Mechanistic Dissection of Cross-Attention Subspaces in Text-to-Image Diffusion Models

Jun-Hyun Bae, Wonyong Jo, Jaehyup Lee, Heechul Jung

Kyungpook National University
{junhyun.bae, whdnjsdyd111, jaehyuplee, heechul}@knu.ac.kr

Abstract

Text-to-image diffusion models utilize cross-attention to integrate textual information into the visual latent space, yet the transformation from text embeddings to latent features remains largely unexplored. We provide a mechanistic analysis of the output-value (OV) circuits within cross-attention layers through spectral analysis via singular value decomposition. Our analysis demonstrates that semantic concepts are encoded in low-dimensional subspaces spanned by singular vectors in OV circuits across cross-attention heads. To verify this, we intervene on concept-related components in the diffusion process, demonstrating that intervention on identified spectral components affects conceptual changes. We further validate these findings by examining visual outputs of isolated subspaces and their alignment with text embedding space. Through this mechanistic understanding, we demonstrate that simply nullifying these spectral components can achieve targeted concept removal with performance comparable to existing methods while providing interpretability. Our work reveals how cross-attention layers encode semantic concepts in spectral subspaces of OV circuits, providing mechanistic insights and enabling precise concept manipulation without retraining.

Code — https://github.com/JunhyunB/diffusion-ov-circuits

Introduction

Diffusion-based text-to-image models have demonstrated remarkable capabilities in generating diverse and high-fidelity imagery from natural language descriptions (Saharia et al. 2022; Rombach et al. 2022). The core mechanism that integrates textual information into the visual latent space is cross-attention. At each denoising step, latent features attend to embeddings from the text encoder to determine how to refine the image. While previous work has extensively studied cross-attention maps to understand how concepts appear in the generated image (Tang et al. 2023; Liu et al. 2024; Park et al. 2025), the semantic transformation from text embeddings to visual features remains largely unexplored. This gap motivates the central question of our study: how do cross-attention layers internally translate text to visual features?

We approach this question through the lens of mechanistic interpretability, which analyzes neural networks by decomposing them into interpretable components (Elhage et al. 2021; Dunefsky, Chlenski, and Nanda 2024). This decomposition reveals that each cross-attention layer operates through two independent circuits: the query-key (QK) circuit and the output-value (OV) circuit. The QK circuit computes attention weights to determine which text tokens influence each spatial position, while the OV circuit transforms the selected tokens' semantic content into visual features. Our focus on the OV circuit stems from a key observation: while the QK circuit controls spatial allocation of attention, the actual semantic-to-visual translation occurs entirely within the OV transformation.

Text embeddings from CLIP and other encoders are known to organize semantic information along intrinsic axes in their high-dimensional space (Yu et al. 2024; Vennam et al. 2024). Since OV matrices perform a direct linear mapping from this semantically-structured text space to the visual residual stream, we hypothesize that OV transformations learn subspaces that align with these semantic axes, enabling coherent translation of textual concepts into visual features. Building on this insight, we apply spectral analysis to OV matrices through singular value decomposition (SVD). This technique decomposes each attention head's OV matrix into a set of orthogonal components, providing a principled framework for analyzing the internal structure of OV matrices. By examining how text embeddings project onto these components, we can identify the internal structure of semantic-to-visual translation.

Our analysis uncovers that semantic concepts are not uniformly distributed across the cross-attention layers, but instead concentrate in specific subspaces spanned by a subset of singular vectors. Systematic modulation of identified spectral components produces corresponding semantic changes in generated imagery, providing causal evidence that the discovered subspaces encode functionally relevant semantic information. Individual concepts are found to occupy limited fractions of the complete spectral space, with distinct organizational patterns observed across different conceptual categories, indicating a compositional architecture underlying semantic encoding.

The mechanistic insights gained from this analysis enable principled interventions into models. Selective nullification of concept-specific spectral components achieves targeted semantic suppression with efficacy comparable to existing

methodologies, while providing interpretability of the underlying transformations. Unlike black-box approaches that require additional training or optimization, our method directly manipulates the model's internal representations based on mechanistic understanding. Our work demonstrates that mechanistic interpretability provides practical insights for model control, achieving competitive performance with existing methods while offering transparency into the underlying mechanisms.

Our contributions are threefold:

- We present a systematic spectral analysis of output-value transformations in cross-attention mechanisms, demonstrating that semantic concepts organize into low-dimensional subspaces with characteristic patterns of spectral concentration.
- We establish a comprehensive framework for characterizing the internal structure of semantic encoding through spectral decomposition, demonstrating that concepts map to subspaces of varying dimensionality in a structured and interpretable manner.
- We demonstrate that targeted manipulation of identified spectral components enables transparent concept control with performance comparable to existing methods, thereby validating the functional relevance of our mechanistic analysis and its practical applicability.

Related Work

Understanding the internal mechanisms of diffusion models has become crucial for text-to-image generation tasks. Tang et al. (2023) focused on analyzing cross-attention maps to understand where concepts appear in generated images. Liu et al. (2024) advanced this understanding by analyzing distinctive functions of cross-attention and self-attention. Recently, Park et al. (2025) introduced head relevance vectors to quantify the importance of individual attention heads in encoding semantic information. However, these approaches mainly analyze attention maps rather than understanding how semantic information is transformed within the model.

Complementary work has explored temporal dynamics and concept manipulation within diffusion models. Zhang et al. (2024a) showed that cross-attention outputs converge to fixed points early in the denoising process, dividing generation into semantic planning and fidelity improvement stages. Gandikota et al. (2024a) proposed low-rank updates for controlling individual concepts, while Hertz et al. (2023) demonstrated how manipulating cross-attention enables text-guided image editing. These findings suggest a structured organization of semantic information, yet the fundamental transformation mechanism remains unclear.

