direction for future work.

Acknowledgements

This work was supported by IITP grant funded by the Korea government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University); No. RS-2022-II220113, Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework), and Air Force Defense Research Sciences Program funded by Air Force Office of Scientific Research.

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.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Shristi Das Biswas, Arani Roy, and Kaushik Roy. 2025. CURE: Concept unlearning via orthogonal representation editing in diffusion models. In *Advances in Neural Information Processing Systems*, pages 91710–91734.

Anh Bui, Thuy-Trang Vu, Long Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. 2025. Fantastic targets for concept erasure in diffusion models and where to find them. In *Proceedings of the International Conference on Learning Representations*, pages 84305–84347.

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. 2025. Normalized attention guidance: Universal negative guidance for diffusion models. In *Advances in Neural Information Processing Systems*, pages 152401–152435.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12606–12633.

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. 2023. Fair diffusion: Instructing text-to-image generation models on fairness. *Transactions on Machine Learning Research*.

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. In *Advances in Neural Information Processing Systems*, pages 29488–29512.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436.

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.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, pages 665–673.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. 2024. Reliable and efficient concept erasure of text-to-image diffusion models. In *Proceedings of the European Conference on Computer Vision*, pages 73–88.

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*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Huayang Huang, Xiangye Jin, Jiaxu Miao, and Yu Wu. 2025. Implicit bias injection attacks against text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28779–28789.

Mintong Kang, Vinayshekhar Bannihatti Kumar, Shamik Roy, Abhishek Kumar, Sopan Khosla, Balakrishnan Murali Narayanaswamy, and Rashmi Gangadharaiah. 2025. FairGen: Controlling sensitive attributes for fair generations in diffusion models via adaptive latent guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25336–25350.

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. In *Advances in Neural Information Processing Systems*, pages 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. *arXiv 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. In *Advances in Neural Information Processing Systems*.

Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Tao Liang, Yanbin Hao, Guojun Ma, and Fuli Feng. 2026. SPEED: Scalable, precise, and efficient concept erasure for diffusion models. In *Proceedings of the International Conference on Learning Representations*.

Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Khan, Qibin Hou, Yaxing Wang, and Jian Yang. 2024. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *Proceedings of the International Conference on Learning Representations*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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.

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 *Advances in Neural Information Processing Systems*.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. 2023. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7061.

Jaeil Park and Sung-Bae Cho. 2026. Multi-feature quantized self-attention for fair large language models. In *The Fourteenth International Conference on Learning Representations*.

Sooyeon Park, Jaeil Park, and Sung-Bae Cho. 2026. Pedit: Pareto-guided image editing via dynamic latent trajectory control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4800–4809.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *Proceedings of the International Conference on Learning Representations*.

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 *Proceedings of the International Conference on Machine Learning*, pages 8748–8763.

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*, pages 1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Shaswati Saha, Sourajit Saha, Manas Gaur, and Tejas Gokhale. 2025. Side effects of erasing concepts from diffusion models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14991–15007.

Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. 2019. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2327–2336.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595.

A Preliminaries

Our method is built upon text-conditioned diffusion models. In text-conditioned diffusion models, an input prompt is transformed into latent representations through a text encoder τ_ϕ, and the generation process is guided by these text representations as conditioning information.

U-Net Based. Earlier U-Net based diffusion models (Ho et al., 2020; Rombach et al., 2022) mainly relied on a single text encoder for conditioning. Given an input prompt p, the prompt is transformed into a text embedding representation x, which is injected into cross-attention layers to guide image generation. Typically, CLIP (Radford et al., 2021) is employed as the text encoder, allowing image features to selectively attend to semantic information from the text representation and enabling text-guided synthesis.

DiT Based. Recent transformer based diffusion models, including FLUX.1 (Labs et al., 2025), SD v3 (Esser et al., 2024), and SDXL (Podell et al., 2024), adopt multiple text encoders to simultaneously capture semantic information at different levels. Instead of relying on a single text representation, these models utilize complementary embeddings obtained from different encoders. Specifically, CLIP mainly captures global semantic and contextual information and produces a pooled representation p_{pool}, which is used in feature modulation to scale and shift intermediate representations before attention operations. In parallel, T5 provides token-wise contextual representations p_{emb}, preserving detailed prompt semantics and compositional information. The complete conditioning set $\mathcal{C} = p_{\text{pool}}, p_{\text{emb}}$ is integrated into transformer blocks, where text and image features interact through independent query (Q), key (K), and value (V) transformations.

Such stronger text conditioning mechanisms significantly improve prompt fidelity by enhancing semantic alignment between text and generated images. However, despite these stronger conditioning mechanisms, generated images frequently retain undesired concepts under explicit negation prompts. In this work, we address this limitation by performing text embedding disentanglement directly within the text representation space prior to diffusion sampling.

B Motivation Details

B.1 Profession Text Embedding Decomposition



Figure 9: Illustration of semantic entanglement of gender in profession-related prompt embeddings.

Diffusion models (Labs, 2024; Podell et al., 2024) frequently exhibit semantic entanglement with demographic attributes in profession-related generation scenarios. For example, under profession-oriented prompts such as "a dentist examining a patient," diffusion models often generate male individuals with a high frequency. This indicates that the profession embedding $\mathbf{x}_{\text{dentist}}$ implicitly contains gender-related semantic priors rather than representing only occupation-specific semantics.

When the bias direction can be explicitly defined by a single reference vector (e.g., $\mathbf{x}_{\text{man}}$), the entangled bias component within a target embedding can be isolated via vector projection:

$$\mathbf{x}^{\text{bias}} = \mathcal{P}_{\mathbf{x}_{\text{man}}}(\mathbf{x}_{\text{dentist}}) \tag{5}$$

We assume that the raw profession embedding (e.g., $\mathbf{x}_{\text{dentist}}$) is linearly composed of its core occupational attributes (such as visual style or attire, denoted as $\mathbf{x}_{\text{style}}$) and the entangled demographic bias ($\mathbf{x}_{\text{bias}}$). This relationship can be formulated as follows:

$$\mathbf{x}_{\text{dentist}} = \mathbf{x}_{\text{style}} + \mathbf{x}^{\text{bias}}. \tag{6}$$

By exploiting this geometric formulation, the unbiased, profession-specific semantic component $\mathbf{x}_{\text{style}}$ can be effectively extracted through vector subtraction:

$$\mathbf{x}^{\text{style}} = \mathbf{x}_{\text{dentist}} - \mathbf{x}_{\text{bias}}. \tag{7}$$

Furthermore, we can synthesize a counterfactual embedding by substituting the original biased axis with an alternative demographic direction (e.g., $\mathbf{x}_{\text{woman}}$):

$$\mathbf{x}'_{\text{dentist}} = \mathbf{x}^{\text{style}} + \mathbf{x}_{\text{woman}}. \tag{8}$$

Table 7: Similarity analysis between text embeddings and padding embeddings across different architectures.

Arch	Max Similarity	Mean Similarity
DiT	0.142	0.024
U-Net	0.992	0.403

The qualitative results generated using the gender-modified counterfactual embedding are shown in Fig. 9.

General Token. However, extending this approach to general object-level concepts is challenging because semantic bias cannot be represented by a single explicit direction such as x_{man} or x_{woman}. In practice, semantic bias is distributed across multiple correlated patterns. For example, in the prompt "bakery with no bread," bread-related semantics are entangled with associated concepts such as cakes, trays, and food structures. Therefore, instead of relying on a single semantic direction, we construct semantic subspaces using multiple auxiliary prompts. Shared semantic components are extracted from related prompt embeddings, enabling more stable estimation of semantic information for negation-aware embedding construction.

B.2 Padding Analysis

SupEOT performs negation-aware generation by suppressing semantic components within padding embeddings. The method assumes that padding regions in text embeddings still contain meaningful semantic information, and that removing such components can improve negation performance. However, unlike U-Net architectures, recent DiT models exhibit relatively cleaner padding embeddings with minimal semantic information. As shown in Tab. 7, U-Net models demonstrate relatively high similarity between text and padding embeddings, suggesting that padding regions retain meaningful semantic content. In contrast, DiT architectures show extremely low similarity between text and padding embeddings, indicating that padding regions are largely disentangled from text semantics. These observations suggest that the padding embedding manipulation strategy proposed in SupEOT may not generalize effectively across different architectures.

C Theoretical Analysis

This section formalizes why Uni-SVD fails to preserve stylistic information under negation, and how the proposed Bi-SVD overcomes this failure through cross-subspace recovery.

Definition 1 (Subspace Structure)**.** Let $\mathcal{P}_{\mathcal{S}}$ and $\mathcal{P}_{\mathcal{B}}$ denote the orthogonal projectors onto the semantic-preserving subspace $\mathcal{S}$ and the bias subspace $\mathcal{B}$, respectively. The principal angles $\theta_1 \leq \cdots \leq \theta_r$ between $\mathcal{S}$ and $\mathcal{B}$ satisfy $\cos\theta_i = \sigma_i(V_S^{(k)\top} V_B^{(k)})$, where $\sigma_i(\cdot)$ denotes the i-th largest singular value. A smaller θ_1 indicates stronger entanglement, and θ_1 is non-increasing in k as larger subspaces admit more overlapping directions.

