EraseSAE: Surgical Concept Erasure in Text-to-Video Diffusion Models via Sparse Autoencoders

Xinghao Wang[1,3*], Dong Li[2†], Wei Yu[2], Yingwei Pan[2], Tao Gong[1,3†], Qi Chu[1,3], Nenghai Yu[1,3], and Ting Yao[2]

[1] University of Science and Technology of China, Hefei, China
[2] HiDream.ai Inc.
[3] Anhui Province Key Laboratory of Digital Security, China
wxhwxhwxh@mail.ustc.edu.cn, {lidong,yuwei,pandy}@hidream.ai, {tgong,qchu,ynh}@ustc.edu.cn, tiyao@hidream.ai

Abstract. Recent advances in text-to-video (T2V) diffusion models have demonstrated remarkable generative capabilities, yet their reliance on loosely curated training data raises pressing safety and copyright concerns. Concept erasure offers a principled remedy by removing unwanted semantics from pretrained models while preserving remaining concepts. However, existing approaches typically operate at a coarse granularity misaligned with the fine-grained, distributed nature of concept representations, leading to incomplete removal or degraded generation quality. We argue that surgical erasure fundamentally requires intervention at the level of monosemantic features, where each unit encodes a single interpretable concept. To this end, we propose *EraseSAE*, a novel framework that leverages sparse autoencoders to achieve surgical concept erasure in DiT-based T2V diffusion models via a principled *decompose-attribute-erase* pipeline. We first introduce the Partitioned Convolutional Sparse Autoencoder, which decomposes dense spatiotemporal activations into disentangled, interpretable sparse features while preserving spatiotemporal coherence. A contrastive attribution mechanism then contrasts activations from paired prompts to isolate concept-specific feature kernels. At inference, timestep-resolved spatiotemporal masks derived from the identified kernels confine erasure to regions where the target concept is active, leaving unrelated content intact. Extensive experiments across diverse diffusion models and concept erasure tasks demonstrate that *EraseSAE* achieves precise and robust concept removal with minimal quality degradation, substantially outperforming state-of-the-art methods. The code is available at https://github.com/HiDream-ai/EraseSAE.

Keywords: Concept Erasure · Sparse Autoencoders · Diffusion Models

---

* This work was performed at HiDream.ai. † Corresponding authors.

1 Introduction

The rapid advancement of T2V diffusion models [21,36] has fundamentally transformed video creation, enabling high-fidelity video synthesis from free-form text prompts. However, this powerful generative capability also introduces significant safety and copyright risks [19, 30]. Trained on massive and loosely curated datasets, these models can readily generate harmful content such as explicit material or deepfakes of public figures. Mitigating such unsafe outputs has emerged as a critical challenge for responsible deployment of T2V models. While retraining on filtered datasets is a straightforward solution [16], the prohibitive computational cost renders it impractical. Concept erasure [26,34,38] provides a viable alternative by selectively removing target semantics from a pretrained model while preserving its capacity to generate remaining content. However, as illustrated in Fig. 2, achieving surgical erasure that precisely removes target concepts without degrading broader generative capabilities remains an open problem.

Existing concept erasure methods, primarily adapted from the text-to-image (T2I) domain, generally fall into two paradigms: *training-free* and *training-based*. *Training-free* methods steer inference away from target concepts through techniques such as negative prompting [29], text embedding manipulation [25,34,38], or conditional guidance [30,39], without altering model weights. While computationally efficient, they merely suppress concepts at the surface level while leaving internal concept representations intact, rendering them vulnerable to adversarial attacks. *Training-based* methods seek stronger guarantees by permanently altering model parameters through global fine-tuning [10, 12, 22, 37], closed-form editing [26], or neuron pruning [35]. Although more robust to input manipulation, these methods suffer from a fundamental granularity mismatch: existing methods operate at a coarse granularity that fails to align with the fine-grained, distributed nature of concept representations within these models. In deep neural networks, concepts are distributed across *polysemantic* representations where individual neurons encode multiple concepts simultaneously. Intervening on such entangled units leads to three distinct failure modes: (i) global weight modifications cause catastrophic forgetting of unrelated concepts; (ii) layer-specific edits targeting only cross-attention projections leave concept traces in other pathways, enabling reconstruction under adversarial probing; and (iii) neuron-level pruning inflicts collateral damage on co-encoded benign concepts. Beyond this granularity mismatch, virtually all existing methods lack spatiotemporal locality: even when a target concept occupies only a local spatial region or temporal segment, current approaches apply erasure globally, unnecessarily degrading unrelated content (Fig. 2). This limitation is particularly acute in video generation, where 3D full-attention in DiT-based T2V models encodes concepts jointly across space and time, making them especially sensitive to coarse-grained intervention.