Recent work has explored alternative approaches to understanding diffusion models with sparse autoencoders (SAEs) (Surkov et al. 2024; Cywiński and Deja 2025; Tian et al. 2025; Shi et al. 2025). While SAEs aim to identify interpretable features, they require additional training and lack clear guidelines on where to apply them within the model architecture. In contrast, we directly dissect the internal structure of cross-attention layers through spectral analysis without auxiliary models.

To understand the fundamental mechanisms of attention, we build upon the mechanistic interpretability framework in transformer models. Early work on *Transformer Circuits* formalized the decomposition of each attention layer into independent QK and OV circuits, establishing a foundation for circuit-level analysis of large models (Elhage et al. 2021). Our work extends the circuit perspective by applying spectral analysis directly to OV transformations in cross-attention, indicating that semantic concepts are encoded in low-dimensional subspaces.

Mechanistic Foundations

Cross-Attention in Diffusion Models

In diffusion models, the U-Net architecture maintains a residual stream, a high-dimensional vector space that attention layers read from and write to. Cross-attention layers serve as the primary mechanism for integrating textual information into this visual residual stream. At each spatial position and resolution, cross-attention reads text embeddings from a pretrained text encoder and writes semantic information that accumulates through the network depth. Following conventional notation for dot-product attention, we use row-wise notation. Specifically, the cross-attention operation computes

$$\text{CrossAttn}(\mathbf{z}, \mathbf{c}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}\mathbf{W}_\text{O}, \qquad (1)$$

where queries $\mathbf{Q} = \mathbf{z}\mathbf{W}_\text{Q}$ are projected from spatial features $\mathbf{z} \in \mathbb{R}^{N\times C}$ with N spatial positions, while keys $\mathbf{K} = \mathbf{c}\mathbf{W}_\text{K}$ and values $\mathbf{V} = \mathbf{c}\mathbf{W}_\text{V}$ are projected from text embeddings $\mathbf{c} \in \mathbb{R}^{L\times d_\text{text}}$ with L tokens. This operation transforms linguistic concepts into visual features that guide the denoising process. To understand how this transformation encodes semantic information, we decompose the operation into its constituent parts.

Circuit Decomposition of Cross-Attention

The standard view treats cross-attention as a monolithic operation, but examining its mathematical structure reveals two functionally distinct computations. Consider the complete cross-attention contribution to the residual stream:

$$\Delta\mathbf{z} = \text{softmax}\left(\frac{(\mathbf{z}\mathbf{W}_\text{Q})(\mathbf{c}\mathbf{W}_\text{K})^T}{\sqrt{d_k}}\right)\cdot(\mathbf{c}\mathbf{W}_\text{V}\mathbf{W}_\text{O}). \qquad (2)$$

This naturally factorizes into two circuits with distinct computational roles:

$$\Delta\mathbf{z} = \underbrace{\mathbf{A}(\mathbf{z},\mathbf{c})}_{\text{QK Circuit}}\cdot\underbrace{\mathbf{c}\mathbf{W}_\text{OV}}_{\text{OV Circuit}}. \qquad (3)$$

The QK circuit computes attention weights $\mathbf{A}(\mathbf{z},\mathbf{c}) = \text{softmax}((\mathbf{z}\mathbf{W}_\text{Q})(\mathbf{c}\mathbf{W}_\text{K})^T/\sqrt{d_k})$. This circuit determines spatial attention allocation: given the current visual state $\mathbf{z}$, which text tokens should influence each spatial position. Importantly, this computation is entirely independent of the semantic content being transmitted.

The OV circuit performs a linear transformation $\mathbf{c}\mathbf{W}_\text{OV}$ where $\mathbf{W}_\text{OV} = \mathbf{W}_\text{V}\mathbf{W}_\text{O}$. This circuit encodes semantic content: given text embeddings, what visual features to generate.

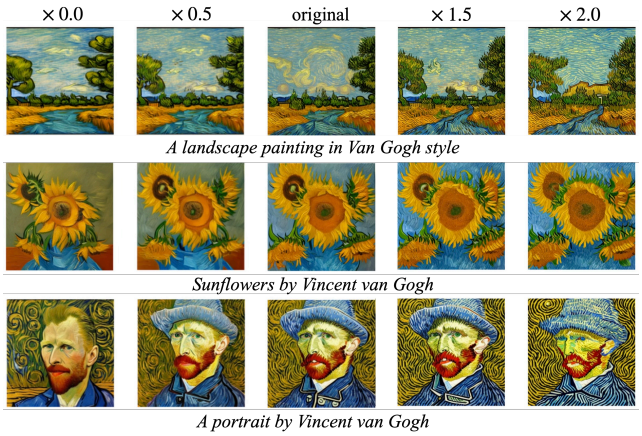

Figure 1: Effect of modulating outputs of high-contribution heads for the "Vincent van Gogh" concept. We select the top-20 heads (about 10.3% out of 195 heads) across all cross-attention layers in Stable Diffusion v2.1 based on their $\rho^{(h)}_{\text{concept}}$ values.

Crucially, this transformation is independent of spatial attention patterns. While attention patterns vary throughout the denoising process as visual features evolve, the OV transformation itself remains fixed, making it particularly amenable to static analysis.

Spectral Analysis of OV Matrix

Modern diffusion models use multi-head cross-attention, where multiple attention heads operate in parallel within each layer. Each head h performs its own OV transformation $\mathbf{W}^{(h)}_{\text{OV}}$, implementing a specialized mapping from text to visual features. While prior work has shown functional specialization across heads (Park et al. 2025), the internal structure of these transformations remains unexplored.