Definition 2 (Embedding Decomposition)**.** A negation prompt embedding $x_n \in \mathbb{R}^d$ is decomposed as

$$x_n = x_n^{\text{sem}} + x_n^{\text{bias}} + x_n^{\text{style},B} + x_n^{\text{style},S}, \quad (9)$$

under the following assumptions:

(A1) $x_n^{\text{sem}} \in \mathcal{S}$ with $\mathcal{P}_{\mathcal{B}}(x_n^{\text{sem}}) = 0$.

(A2) $x_n^{\text{bias}} \in \mathcal{B}$ with $\mathcal{P}_{\mathcal{S}}(x_n^{\text{bias}}) = 0$.

(A3) $x_n^{\text{style},B} \in \mathcal{B}$ with $\mathcal{P}_{\mathcal{S}}(x_n^{\text{style},B}) = \eta_b \neq 0$, representing bias-entangled style.

(A4) $x_n^{\text{style},S} \in \mathcal{S}$ with $\mathcal{P}_{\mathcal{B}}(x_n^{\text{style},S}) = \eta_s \neq 0$, representing semantic-aligned style.

Definition 3 (Operators)**.**

$$\begin{aligned}
\mathcal{T}_{\text{uni}} &= \mathbb{I} - \mathcal{P}_{\mathcal{B}}, \\
\mathcal{P}^{\text{sem}} &= (\mathbb{I} - \mathcal{P}_{\mathcal{B}})\,\mathcal{P}_{\mathcal{S}}, \\
\mathcal{P}^{\text{style}} &= \mathcal{P}_{\mathcal{S}}\,(\mathbb{I} - \mathcal{P}_{\mathcal{B}}), \\
\mathcal{T}_{\text{bi}} &= \mathcal{P}^{\text{sem}} + \mathcal{P}^{\text{style}} \\
&= 2\mathcal{P}_{\mathcal{S}} - \mathcal{P}_{\mathcal{B}}\mathcal{P}_{\mathcal{S}} - \mathcal{P}_{\mathcal{S}}\mathcal{P}_{\mathcal{B}}.
\end{aligned}$$

We define $\Delta_{\text{style}} = \mathcal{P}_{\mathcal{S}}\mathcal{P}^{\text{sem}}(x_n^{\text{style},B})$ and $\Delta_{\text{bias}} = \mathcal{P}_{\mathcal{B}}\mathcal{P}^{\text{style}}(x_n^{\text{style},S})$ to denote the recovered style and residual bias terms, respectively.

C.1 Limitations of Uni-SVD

Theorem 1 (Uni-SVD: Style Degradation)**.** Under **(A1)–(A4)**, $\mathcal{T}_{\text{uni}} = \mathbb{I} - \mathcal{P}_{\mathcal{B}}$ yields:

$$\mathcal{T}_{\text{uni}}(x_n^{\text{bias}}) = 0, \qquad \mathcal{T}_{\text{uni}}(x_n^{\text{style},B}) = 0. \quad (10)$$

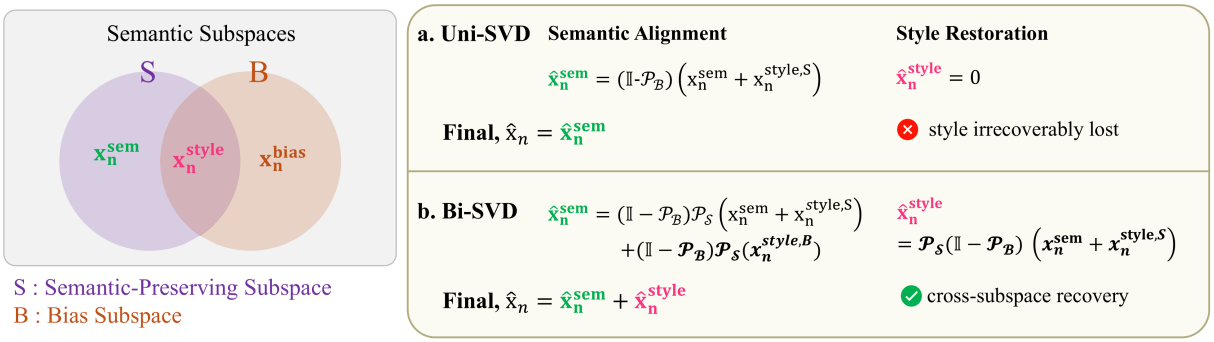


Figure 10: Comparison between Uni-SVD and Bi-SVD for negation embedding decomposition. Unlike Uni-SVD, Bi-SVD selectively preserves contextual semantics while suppressing bias-related representations.

Proof.

For x_n^{bias}: by **(A2)**, $\mathcal{P}_\mathcal{B}(x_n^{\text{bias}}) = x_n^{\text{bias}}$, thus
$\mathcal{T}_{\text{uni}}(x_n^{\text{bias}}) = (\mathbb{I} - \mathcal{P}_\mathcal{B})(x_n^{\text{bias}}) = 0$

For $x_n^{\text{style,B}}$: by **(A3)**, $\mathcal{P}_\mathcal{B}(x_n^{\text{style,B}}) = x_n^{\text{style,B}}$, thus
$\mathcal{T}_{\text{uni}}(x_n^{\text{style,B}}) = (\mathbb{I} - \mathcal{P}_\mathcal{B})(x_n^{\text{style,B}}) = 0$

Because $\mathcal{T}_{\text{uni}}$ treats the entire subspace $\mathcal{B}$ uniformly, it cannot distinguish bias from style entangled within $\mathcal{B}$. Consequently, the stylistic component x_n^{style} is irrecoverably lost together with the bias, directly causing the visual fidelity collapse observed in Fig. 2(c). Stronger entanglement (larger $\|x_n^{\text{style}}\|$) intensifies this loss, and no choice of suppression strength within the uni-directional framework can simultaneously remove bias and preserve x_n^{style}.

C.2 Properties of Bi-SVD

Theorem 2 (Bi-SVD: Bounded Bias and Style Recovery)**.** Under **(A1)**–**(A3)**, the bidirectional operator $\mathcal{T}_{\text{bi}}$ satisfies:
(a) Bias suppression:

$$\mathcal{T}_{\text{bi}}(x_n^{\text{bias}}) = 0. \tag{11}$$

(b) Bounded bias residual:

$$\mathcal{P}_\mathcal{B}(\mathcal{T}_{\text{bi}}(x_n)) = \Delta_{\text{bias}}, \tag{12}$$

where $\|\Delta_{\text{bias}}\| \le \cos\theta_1 \, \|(\mathbb{I} - \mathcal{P}_\mathcal{B})\, x_n\|$.

The residual Δ_{bias} arises solely from η_s through a three-step attenuation chain $\eta_s \to \mathcal{P}_\mathcal{S}(\eta_s) \to \mathcal{P}_\mathcal{B}\mathcal{P}_\mathcal{S}(\eta_s)$, each step reducing energy by $\cos\theta_1$. It becomes negligible when auxiliary prompts yield well-separated subspaces, and vanishes exactly when $\eta_s = 0$.

(c) Cross-subspace style recovery:

$$\mathcal{T}_{\text{bi}}(x_n^{\text{style},B}) = \mathcal{P}^{\text{sem}}(x_n^{\text{style},B}) \neq 0, \tag{13}$$

$$\mathcal{P}^{\text{style}}(x_n^{\text{style},S}) \neq 0. \tag{14}$$

(d) Additional semantic content vs. Uni-SVD:

$$\mathcal{P}_\mathcal{S}(\mathcal{T}_{\text{bi}}(x_n)) = 2\,\mathcal{P}_\mathcal{S}(\mathcal{T}_{\text{uni}}(x_n)) + \Delta_{\text{style}}. \tag{15}$$

While $\mathcal{T}_{\text{uni}}$ unconditionally annihilates x_n^{style} (Theorem 1), $\mathcal{T}_{\text{bi}}$ rescues its semantic-aligned content via cross-subspace recovery. The total preserved semantic content of Bi-SVD strictly exceeds that of Uni-SVD by Δ_{style}, which represents bias-free stylistic knowledge extracted from $\mathcal{B}$ that is structurally inaccessible to any uni-directional operator.