These limitations converge on a fundamental requirement: surgical concept erasure demands intervention at the level of *monosemantic* units, where each feature encodes a single, interpretable concept. Sparse autoencoders (SAEs) [28], recently developed for mechanistic interpretability of large language models (LLMs) [8,31], provide exactly such a substrate by decomposing high-dimensional

activations into sparse linear combinations of monosemantic features. This property offers three key advantages that directly address the above limitations. First, projecting dense activations into a disentangled sparse space enables isolating specific features without disturbing others, resolving the polysemanticity problem that plagues neuron-level pruning. Second, SAE features capture distributed representations spanning arbitrary network components, overcoming the incomplete coverage of layer-specific editing. Third, because SAE activations naturally vary across spatial positions and frames, suppressing a target feature automatically confines erasure to where the concept is active, thereby achieving spatiotemporal locality that global fine-tuning cannot provide. Despite these advantages, transferring SAEs from LLMs to DiT-based T2V models presents substantial challenges. Conventional linear SAEs neglect local dependencies across neighboring positions and frames, risking degradation of spatiotemporal coherence. Furthermore, whereas LLM activations maintain stable semantics once generated, diffusion features progressively transition from coarse structure to fine detail across denoising timesteps, rendering single-timestep attribution insufficient.

Motivated by these insights, we propose *EraseSAE*, a novel framework that leverages SAEs to achieve surgical concept erasure in DiT-based T2V diffusion models via a principled *decompose–attribute–erase* pipeline. In the *decompose* stage, we introduce the *Partitioned Convolutional Sparse Autoencoder (PConvSAE)* to address the challenge of spatiotemporal coupling. By employing multi-dimensional convolutions and a partitioned architecture, PConvSAE decomposes coupled spatiotemporal activations into interpretable, concept-specific sparse features while preserving spatial structure and temporal coherence. In the *attribute* stage, we feed paired prompts that include and exclude the target concept into the frozen diffusion model and contrast the resulting activation distributions within PConvSAE, isolating a compact set of feature kernels highly correlated with the target concept. In the *erase* stage, we derive spatiotemporal masks from the activation maps of the identified kernels at each denoising timestep, and selectively suppress the target concept by modifying PConvSAE features only within the masked regions. This strategy restricts interventions exclusively to regions where the target concept is active while preserving the original feature evolution elsewhere. By operating in the monosemantic feature space, *EraseSAE* unifies the strengths of both paradigms: it achieves the robust and permanent erasure of training-based methods while maintaining the surgical precision and minimal side effects that training-free methods aspire to but cannot guarantee. More broadly, our work establishes that mechanistic interpretability tools can serve as a foundation for precise and controllable generation in the video domain.

We summarize our main contributions as follows:

- We propose *EraseSAE*, the first framework to leverage SAEs for concept erasure in T2V diffusion models, which surgically removes target concepts through a *decompose–attribute–erase* pipeline in monosemantic feature space.
- We introduce PConvSAE, a convolutional sparse autoencoder with a partitioned architecture that decomposes dense visual representations into interpretable sparse features while preserving spatiotemporal coherence.

- We design a contrastive attribution mechanism that isolates concept-correlated features and derives timestep-resolved spatiotemporal masks, confining erasure to where target concepts are active while preserving non-target content.
- Extensive experiments across multiple T2V diffusion models and diverse concept erasure tasks demonstrate that *EraseSAE* achieves state-of-the-art erasure effectiveness while preserving generation quality, outperforming the strongest baseline by 34.5% in erasure accuracy and 8.3% in SSIM.