To uncover this structure, we perform spectral decomposition via Singular Value Decomposition (SVD) on each head's OV matrix:

$$\mathbf{W}^{(h)}_{\text{OV}} = \sum_{i=1}^{r} \sigma^{(h)}_i \mathbf{u}^{(h)}_i (\mathbf{v}^{(h)}_i)^T, \tag{4}$$

where $r = \min(d_{\text{text}}, d_{\text{head}})$ is the rank. The left singular vectors $\{\mathbf{u}^{(h)}_i\}$ form an orthonormal basis in text embedding space, while the right singular vectors $\{\mathbf{v}^{(h)}_i\}$ define corresponding directions in the head's output space. The singular values $\{\sigma^{(h)}_i\}$ determine the strength of each mapping. When text embeddings have high projection onto specific $\mathbf{u}^{(h)}_i$ directions, the corresponding output features along $\mathbf{v}^{(h)}_i$ are amplified in the transformation.

Identifying Concept-Contributing Heads

Each head's left singular vectors $\{\mathbf{u}^{(h)}_i\}$ form an orthonormal basis in text embedding space. For a given embedding $\bar{\mathbf{c}}$, the transformation through head h produces an intermediate representation that directly determines the head's contribution to the residual stream.

For each prompt, the text encoder produces a sequence of token embeddings which we aggregate into a single vector $\bar{\mathbf{c}} \in \mathbb{R}^{d_{\text{text}}}$ via mean pooling across the token dimension. We define the spectral representation of embedding $\bar{\mathbf{c}}$ for head h:

$$\mathbf{s}^{(h)}(\bar{\mathbf{c}}) = \bar{\mathbf{c}}\mathbf{U}^{(h)}\mathbf{\Sigma}^{(h)} \in \mathbb{R}^{d_{\text{head}}}. \tag{5}$$

Here, $\mathbf{U}^{(h)}$ and $\mathbf{\Sigma}^{(h)}$ denote the matrices of left singular vectors and singular values from the SVD of $\mathbf{W}^{(h)}_{\text{OV}}$. This representation captures how the text embedding projects onto head h's singular vector basis, weighted by the corresponding singular values. Crucially, this vector directly determines the head's output through multiplication with $(\mathbf{V}^{(h)})^T$, where $\mathbf{V}^{(h)}$ is the matrix of right singular vectors. This output contributes to the residual stream after being weighted by attention scores.

To identify heads that strongly respond to specific conceptual differences, we measure how each head's spectral representation changes between base prompts and their conceptual variations. Given a base prompt and its conceptual variation (e.g., "a mountain" vs. "a mountain in Van Gogh style"), we measure the change in each head's spectral representation:

$$\Delta\mathbf{s}^{(h)}_{\text{concept}} = \mathbf{s}^{(h)}(\bar{\mathbf{c}}_{\text{concept}}) - \mathbf{s}^{(h)}(\bar{\mathbf{c}}_{\text{base}}), \tag{6}$$

$$\rho^{(h)}_{\text{concept}} = \|\Delta\mathbf{s}^{(h)}_{\text{concept}}\|_2. \tag{7}$$

High values of $\rho^{(h)}_{\text{concept}}$ indicate that head h strongly amplifies the conceptual difference. Since $\Delta\mathbf{s}^{(h)}_{\text{concept}}$ directly translates to differences in the residual stream contribution via multiplication with $(\mathbf{V}^{(h)})^T$, heads with large $\rho^{(h)}_{\text{concept}}$ values are the primary contributors of concept-specific visual features.

Figure 1 demonstrates the effect of modulating high-contribution heads. For the top-k heads ranked by $\rho^{(h)}_{\text{concept}}$ values, we scale their contributions to the residual stream during generation:

$$\Delta\tilde{\mathbf{z}}^{(h)} = \alpha \cdot \Delta\mathbf{z}^{(h)}, \quad h \in \mathcal{H}_k, \tag{8}$$

where $\Delta\mathbf{z}^{(h)}$ is head h's original contribution to the residual stream, $\mathcal{H}_k$ denotes the set of top-k heads, and $\alpha \geq 0$ controls the intervention strength. Setting $\alpha = 0$ completely removes these heads' contributions, while $\alpha > 1$ amplifies their effect.

From Heads to Subspaces: Spectral Localization of Concepts

The previous analysis identifies concept-contributing heads but does not examine how concepts are organized within each head. Neural networks often exhibit polysemanticity, the phenomenon where individual components encode multiple, distinct features (Elhage et al. 2022; Scherlis et al. 2022). In the context of cross-attention, this raises the question of whether concepts fully utilize entire heads or localize to specific spectral components. This motivates us to move beyond head-level analysis to investigate how concepts are structured within the spectral geometry of these transformations.

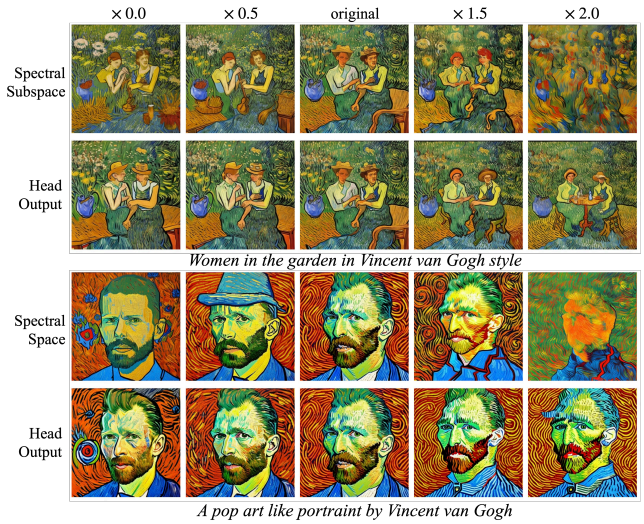

Figure 2: Spectral modulation versus head-level modulation. (Top) Scaling the identified spectral components (10%) in $\mathcal{S}_c$ versus (bottom) scaling outputs of top-10% high-contribution heads, with varying α.