Proof. **(a)** By **(A2)**, $\mathcal{P}_\mathcal{S}(x_n^{\text{bias}}) = 0$ and $(\mathbb{I} - \mathcal{P}_\mathcal{B})(x_n^{\text{bias}}) = 0$. Thus:

$$\begin{aligned}
\mathcal{P}^{\text{sem}}(x_n^{\text{bias}}) &= (\mathbb{I} - \mathcal{P}_\mathcal{B})\mathcal{P}_\mathcal{S}(x_n^{\text{bias}}) = 0 \\
\mathcal{P}^{\text{style}}(x_n^{\text{bias}}) &= \mathcal{P}_\mathcal{S}(\mathbb{I} - \mathcal{P}_\mathcal{B})(x_n^{\text{bias}}) = 0 \\
\therefore\ \mathcal{T}_{\text{bi}}(x_n^{\text{bias}}) &= \mathcal{P}^{\text{sem}}(x_n^{\text{bias}}) + \mathcal{P}^{\text{style}}(x_n^{\text{bias}}) = 0
\end{aligned}$$

(c) For $x_n^{\text{style,B}}$: by **(A3)**, $\mathcal{P}_\mathcal{S}(x_n^{\text{style,B}}) = \eta_b$ and $(\mathbb{I} - \mathcal{P}_\mathcal{B})(x_n^{\text{style,B}}) = 0$. Thus:

$$\begin{aligned}
\mathcal{P}^{\text{style}}(x_n^{\text{style,B}}) &= \mathcal{P}_\mathcal{S}(\mathbb{I} - \mathcal{P}_\mathcal{B})(x_n^{\text{style,B}}) = 0 \\
\mathcal{P}^{\text{sem}}(x_n^{\text{style,B}}) &= (\mathbb{I} - \mathcal{P}_\mathcal{B})\mathcal{P}_\mathcal{S}(x_n^{\text{style,B}}) \\
&= (\mathbb{I} - \mathcal{P}_\mathcal{B})\,\eta_b \neq 0 \\
\therefore\ \mathcal{T}_{\text{bi}}(x_n^{\text{style,B}}) &= (\mathbb{I} - \mathcal{P}_\mathcal{B})\,\eta_b \neq 0
\end{aligned}$$

For $x_n^{\text{style,S}}$: by **(A4)**, $\mathcal{P}_\mathcal{B}(x_n^{\text{style,S}}) = \eta_s$ and $\mathcal{P}_\mathcal{S}(x_n^{\text{style,S}}) = x_n^{\text{style,S}}$. Thus:

$$\begin{aligned}
\mathcal{P}^{\text{style}}(x_n^{\text{style,S}}) &= \mathcal{P}_\mathcal{S}(\mathbb{I} - \mathcal{P}_\mathcal{B})(x_n^{\text{style,S}}) \\
&= \mathcal{P}_\mathcal{S}(x_n^{\text{style,S}} - \eta_s) \\
&= x_n^{\text{style,S}} - \mathcal{P}_\mathcal{S}(\eta_s) \neq 0
\end{aligned}$$

(b) Since $\mathcal{P}_{\mathcal{B}}(\mathbb{I}-\mathcal{P}_{\mathcal{B}})=0$, we have $\mathcal{P}_{\mathcal{B}}\mathcal{P}^{\text{sem}}=0$. Thus:

$$\begin{aligned}
&\mathcal{P}_{\mathcal{B}}(\mathcal{T}_{\text{bi}}(x_n)) \\
&=\mathcal{P}_{\mathcal{B}}(\mathcal{T}_{\text{bi}}(x_n^{\text{style,S}}+x_n^{\text{style,B}})) \quad \text{(by Thm. 2a, (A1))} \\
&=\mathcal{P}_{\mathcal{B}}(\mathcal{T}_{\text{bi}}(x_n^{\text{style,S}})+\mathcal{P}^{\text{sem}}(x_n^{\text{style,B}})) \text{ (by Thm. 2c)} \\
&=\mathcal{P}_{\mathcal{B}}(\mathcal{P}^{\text{style}}(x_n^{\text{style,S}})) \quad (\text{since } \mathcal{P}_{\mathcal{B}}\mathcal{P}^{\text{sem}}=0) \\
&=\Delta_{\text{bias}}.
\end{aligned}$$

The bound follows from $\|\mathcal{P}_{\mathcal{B}}\mathcal{P}_{\mathcal{S}}\|_{\text{op}}=\cos\theta_1$.

(d) First, by Theorem 1, $\mathcal{T}_{\text{uni}}(x_n^{\text{bias}}) = \mathcal{T}_{\text{uni}}(x_n^{\text{style,B}})=0$:

$$\mathcal{P}_{\mathcal{S}}(\mathcal{T}_{\text{uni}}(x_n)) = \mathcal{P}^{\text{style}}(x_n^{\text{sem}}) + \mathcal{P}^{\text{style}}(x_n^{\text{style,S}})$$

For $\mathcal{T}_{\text{bi}}$, applying $\mathcal{T}_{\text{bi}} = 2\mathcal{P}_{\mathcal{S}} - \mathcal{P}_{\mathcal{B}}\mathcal{P}_{\mathcal{S}} - \mathcal{P}_{\mathcal{S}}\mathcal{P}_{\mathcal{B}}$:

$$\begin{aligned}
\mathcal{P}_{\mathcal{S}}(\mathcal{T}_{\text{bi}}(x_n^{\text{sem}})) &= 2\mathcal{P}_{\mathcal{S}}x_n^{\text{sem}} - 2\mathcal{P}_{\mathcal{S}}\mathcal{P}_{\mathcal{B}}(x_n^{\text{sem}}) \\
&= 2\mathcal{P}^{\text{style}}(x_n^{\text{sem}})
\end{aligned}$$

$$\begin{aligned}
\mathcal{P}_{\mathcal{S}}(\mathcal{T}_{\text{bi}}(x_n^{\text{style,S}})) &= \mathcal{P}_{\mathcal{S}}(2x_n^{\text{style,S}} - \eta_s - \mathcal{P}_{\mathcal{S}}(\eta_s)) \\
&= 2\mathcal{P}^{\text{style}}(x_n^{\text{style,S}})
\end{aligned}$$

$$\begin{aligned}
&\mathcal{P}_{\mathcal{S}}(\mathcal{T}_{\text{bi}}(x_n^{\text{style,B}})) \\
&= \mathcal{P}_{\mathcal{S}}\mathcal{P}^{\text{sem}}(x_n^{\text{style,B}}) + \mathcal{P}_{\mathcal{S}}\underbrace{(\mathbb{I}-\mathcal{P}_{\mathcal{B}})(x_n^{\text{style,B}})}_{=0 \text{ (by (A3))}} \\
&= \Delta_{\text{style}}
\end{aligned}$$

Combining with $\mathcal{P}_{\mathcal{S}}(\mathcal{T}_{\text{bi}}(x_n^{\text{bias}})) = 0$ (by Theorem 2a), we obtain:

$$\mathcal{P}_{\mathcal{S}}(\mathcal{T}_{\text{bi}}(x_n)) = 2\,\mathcal{P}_{\mathcal{S}}(\mathcal{T}_{\text{uni}}(x_n)) + \Delta_{\text{style}}.$$

D Method Detail

D.1 Auxiliary Prompts

Auxiliary prompts are generated using Qwen2.5-Instruct (Bai et al., 2025). Given a negation prompt p_n, we construct two auxiliary prompt sets: a biased semantic set A_B that strongly represents the target concept and a semantic set A_S that preserves the original semantics while avoiding explicit emphasis on the target concept.

To improve prompt quality and semantic consistency, we employ Sentence-BERT (Reimers and Gurevych, 2019) to compute sentence-level semantic similarities between generated prompts and target semantic concepts. In each round, $C = 20$ candidate prompts are generated and the one maximizing semantic similarity to its corresponding target concept is selected. In addition, prompt length is constrained to maintain concise and visually descriptive representations. Detailed LLM instructions and prompt selection strategy are summarized in Fig. 11 and Alg. 1, respectively.

Algorithm 1 Iterative Auxiliary Prompt Selection Strategy

Require: Negation prompt p_n, biased concept p_b, preserved semantic concept p_s, number of rounds N, candidates per round C
Ensure: Biased prompt set A_B, Semantic prompt set A_S
1: Initialize prompt sets: $A_B \leftarrow \emptyset$, $A_S \leftarrow \emptyset$.
2: **for** $i = 1$ to N **do**
3: Generate C candidate prompts for round i via LLM:

$$A_B^i = \{a_b^{i,1}, a_b^{i,2}, \ldots, a_b^{i,C}\} \sim \text{LLM}(p_b)$$

$$A_S^i = \{a_s^{i,1}, a_s^{i,2}, \ldots, a_s^{i,C}\} \sim \text{LLM}(p_s)$$

4: Select the optimal prompt maximizing semantic similarity via text encoder $\mathcal{E}$:

$$a_b^i = \arg\max_{a \in A_B^i} \text{Sim}(\mathcal{E}(a), \mathcal{E}(p_b))$$

$$a_s^i = \arg\max_{a \in A_S^i} \text{Sim}(\mathcal{E}(a), \mathcal{E}(p_s))$$

5: Accumulate selected prompts: $A_B \leftarrow A_B \cup \{a_b^i\}$, $A_S \leftarrow A_S \cup \{a_s^i\}$
6: **end for**
7: **return** A_B, A_S

Instruction

You are a model responsible for constructing auxiliary prompts for semantic subspace decomposition. Given a negation and semantic concept, generate two types of auxiliary prompt sets.

Example Input

[BIASED]
bread

[SEMANTIC]
an empty shop interior with clean shelves

Definition

[BIASED]
Prompts that strongly contain or imply the target negation concept. These prompts are used to construct the bias subspace $\mathcal{B}$.

[SEMANTIC]
Prompts that avoid the target negation concept while preserving the original semantic. These prompts are used to construct the semantic-preserving subspace $\mathcal{S}$.

Generation Guidelines

[BIASED]
- Strongly activate the target negation concept.
- Describe scenes where the base semantic naturally contains the target negation concept.
- Use concrete and visually descriptive language.
- Preserve consistency with the original semantic concept.

[SEMANTIC]
- Avoid directly mentioning the target negation concept.
- Preserve the original semantic.
- Focus on visually similar, clean, empty, minimal, or uncluttered scenes.
- Do not explicitly use negation expressions such as “without”, “no”, “missing”, or “lacks”.

Example Output

[BIASED]
- A bakery filled with freshly baked bread on display shelves.
- A bakery counter stacked with warm loaves of bread.
- Fresh bread arranged inside a bakery display case.

[SEMANTIC]
- A clean bakery interior with organized display shelves.
- A quiet bakery shop with minimal interior structure.
- A bakery counter with neatly arranged empty display racks.

Figure 11: LLM Instructions for Auxiliary Prompt Construction

D.2 Bidirectional Projection Algorithm

Given a negation prompt p_n, the proposed method adaptively refines the text embedding by geometrically disentangling the pure semantics and residual styles from implicit biases, as summarized in Alg. 2.

Algorithm 2 Negation-aware Text Embedding via Bi-SVD

Require: Negation prompt p_n (embedding x_n), auxiliary embeddings $X_S, X_B \in \mathbb{R}^{N\times d}$
Ensure: Negation-aware embedding $\hat{x}_n$

// Step 1: Subspace Estimation via Mean-Centered SVD

1: Calculate structural means:

$$\mu_{\mathcal{S}} = \mathrm{Mean}(X_S), \qquad \mu_{\mathcal{B}} = \mathrm{Mean}(X_B)$$

2: Perform SVD on mean-centered matrices (Eq. 1):

$$X_S - \mu_{\mathcal{S}} = U_S \Sigma_S V_S^\top,$$
$$X_B - \mu_{\mathcal{B}} = U_B \Sigma_B V_B^\top.$$

3: Define affine projection operators with top-k bases $V_S^{(k)}$ and $V_B^{(k)}$:

$$\mathcal{P}_{\mathcal{S}}(x) = \mu_{\mathcal{S}} + V_S^{(k)} {V_S^{(k)}}^\top (x - \mu_{\mathcal{S}})$$

$$\mathcal{P}_{\mathcal{B}}(x) = \mu_{\mathcal{B}} + V_B^{(k)} {V_B^{(k)}}^\top (x - \mu_{\mathcal{B}})$$

// Step 2: Bidirectional Projection & Residual Bias Removal

4: Extract the bias-reduced semantic representation (Eq. 2):

$$\hat{x}_n^{\mathrm{sem}} = (\mathbb{I} - \mathcal{P}_{\mathcal{B}})\mathcal{P}_{\mathcal{S}}(x_n)$$

// Step 3: Residual Style Restoration & Final Recomposition

5: Restore residual style details onto $\mathcal{S}$ (Eq. 3):

$$\hat{x}_n^{\mathrm{style}} = \mathcal{P}_{\mathcal{S}}(\mathbb{I} - \mathcal{P}_{\mathcal{B}})(x_n)$$

6: Recompose the final negation-aware text embedding (Eq. 4):

$$\hat{x}_n = \hat{x}_n^{\mathrm{sem}} + \hat{x}_n^{\mathrm{style}}$$

7: **return** $\hat{x}_n$

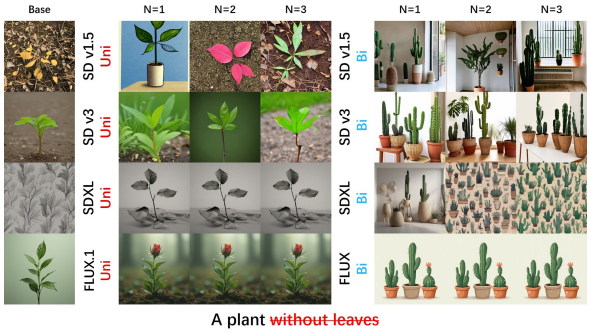


Figure 12: Qualitative comparison of Base, Uni-SVD, and Bi-SVD under varying numbers of auxiliary prompts.

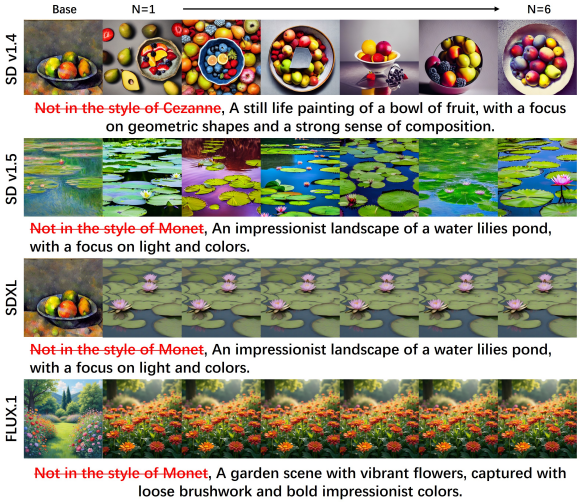


Figure 13: Qualitative results of concept negation with varying numbers of auxiliary prompts.

E Experimental Details

E.1 Experimental Setup

All experiments were conducted on an NVIDIA A100 GPU. To ensure a fair comparison, the number of inference steps was kept identical to the default settings of each corresponding base model, and a consistent random seed was used across all methods. Furthermore, all models utilized in this study were used in strict compliance with their respective licenses.

E.2 Uni-SVD vs Bi-SVD

Fig. 12 provides additional qualitative results supporting the analysis in § 4.3 regarding the effectiveness of Bi-SVD. U-Net based models such as SD v1.4 and SD v1.5 exhibit unstable removal of target negation semantics when **Uni-SVD** is applied. As discussed in prior work such as SupEOT, the generation process of diffusion models is influenced not only by text embeddings but also by multiple factors including padding tokens and other latent interactions. Consequently, controlling only a single semantic direction through **Uni-SVD** becomes insufficient for resolving such complex semantic entanglement in a stable manner.

In contrast, under FLUX.1, **Uni-SVD** gradually estimates more meaningful semantic directions as the number of auxiliary prompts increases, resulting in improved semantic disentanglement performance. This behavior becomes particularly evident in FLUX.1 because, as reported in previous studies (Dalva et al., 2024), the embedding representations of FLUX.1 exhibit relatively disentangled properties compared to other diffusion models, leading to more clearly separated semantic structures. As a result, dominant semantic components can be estimated more reliably even through simple unidirectional semantic decomposition, allowing **Uni-SVD** to achieve a certain level of semantic manipulation performance.

However, for another Transformer based architecture, SD v3, **Uni-SVD** still fails to reliably preserve negation behavior despite its stronger text understanding capability. This observation suggests that the effectiveness of **Uni-SVD** is not determined solely by the backbone architecture itself. Unless the underlying embedding representations already possess disentangled semantic structures similar to FLUX.1, **Uni-SVD** remains limited in handling strongly entangled semantic scenarios.

In contrast, the proposed **Bi-SVD** forms a more robust semantic subspace as the number of auxiliary prompts increases, allowing semantic directions to be estimated more reliably while simultaneously separating semantic and stylistic information through bidirectional decomposition. These results suggest that simple unidirectional suppression strategies are difficult to generalize across diverse diffusion backbones, whereas the proposed bidirectional decomposition framework provides a more robust and generalizable solution for negation-aware generation.

E.3 Concept Negation

For concept-level negation, DeRO uses only a single auxiliary semantic prompt (A_S) and a single manually defined biased prompt (A_B) by default. Unlike object-level negation, where target semantics are often strongly entangled with surrounding contextual information, concept-level semantics are generally more directly associated with the tar-

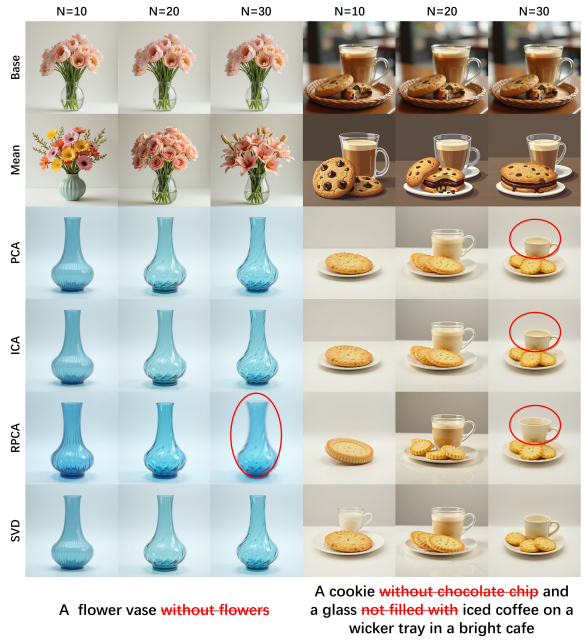


Figure 14: Qualitative comparisons of different component decomposition methods under single and multi-negation settings.

get concept itself. As a result, the dependency on multiple auxiliary prompts becomes substantially lower, and stable semantic removal can already be achieved using a single auxiliary prompt pair across different diffusion backbones. Additional qualitative comparisons are provided in Fig. 13.