Our spectral difference vector $\Delta\mathbf{s}^{(h)}_{\text{concept}} \in \mathbb{R}^{d_{\text{head}}}$ captures how head h's spectral representation differs between base and concept prompts. To identify which spectral components drive the high $\rho^{(h)}_{\text{concept}}$ values, we decompose the contribution of each singular vector. Since $\rho^{(h)}_{\text{concept}} = \|\Delta\mathbf{s}^{(h)}_{\text{concept}}\|_2$, we examine each element of $\Delta\mathbf{s}^{(h)}_{\text{concept}}$:

$$[\Delta\mathbf{s}^{(h)}_{\text{concept}}]_i = \sigma^{(h)}_i \left(\langle \bar{\mathbf{c}}_{\text{concept}}, \mathbf{u}^{(h)}_i \rangle - \langle \bar{\mathbf{c}}_{\text{base}}, \mathbf{u}^{(h)}_i \rangle \right), \quad (9)$$

where large $|[\Delta\mathbf{s}^{(h)}_{\text{concept}}]_i|$ values indicate that the i-th singular vector contributes significantly to the head's overall concept response $\rho^{(h)}_{\text{concept}}$. Since the residual stream contribution from head h is $\Delta\mathbf{s}^{(h)}_{\text{concept}}(\mathbf{V}^{(h)})^T$, each singular component contributes $[\Delta\mathbf{s}^{(h)}_{\text{concept}}]_i \mathbf{v}^{(h)}_i$ to the output.

Having identified significant components within high-contribution heads, we now extend our analysis to all heads in the model. Some heads with low overall $\rho^{(h)}_{\text{concept}}$ values might still contain individual components with strong concept encoding. To capture all concept-relevant spectral components across the model, we rank all singular components by their absolute contributions $|[\Delta\mathbf{s}^{(h)}_{\text{concept}}]_i|$ and select the top-k% to form the concept's spectral signature $\mathcal{S}_c$.

The selected components $\mathcal{S}_c$ enable targeted concept manipulation through spectral modulation. We modify the OV matrices by scaling only these identified components:

$$\widetilde{\mathbf{W}}^{(h)}_{\text{OV}} = \mathbf{W}^{(h)}_{\text{OV}} + (\alpha - 1) \sum_{i:(h,i)\in\mathcal{S}_c} \sigma^{(h)}_i \mathbf{u}^{(h)}_i (\mathbf{v}^{(h)}_i)^T, \quad (10)$$

where $\alpha \geq 0$ controls the modulation strength. Setting $\alpha = 0$ nullifies the selected components (removing the concept), $\alpha = 1$ preserves the original model, and $\alpha > 1$ amplifies the concept's encoding.

Figure 2 illustrates a fundamental difference in how concepts are organized within attention heads. These results demonstrate that semantic concepts can be localized to specific spectral spaces, enabling precise identification and manipulation of concept-encoding components through spectral analysis.

Experimental Results

Having established that semantic concepts concentrate in specific spectral components of the OV matrices, we now test how this dissection of cross-attention subspaces affects model behavior. We do this through qualitative and quantitative experiments designed to verify whether manipulating those components produces the predicted changes.

We first probe the identified subspaces in Stable Diffusion v2.1 by selectively removing or amplifying the spectral components. Through this spectral modulation, we observe how the cross-attention mechanism changes when specific components are activated, thereby confirming that these subspaces represent the pathways by which concepts are encoded.

We then validate the effectiveness of our mechanistic findings through concept removal experiments, demonstrating that spectral manipulation achieves practical results comparable to existing methods. These findings validate that spectral analysis provides a principled framework for understanding how diffusion models encode semantics, bridging mechanistic insights with practical applications.

Analyzing Concept-Related Spectral Subspaces

A key finding emerges when generating images using only concept-related spectral components. Figure 3 illustrates generated images across different concept types including artistic styles (*Van Gogh*, *Picasso*), content attributes (*nudity*), and lighting conditions (*sunset*, *neon*) for varying dimensionalities of concept-related subspaces, $\mathcal{S}_c$.

To understand what semantic information these components encode, we generate images using only the spectral components in $\mathcal{S}_c$ while setting all other components to zero. The results uncover the semantic primitives underlying each concept. For artistic styles, these components encode pure stylistic patterns separated from scene content. The *neon light* concept demonstrates particularly notable behavior. The spectral components selectively capture luminescent boundaries and edge glows characteristic of neon signage, producing ethereal outlines where neon illumination would typically appear in urban scenes.

In contrast to style and lighting concepts, nudity-related concepts exhibit different encoding characteristics. Rather than decomposing into visual patterns, these spectral components reconstruct complete human forms with contextual details preserved. This distinction suggests that content-based concepts require more holistic representation compared to stylistic attributes.