E.4 Extraction Method

Concept negation evaluates whether concept-specific characteristics remain despite explicit negation prompts; for example, given a prompt p_n such as "a garden not in the style of Monet", we assess whether Monet-style characteristics are successfully removed from generated images using artistic concepts.

Fig. 14 provides additional qualitative results for the analysis in § 4.3 (Extraction Method). To investigate the effect of component extraction methods for semantic basis decomposition in the prompt embedding space, we compare Mean, Principal Component Analysis (**PCA**), Independent Component Analysis (**ICA**), Robust Principal Component Analysis (**RPCA**), and Singular Value Decomposition (**SVD**) under both single-negation and multi-negation settings.

Mean exhibits the weakest performance because simple averaging fails to effectively suppress shared biased semantics, resulting in residual bias in generated images.

RPCA exhibits the most unstable behavior. **RPCA** assumes that the input representation can be decomposed into low-rank structures and sparse residual components. However, auxiliary prompts often contain meaningful semantic variations rather than sparse noise. Consequently, some semantic information may be incorrectly separated into sparse residual components, distorting the underlying semantic structure of the embedding space and degrading both semantic consistency and image quality. As shown on the left side of Fig. 14, this effect becomes more noticeable as the number of auxiliary prompts increases, leading to increasingly blurry generated results.

While **PCA** and **ICA** are highly effective at capturing simple, dominant semantic structures, they fail to maintain semantic consistency in complex, multi-negation scenarios. In single-negation settings, **PCA** extracts dominant semantic directions through variance maximization, whereas **ICA** aims to separate statistically independent latent components; both methods perform reliably because the underlying semantic patterns are straightforward. However, in multi-negation scenarios, multiple semantic attributes become simultaneously entangled and interact with each other within the text representation. For instance, under complex prompts such as "*not filled with iced coffee*," both methods struggle to consistently satisfy multiple semantic conditions. This demonstrates that relying solely on dominant directions or independent components is insufficient for representing intertwined semantic factors, ultimately leading to a degradation of semantic consistency.

SVD consistently demonstrates the most robust behavior across different settings. Unlike **PCA**, which tends to be sensitive to dominant, high-variance semantic factors, **SVD** estimates the principal semantic structures repeatedly shared across auxiliary prompts in a much more stable manner. Given that multiple entangled semantic directions coexist in multi-negation scenarios, the stable estimation of these shared structures becomes increasingly critical. Consequently, **SVD** achieves more reliable semantic disentanglement and better preserves semantic consistency across diverse and complex conditions.

E.5 SVD Component

Fig. 15 provides additional qualitative results for the analysis in § 4.3 (Effect of SVD Component). The results show that using only $k = 1$ is sufficient

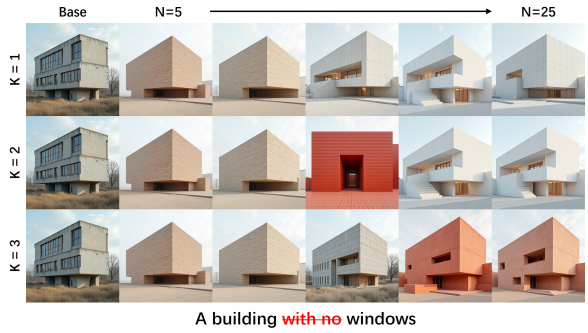


Figure 15: Qualitative comparisons with different numbers of SVD components (k).

to capture the dominant semantic direction shared across auxiliary prompts, enabling stable semantic disentanglement. In contrast, including additional components tends to introduce finer semantic variations, contextual information, and residual stylistic representations. Since these components are not always directly related to negation semantics, they may distort the original semantic structure or weaken negation performance. Therefore, we use only the first singular component in all experiments.

F Additional Experiment

F.1 Nude Concept Negation

** As this section involves sensitive visual concepts, the proposed framework is intended solely for research purposes to evaluate semantic disentanglement and negation understanding, rather than for generating or modifying inappropriate content. **

To evaluate the generalization capability of DeRO, we conduct additional nude concept negation experiments using a dataset adapted from (Gong et al., 2024) across SD v1.4, SD v1.5, FLUX.1, and SDXL. Similar to artist concept negation, where strong semantic associations are induced by artist identities, nude concept negation also exhibits strong concept-specific priors embedded within pretrained text representations. Consequently, diffusion models often retain nudity-related semantics even under explicit negation prompts such as *without nudity*.

To examine whether generated images correctly follow prompts containing explicit negation expressions, we apply DeRO and present qualitative results in Fig. 16. Compared to the base models, DeRO demonstrates substantially improved adherence to the input prompts.

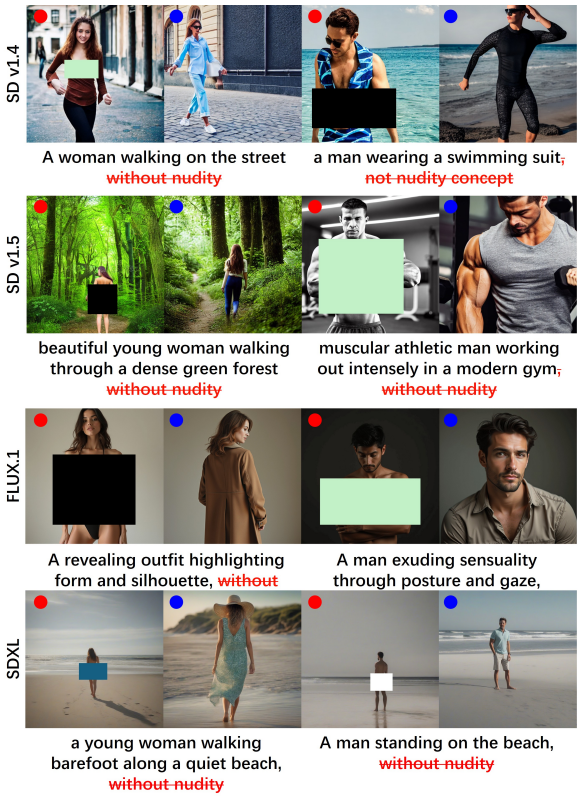


Figure 16: Qualitative comparisons for nude concept negation across SD v1.4, SD v1.5, FLUX.1, and SDXL. Regions highlighted with red markers indicate failure cases in the base model, whereas regions highlighted with blue markers correspond to improvements achieved by DeRO.

Specifically, DeRO consistently suppresses nudity-related biased semantics while preserving clothing structure, human posture, and overall visual realism across different diffusion backbones. These results suggest that the proposed Bi-SVD based decomposition effectively generalizes beyond artist-style negation toward broader semantic concept negation scenarios.

F.2 Additional Application

DeRO is additionally evaluated on FLUX.1-schnell (Labs, 2024), a distilled version of FLUX.1. The results show that DeRO can also be effectively applied to distilled diffusion models. However, due to compressed semantic representations introduced by the distillation process, semantic subspace estimation becomes relatively unstable when only a limited number of auxiliary prompts are available. To compensate for this limitation, we use 12 auxiliary prompts in our experiments. Qualitative examples are in Fig. 22.

G Computational Cost

To compare computational efficiency across different backbones with varying image resolutions and generation settings, we report inference latency normalized by the corresponding base model. As shown in Tab. 8, optimization-based methods introduce substantial computational costs. DATE (Na et al., 2025) and SupEOT (Li et al., 2024) significantly increase inference time, reaching 13.07× and 6.64× relative latency, respectively. Model editing approaches such as MACE (Lu et al., 2024), SPEED (Li et al., 2026), and UCE (Gandikota et al., 2024) also incur considerable overhead due to parameter manipulation or additional optimization procedures. In contrast, DeRO performs only a lightweight SVD operation in the text embedding space, introducing negligible computational overhead. Across diverse backbone architectures, it consistently requires only 1.004× to 1.029× relative latency compared to the corresponding base models.

Table 8: Comparison of inference latency based on the different diffusion backbones.

Arch	Base	Model	Latency	
			Time(s)	Relative
SD v1.4	U-Net	Base	2.001	-
		MACE	4.389	2.194×
		SPEED	4.534	2.266×
		Ours	**2.060**	**1.029**×
SD v1.5	U-Net	Base	2.124	-
		DATE	27.761	13.070×
		SupEOT	14.103	6.640×
		RECE	2.795	1.316×
		Ours	**2.060**	**1.029**×
SD v3	DiT	Base	4.947	-
		SoftREPA	5.343	1.080×
		Ours	**4.968**	**1.004**×
SDXL	DiT	Base	5.991	-
		Ours	**6.052**	**1.010**×
FLUX.1	DiT	Base	99.067	-
		UCE	199.026	2.009×
		Ours	**101.356**	**1.023**×
FLUX.1-schnell	DiT	Base	78.729	-
		NAG	80.540	1.023×
		Ours	**80.382**	**1.021**×

H User Study Details

We conduct a human evaluation of object- and concept-level negation using FLUX.1, which achieved the best overall performance in the main experiments. For object negation, we compare the FLUX.1 base model, NAG, and DeRO, while for concept negation, we compare the FLUX.1 base model, Erasing, and DeRO. A total of 32 volunteers participated without financial compensation. All participants provided informed consent prior to the study and were informed that their responses would be used solely for research purposes. The sample size was determined through a power analysis following Cohen (Cohen, 1960), assuming a large effect size ($d = 0.8$), power of 0.8, and a significance level of $\alpha = 0.05$. Each participant evaluated 12 randomly sampled prompts for each negation type and ranked the three generated images based on *Prompt Alignment* and *Visual Quality*, measuring adherence to the prompt including the specified negation and overall perceptual quality, respectively. The rankings were converted into Borda scores of 1.0, 0.5, and 0.0 for first, second, and third place, respectively, accumulated across all participants and prompts, and normalized by the maximum possible score. The resulting preference scores are reported in Tab. 4. Wilcoxon signed-rank tests on participant-level responses show that DeRO significantly outperforms the competing methods across all four criteria: Object Alignment ($p = 3.6 \times 10^{-7}$), Object Visual Quality ($p = 3.8 \times 10^{-7}$), Concept Alignment ($p = 3.7 \times 10^{-7}$), and Concept Visual Quality ($p = 3.8 \times 10^{-7}$), with all differences significant at $p < 0.001$. The evaluation instructions and example questions are provided in Fig. 17.

I Additional Qualitative Results

Additional qualitative results are presented from Fig. 18 to Fig. 23. Each image set compares the generation results of the base model and the proposed method (DeRO). Fig. 18 and Fig. 19 present results on U-Net based diffusion models, including SD v1.4 and SD v1.5. Fig. 20 to Fig. 23 present results on DiT based diffusion models, including SD v3, SDXL, FLUX.1-schnell, and FLUX.1.

J Use of AI Assistants

We utilized LLMs (e.g., ChatGPT, Gemini) to support manuscript writing refinement, grammatical editing, and code debugging during the research process. All research claims, interpretations, experimental design, implementation details, and scientific content remain the sole responsibility of the authors.

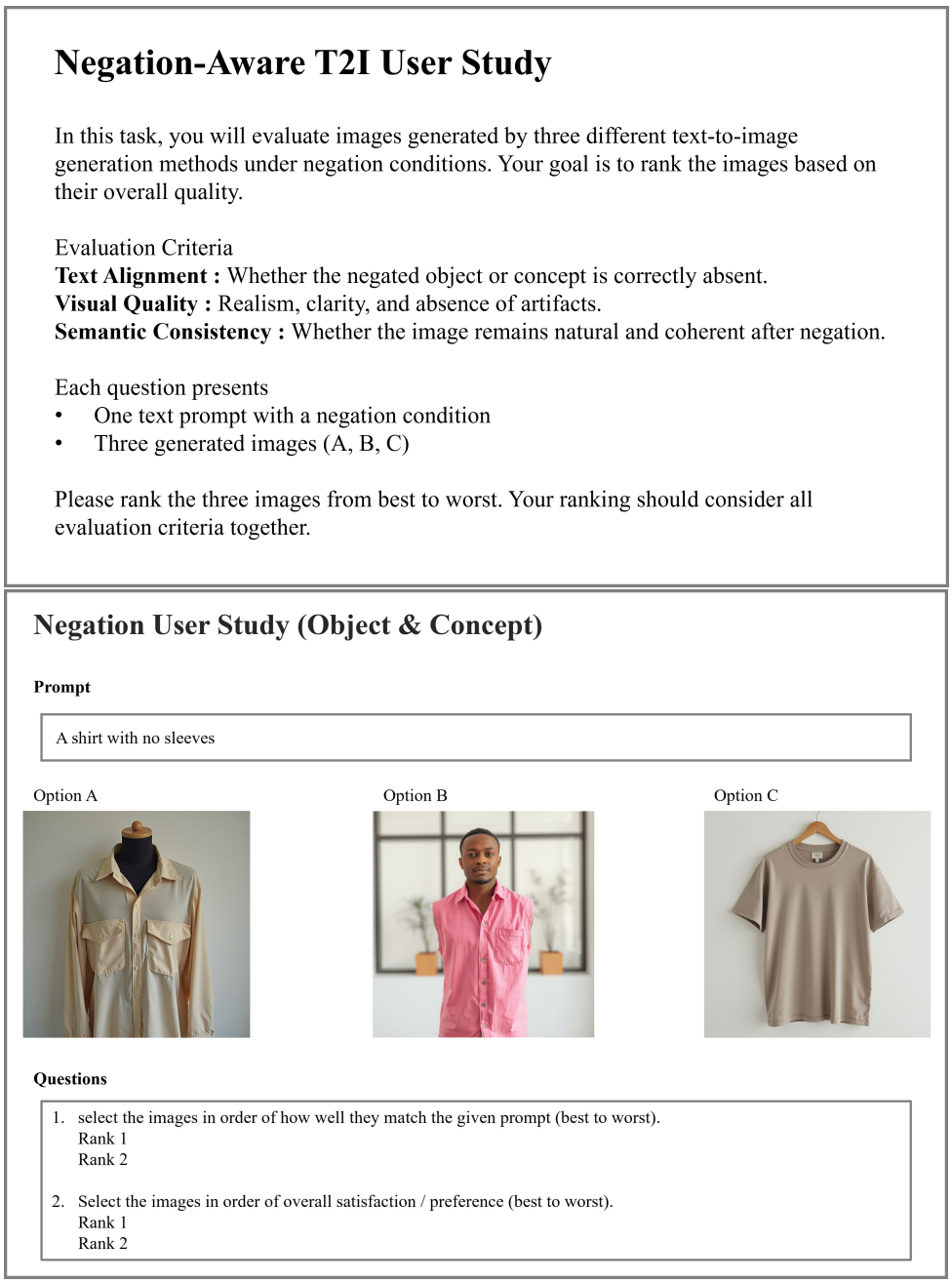


Figure 17: Example of the negation-aware user study. Participants compare generated images from three methods and evaluate negation fidelity and overall visual quality.

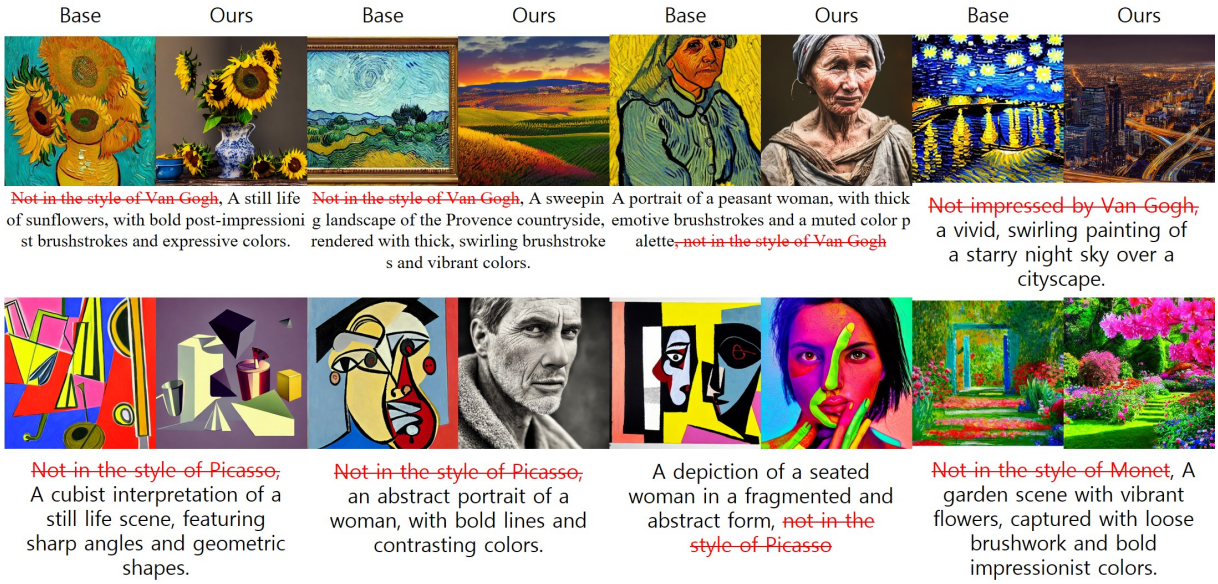


Figure 18: Qualitative examples on SD v1.4 on concept negation.