Nevertheless, the effectiveness of spectral modulation remains consistent. Removing the top-15-20% of concept-related spectral vectors successfully eliminates the concept from generated images. As shown in Figure 3, when these components are excluded during generation, the model produces fully clothed figures or empty scenes while maintaining

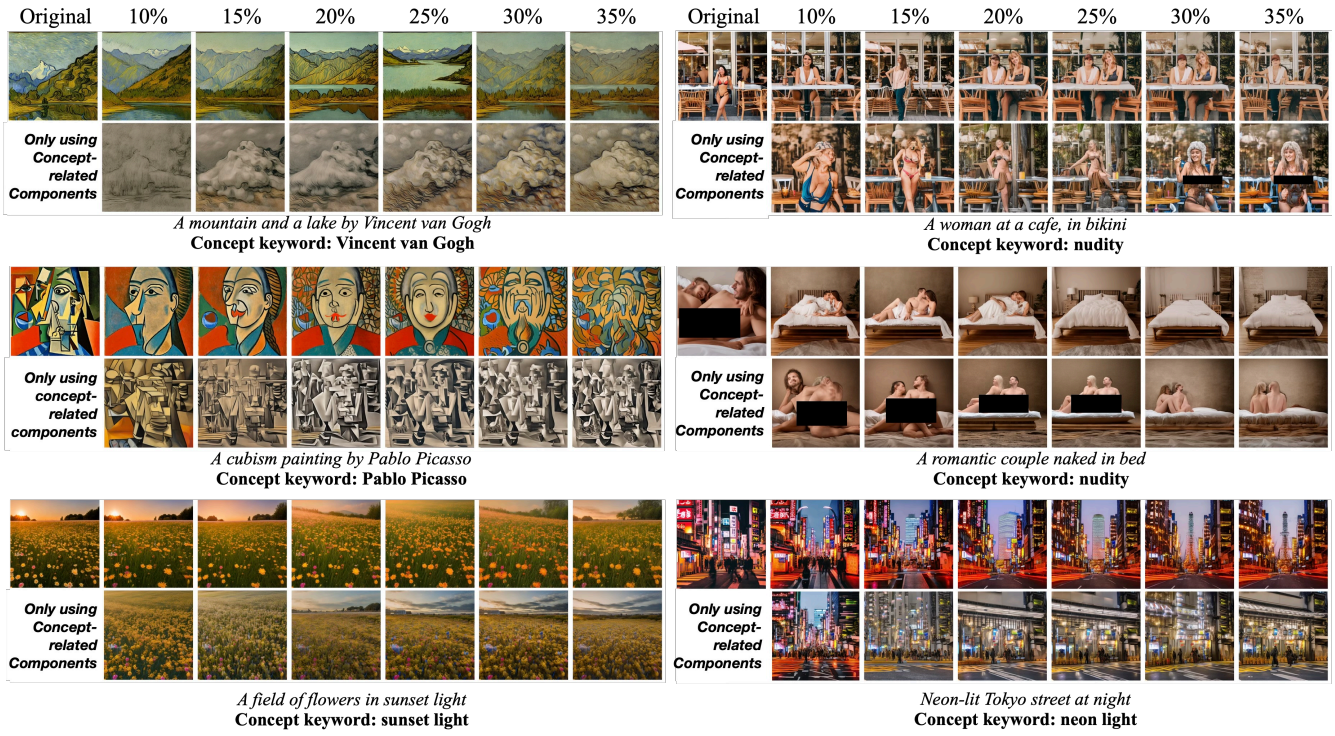

Figure 3: Spectral isolation of semantic concepts. For each concept, we show generated images with concept components removed (top) and with only concept components activated (bottom), using the top-$k\%$ of components from $\mathcal{S}_c$.

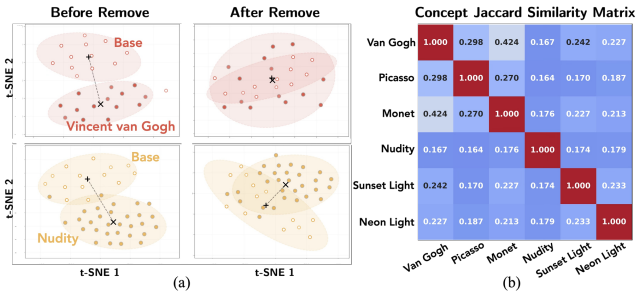

Figure 4: (a) t-SNE visualization of head outputs $\Delta\mathbf{z}^{(h)}$ before and after removing top-10% concept-related spectral components. (b) Jaccard similarity between top-10% spectral component sets selected for different concepts.

all other visual attributes. This demonstrates that despite the complex semantics for the *nudity* concept, content-based concepts still encode in compact spectral subspaces. However, removing more than 30% of spectral components can degrade the overall quality of generated images.

To validate that the identified spectral components encode semantic concepts, we examine how their removal affects the residual stream contributions. Figure 4(a) shows a t-SNE visualization of head outputs $\Delta\mathbf{z}^{(h)}$, before and after removing concept-related spectral components. Before removal, clear separation exists between base and concept prompt representations. After removal, these clusters overlap, demonstrating that the heads no longer distinguish between the two prompts. This confirms that our method precisely identifies the components responsible for concept encoding.

We also examine the organization of concept subspaces by analyzing overlap between different concepts' spectral signatures. Figure 4(b) presents Jaccard similarity between singular vectors selected for different concepts. The relatively low off-diagonal values show that each concept primarily occupies its own spectral territory, though complete separation is not observed. The observed overlap between related artistic styles (*Van Gogh* and *Monet*) indicates that concept encoding operates through combinatorial assembly of shared spectral components. Individual singular vectors participate in multiple concept representations, yet each concept retains a unique spectral signature defined by its specific component combination. This mechanism represents an efficient encoding strategy wherein semantic similarity translates to partial overlap in spectral space while maintaining sufficient separation for concept disambiguation.