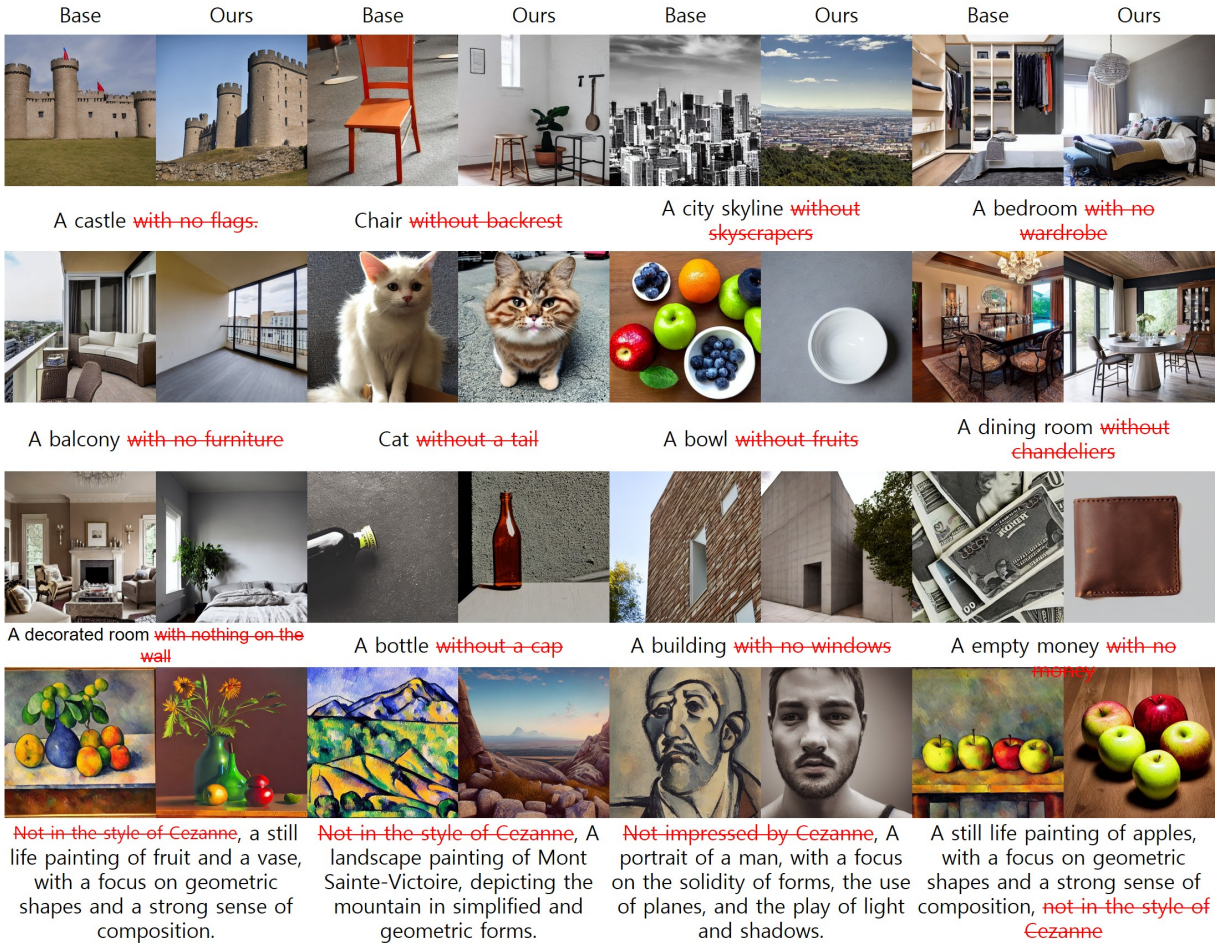


Figure 19: Qualitative results on SD v1.5 for object and concept negation.

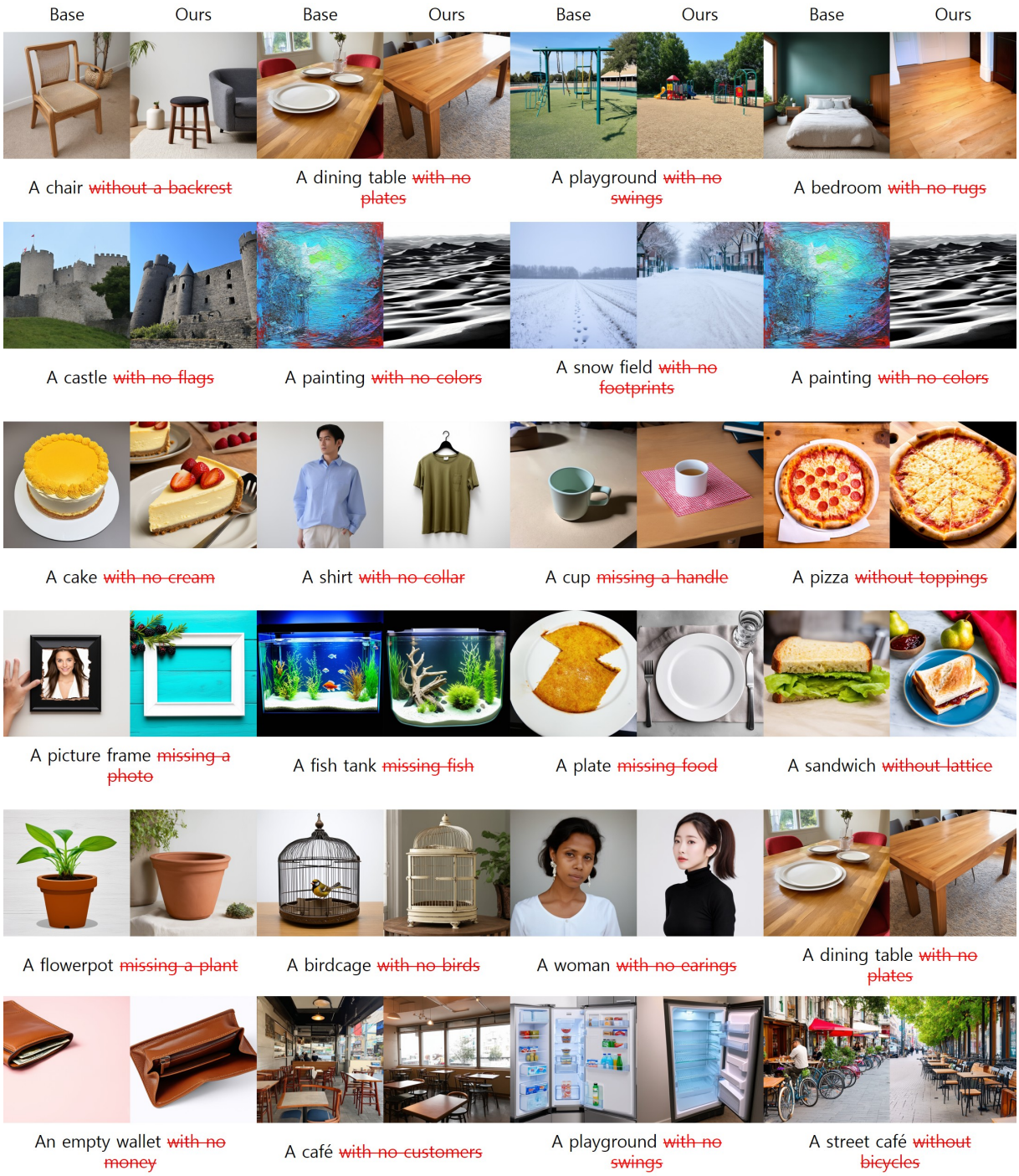


Figure 20: Object negation examples on SD v3.

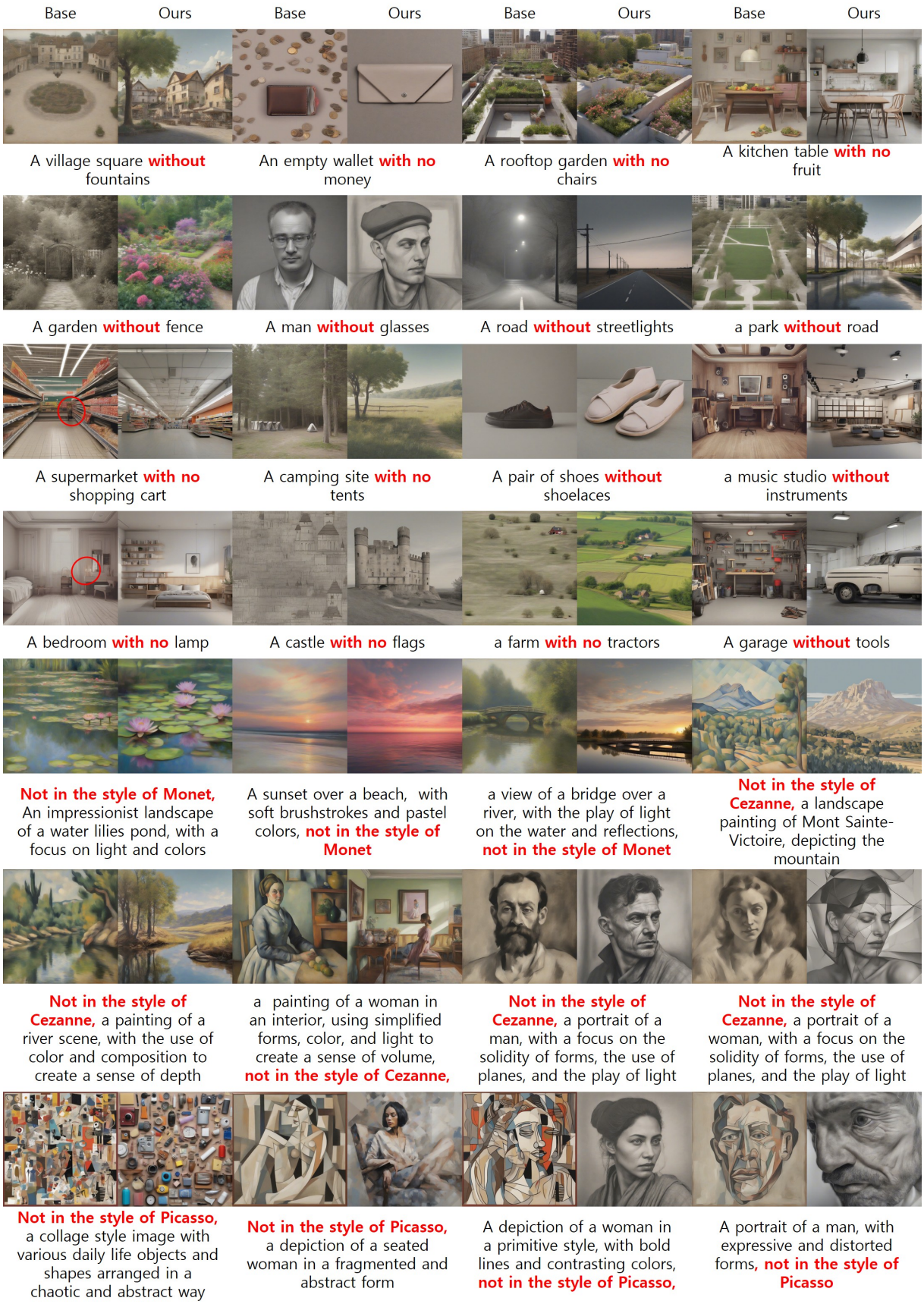


Figure 21: Object and artistic concept negation results on SDXL.

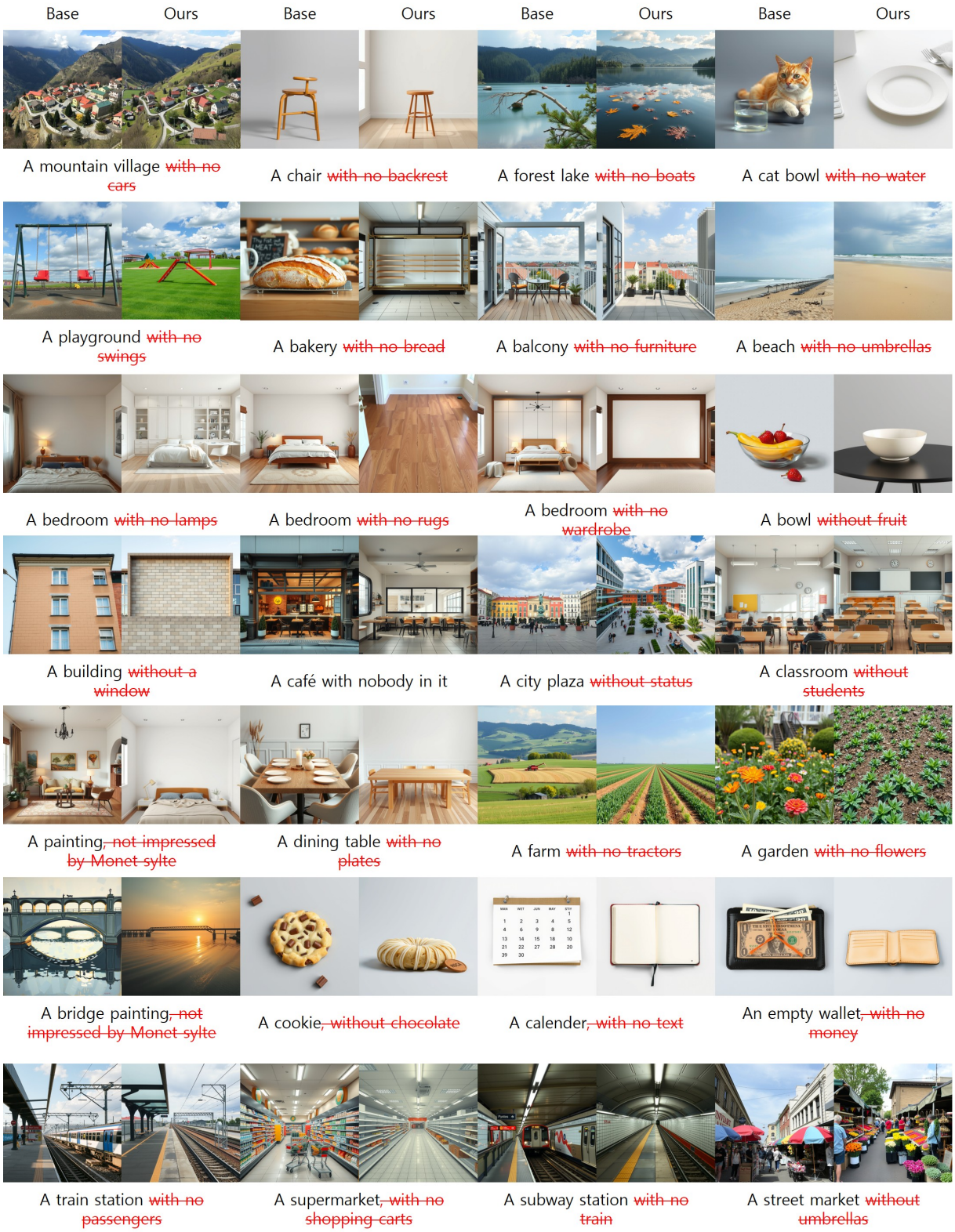


Figure 22: Object and concept negation results on FLUX.1-schnell.

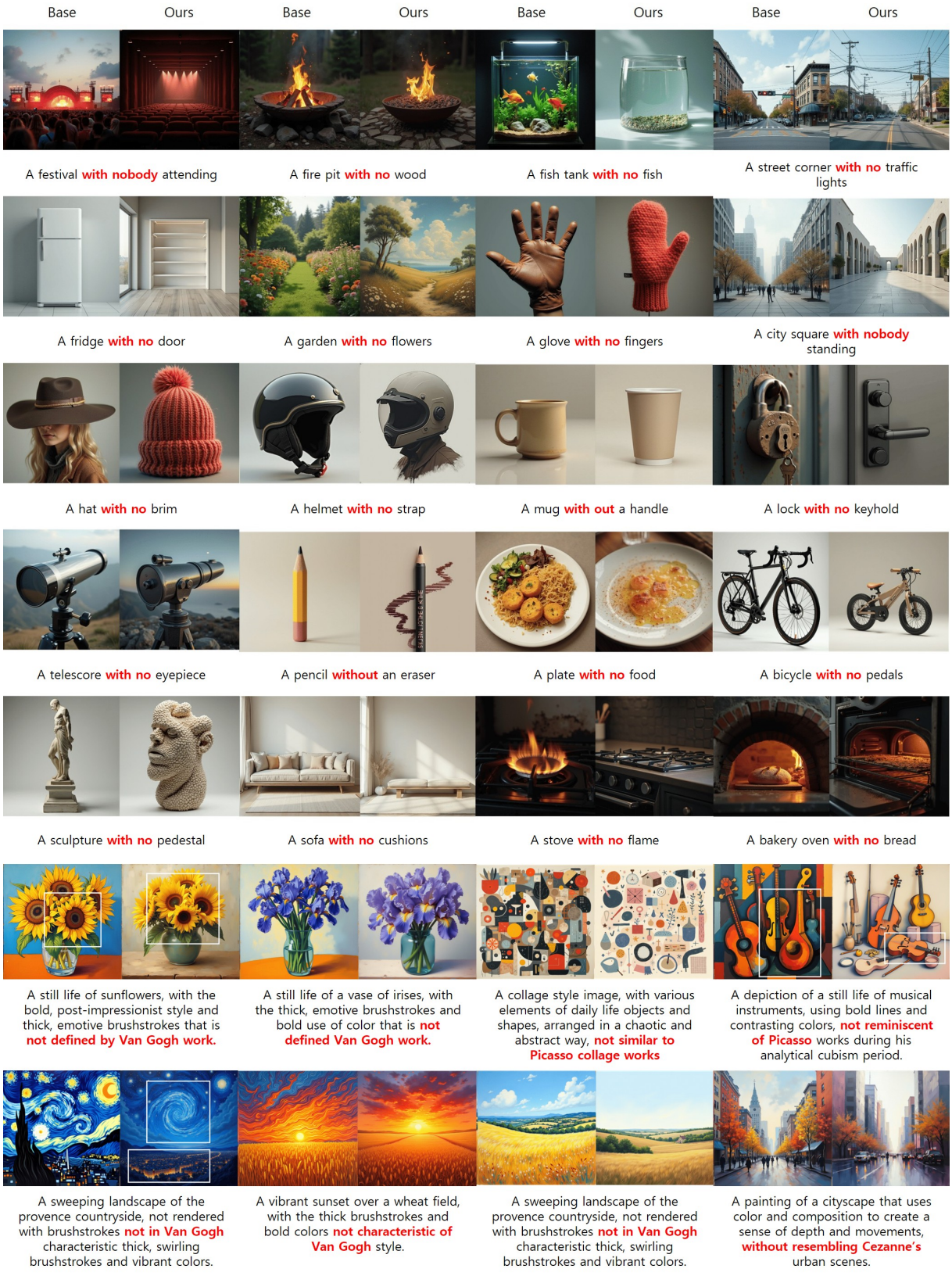


Figure 23: Object and artistic concept negation results on FLUX.1.