Figure 5 illustrates the spectral distribution of concept encoding within high-contribution heads. We analyze heads exhibiting consistently high $\rho^{(h)}_{\text{concept}}$ values across three artistic style concepts: *Van Gogh*, *Monet*, and *Picasso*. The distribution of contribution scores $|[\Delta\mathbf{s}^{(h)}_{\text{concept}}]_i|$ demonstrates that each concept shows characteristic patterns of singular vector activation. Furthermore, the singular vectors exhibiting substantial contributions are not confined to those with the largest singular values (low indices), but are distributed throughout

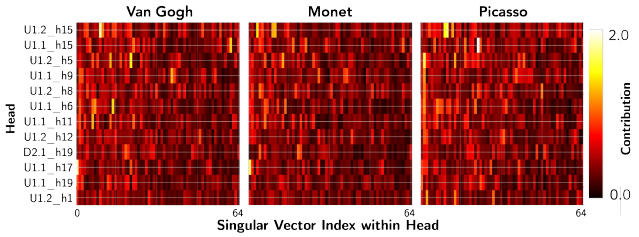

Figure 5: Distribution of spectral contributions across singular vectors for artistic style concepts. For heads with high $\rho^{(h)}_{\text{concept}}$ values shared across *Van Gogh*, *Monet*, and *Picasso* concepts, we visualize $|[\Delta \mathbf{s}^{(h)}_{\text{concept}}]_i|$ for each singular vector i. Different concepts activate distinct patterns of singular vectors within the same heads, demonstrating concept-specific spectral localization.

the spectral range in concept-specific configurations. These findings indicate that within individual attention heads, distinct concepts exhibit differentiated activation patterns across dominant singular vectors, demonstrating how diffusion models achieve efficient multi-concept encoding through partially overlapping but distinguishable spectral patterns.

Semantic Content of Spectral Subspaces

We examine the semantic information encoded in concept-specific subspaces $\mathcal{S}_c$ by analyzing their alignment with vocabulary tokens through a reconstruction-based approach. We first isolate the semantic direction in text embedding space by computing difference vectors between concept and base prompts:

$$\Delta \mathbf{c} = \bar{\mathbf{c}}_{\text{concept}} - \bar{\mathbf{c}}_{\text{base}}. \tag{11}$$

The difference vector $\Delta \mathbf{c}$ contains various semantic changes. By projecting this vector onto $\mathcal{S}_c$ and reconstructing it, we can observe what semantic information these spectral components actually encode.

For the top-k heads ranked by $\rho^{(h)}_{\text{concept}}$ values, we perform head-specific reconstruction of the semantic difference vector. For each head h, we project $\Delta \mathbf{c}$ onto the subspace spanned by its concept-related singular vectors:

$$\Delta \hat{\mathbf{c}}^{(h)} = \sum_{i:(h,i)\in\mathcal{S}_c} \langle \Delta \mathbf{c}, \mathbf{u}^{(h)}_i \rangle \mathbf{u}^{(h)}_i. \tag{12}$$

This reconstruction is analogous to SVD-based image compression, where reconstructing with only the top singular components preserves the main visual content while discarding high-frequency noise. By reconstructing $\Delta \mathbf{c}$ using only the components $(h, i) \in \mathcal{S}_c$, we extract the portion of the semantic difference that these spectral components actually encode. The resulting $\Delta \hat{\mathbf{c}}^{(h)}$ for head h represents only the semantic information encoded by those selected components of head h. Under the hypothesis that $\mathcal{S}_c$ captures concept-relevant information, concept-relevant tokens should rank among the highest in similarity to these reconstructed vectors.

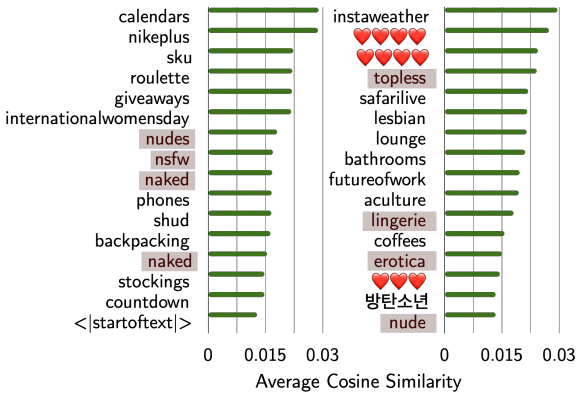

Figure 6: Tokens most aligned with nudity-related spectral components. Each panel displays the 16 tokens with highest cosine similarity to concept-specific singular vectors from one of the top-2 heads (selected by $\rho^{(h)}_{\text{nudity}}$ values). Highlighting indicates explicitly nudity-related tokens.

We focus our analysis on the top-2 heads with highest $\rho^{(h)}_{\text{concept}}$ values. For each head, we compute cosine similarities between $\Delta \hat{\mathbf{c}}^{(h)}$ and all 49,408 CLIP token embeddings, averaging across multiple prompt pairs for robustness. We experiment with the *nudity* concept and only use top-10% of *nudity*-related singular vectors.

Figure 6 illustrates what semantic information these spectral components encode by showing tokens with highest alignment to the reconstructed direction. The identified components primarily encode explicit nudity-related semantics (”nude”, ”naked”, ”topless”, ”erotica”, ”nsfw”), confirming their functional role in generating such content.

This reconstruction-based analysis provides two critical insights. First, concept-relevant tokens consistently rank highest in similarity despite the high dimensionality of the embedding space, confirming that our identified spectral components encode the target semantic information. Second, the presence of adjacent tokens shows the polysemantic nature of these subspaces, where concept-specific components capture semantic neighborhoods rather than perfectly isolated concepts.

Effectiveness of Spectral Nullification

Our mechanistic findings motivate a critical test of whether effective concept removal can be achieved exclusively through manipulation of spectral components, without any additional model retraining. We refer to our approach as **Spectral Nullification (SN)**, which nullifies identified concept-related spectral components during the generation process. This evaluation addresses the practical implications of our analysis by comparing SN against recently proposed concept removal methods.

To verify that the dissected subspaces truly mediate concept generation, we conduct nudity-removal experiments on five benchmark datasets with adversarial prompts, including Ring-A-Bell (Tsai et al. 2024), I2P (Schramowski et al. 2023),

Method	Ring-A-Bell			I2P	MMA	P4D	UnLearn
	K16	K38	K77				
SDv1.4	97.89	94.74	87.37	25.03	68.10	69.76	50.70
ESD	76.84	78.95	74.74	13.04	24.80	50.24	26.06
CA	88.42	88.42	84.21	19.30	58.50	63.41	44.37
MACE	89.47	95.79	93.68	25.56	66.00	68.29	50.70
SDID	95.79	91.58	84.21	23.12	62.00	66.83	48.59
UCE	22.11	18.95	21.05	8.06	41.00	38.05	21.13
RECE	**10.53**	**9.47**	7.37	4.24	25.00	21.46	9.15
SLD-Medium	68.42	60.00	50.53	8.38	48.70	43.90	23.94
SLD-Strong	18.95	10.53	**6.32**	**2.33**	**7.70**	**11.71**	**7.04**
SAFREE	65.26	55.79	45.26	6.26	29.90	38.54	14.79
SN (Ours)	41.05	35.79	30.53	4.24	17.60	18.54	8.45

Table 1: Evaluation of concept removal methods on nudity-related content. We report ASR across different benchmarks. Best results are shown in **bold** and second-best results are underlined. Highlight colors denote method types: training-based (gray), closed-form update (blue), and inference-time guiding (green).

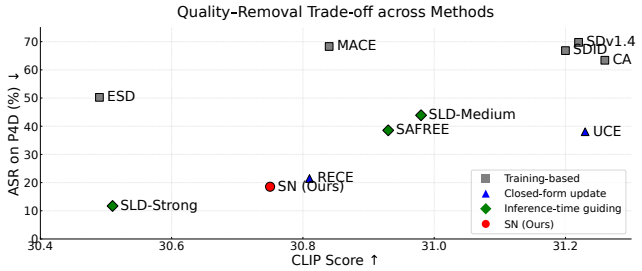

Figure 7: Quality-removal trade-off across different concept removal methods on P4D benchmark and COCO dataset. Lower ASR indicates better concept removal, while higher CLIP score indicates better preservation of generation quality.

MMA-Diffusion (Yang et al. 2024), P4D (Chin et al. 2024), and UnlearnDiffAtk (Zhang et al. 2024b). These datasets provide comprehensive coverage of prompts designed to elicit inappropriate content, serving as a rigorous test for concept removal effectiveness. For images generated using these adversarial prompts, we employ pretrained NudeNet (Bedapudi 2019) with threshold 0.6 to detect inappropriate content and compute Attack Success Rate (ASR), where lower ASR indicates better concept removal.

We compare against three categories of baseline methods. Training-based approaches include ESD (Gandikota et al. 2023), CA (Kumari et al. 2023), MACE (Lu et al. 2024), SDID (Li et al. 2024), which modify model parameters through additional training. Closed-form methods include UCE (Gandikota et al. 2024b) and RECE (Gong et al. 2024), which update model weights analytically without iterative training. Inference-time methods include SLD (Schramowski et al. 2023) and SAFREE (Yoon et al. 2025), which guide the generation process without retraining. All experiments are conducted on Stable Diffusion v1.4 to ensure fair comparison across methods. For our implementation, we nullify the top-20% of concept-related spectral components, based on our observation that this range effectively removes concepts while preserving generation quality.

Figure 8: Comparison of generated images for the I2P adversarial prompts (sample #1045 and #3147).

Table 1 provides empirical validation of our mechanistic analysis. If our hypothesis is correct that concepts are encoded in specific spectral components that flow through the residual stream, then nullifying these components should prevent concept-related visual features from being generated. The results confirm this prediction as removing identified spectral components significantly reduces inappropriate content generation across all benchmarks.

Our method's performance varies across datasets, with detection rates ranging from 4.24% on I2P to 41.05% on Ring-A-Bell K16, reflecting differences in adversarial prompt design and attack sophistication across benchmarks. However, the consistent suppression across all benchmarks suggests that our method identifies fundamental pathways through which inappropriate concepts flow from text embeddings to visual outputs.

To evaluate general image generation quality after concept removal, we test each method on 1,000 randomly sampled COCO captions (Lin et al. 2014), and measure semantic alignment using CLIP scores (Hessel et al. 2021). Figure 7 illustrates the inherent trade-off between concept removal and generation quality across different method categories. Our method occupies a competitive position in this trade-off space, demonstrating that mechanistic identification and targeted nullification of concept-encoding spectral components provides a principled approach to concept control. Figure 8 provides qualitative results supporting these quantitative results.

Conclusion

In this work, we studied how cross-attention layers encode textual semantics through spectral analysis of OV circuits. Our key finding is that semantic concepts localize to distinct spectral subspaces across cross-attention heads. Intervention on these subspaces enables interpretable concept manipulation in the diffusion process. Furthermore, we introduced a reconstruction-based approach to probe the semantic content of spectral subspaces in OV matrices, providing a systematic method to interpret what each identified component encodes. We demonstrated the practical utility of these insights through Spectral Nullification, achieving concept removal performance comparable to existing methods without model training, while providing transparency into the underlying mechanisms.

Acknowledgments

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2025-02283048, Developing the Next-Generation General AI with Reliability, Ethics, and Adaptability, 50%) and the Core Research Institute Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A1A03043144, 50%).

References

Bedapudi, P. 2019. Nudenet: Neural nets for nudity classification, detection and selective censoring.

Chin, Z.-Y.; Jiang, C. M.; Huang, C.-C.; Chen, P.-Y.; and Chiu, W.-C. 2024. Prompting4Debugging: Red-Teaming Text-to-Image Diffusion Models by Finding Problematic Prompts. In *Forty-first International Conference on Machine Learning*.

Cywiński, B.; and Deja, K. 2025. SAeUron: Interpretable Concept Unlearning in Diffusion Models with Sparse Autoencoders. In *Forty-second International Conference on Machine Learning*.

Dunefsky, J.; Chlenski, P.; and Nanda, N. 2024. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37: 24375–24410.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; Grosse, R.; McCandlish, S.; Kaplan, J.; Amodei, D.; Wattenberg, M.; and Olah, C. 2022. Toy Models of Superposition. *Transformer Circuits Thread*.

Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; DasSarma, N.; Drain, D.; Ganguli, D.; Hatfield-Dodds, Z.; Hernandez, D.; Jones, A.; Kernion, J.; Lovitt, L.; Ndousse, K.; Amodei, D.; Brown, T.; Clark, J.; Kaplan, J.; McCandlish, S.; and Olah, C. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2426–2436.

Gandikota, R.; Materzyńska, J.; Zhou, T.; Torralba, A.; and Bau, D. 2024a. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, 172–188. Springer.

Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024b. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5111–5120.

Gong, C.; Chen, K.; Wei, Z.; Chen, J.; and Jiang, Y.-G. 2024. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, 73–88. Springer.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-or, D. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations*.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528.

Kumari, N.; Zhang, B.; Wang, S.-Y.; Shechtman, E.; Zhang, R.; and Zhu, J.-Y. 2023. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22691–22702.

Li, H.; Shen, C.; Torr, P.; Tresp, V.; and Gu, J. 2024. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12006–12016.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, B.; Wang, C.; Cao, T.; Jia, K.; and Huang, J. 2024. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7817–7826.

Lu, S.; Wang, Z.; Li, L.; Liu, Y.; and Kong, A. W.-K. 2024. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6430–6440.

Park, J.; Ko, J.; Byun, D.; Suh, J.; and Rhee, W. 2025. Cross-Attention Head Position Patterns Can Align with Human Visual Concepts in Text-to-Image Generative Models. In *The Thirteenth International Conference on Learning Representations*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Scherlis, A.; Sachan, K.; Jermyn, A. S.; Benton, J.; and Shlegeris, B. 2022. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22522–22531.

Shi, Y.; Li, C.; Wang, Y.; Zhao, Y.; Pang, A.; Yang, S.; Yu, J.; and Ren, K. 2025. Dissecting and mitigating diffusion bias via mechanistic interpretability. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8192–8202.

Surkov, V.; Wendler, C.; Mari, A.; Terekhov, M.; Deschenaux, J.; West, R.; Gulcehre, C.; and Bau, D. 2024. One-Step is Enough: Sparse Autoencoders for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2410.22366*.

Tang, R.; Liu, L.; Pandey, A.; Jiang, Z.; Yang, G.; Kumar, K.; Stenetorp, P.; Lin, J.; and Türe, F. 2023. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5644–5659.

Tian, Z.; Nan, S.; Xu, M.; Zhai, S.; Qu, W.; Liu, J.; Jia, R.; and Zhang, J. 2025. Sparse autoencoder as a zero-shot classifier for concept erasing in text-to-image diffusion models. *arXiv preprint arXiv:2503.09446*.

Tsai, Y.-L.; Hsu, C.-Y.; Xie, C.; Lin, C.-H.; Chen, J. Y.; Li, B.; Chen, P.-Y.; Yu, C.-M.; and Huang, C.-Y. 2024. Ring-A-Bell! How Reliable are Concept Removal Methods For Diffusion Models? In *The Twelfth International Conference on Learning Representations*.

Vennam, S.; Singh, S.; Govil, A.; and Kumaraguru, P. 2024. Emergence of Text Semantics in CLIP Image Encoders. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*.

Yang, Y.; Gao, R.; Wang, X.; Ho, T.-Y.; Xu, N.; and Xu, Q. 2024. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7737–7746.

Yoon, J.; Yu, S.; Patil, V.; Yao, H.; and Bansal, M. 2025. SAFREE: Training-Free and Adaptive Guard for Safe Text-to-Image And Video Generation. In *The Thirteenth International Conference on Learning Representations*.

Yu, H.; Luo, H.; Wang, F.; and Zhao, F. 2024. Uncovering the text embedding in text-to-image diffusion models. *arXiv preprint arXiv:2404.01154*.

Zhang, W.; Liu, H.; Xie, J.; Faccio, F.; Shou, M. Z.; and Schmidhuber, J. 2024a. Cross-attention makes inference cumbersome in text-to-image diffusion models. *arXiv e-prints*, arXiv–2404.

Zhang, Y.; Jia, J.; Chen, X.; Chen, A.; Zhang, Y.; Liu, J.; Ding, K.; and Liu, S. 2024b. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, 385–403. Springer.