2 Related Works

2.1 Concept Erasure in Diffusion Models

Concept erasure has been extensively studied in T2I diffusion models [5,6,23,29] along two main paradigms. *Training-free* methods [24, 25, 27, 39] redirect inference without modifying model weights. SLD [30] introduces a safety guidance term to steer denoising away from unsafe content, while SAFREE [38] projects text embeddings orthogonally to toxic concept directions. Though efficient, these approaches suppress concepts only at the surface level, leaving internal representations vulnerable to adversarial attacks [7,15]. *Training-based* methods offer stronger guarantees by permanently altering model parameters. ESD [10] fine-tunes cross-attention layers with negative-prompt guidance, and CA [22] ablates target concepts by anchoring outputs to a reference distribution. UCE [11] unifies multi-concept editing via closed-form projection updates, while AdvUnlearn [40] and Receler [17] augment fine-tuning with adversarial training for improved robustness. MACE [26] further scales erasure to massive concept sets through closed-form parameter mapping, and EraseAnything [12] extends the paradigm to flow-matching architectures via bi-level optimization. However, these methods operate at a coarse granularity and apply erasure globally, lacking the locality needed to preserve unrelated content. Extension to T2V models remains nascent. VideoEraser [34] projects prompt embeddings away from unsafe directions, and T2VUnlearning [37] fine-tunes velocity predictions with negative guidance. However, the spatiotemporal entanglement inherent in video generation amplifies the inherited limitations from T2I methods, and no existing approach achieves fine-grained, localized erasure without degrading overall generation quality.

2.2 Sparse Autoencoders for Mechanistic Interpretability

SAEs decompose dense neural activations into sparse linear combinations of monosemantic features, providing a principled tool for mechanistic interpretability [28]. Originally developed for modeling biological vision, SAEs have recently shown notable success in interpreting large-scale transformer-based language models [8], with subsequent work scaling to frontier models [32] and enabling feature-level steering for controllable generation [31]. Recent efforts have extended SAEs to visual generative models. SAUCE [13] applies SAEs within autoregressive vision-language models to suppress undesirable features for selective

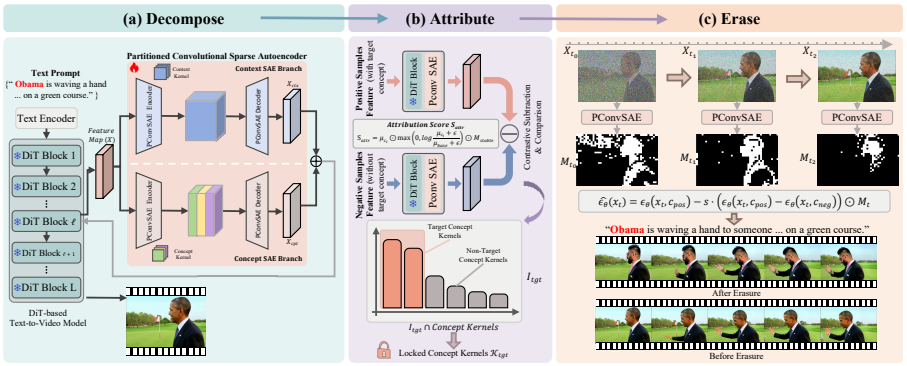

Fig. 1: Overview of *EraseSAE* framework. *EraseSAE* follows a *decompose–attribute–erase* pipeline. (a) *Decompose*: PConvSAE decomposes dense spatiotemporal activations into disentangled monosemantic features via a partitioned dual-branch convolutional architecture with spatial-aware activation. (b) *Attribute*: a contrastive log-ratio scoring mechanism isolates concept-specific feature kernels against hard-negative baselines in a one-time offline procedure. (c) *Erase*: timestep-resolved spatiotemporal masks derived from the locked kernels guide spatially-modulated classifier-free guidance, confining suppression to active concept regions while preserving unrelated content intact.

concept unlearning. SAeUron [9] integrates SAEs into image diffusion models, achieving interpretable concept erasure through feature-level interventions. However, no prior work has investigated SAEs in T2V diffusion models. Conventional MLP-based SAE architectures flatten high-dimensional hidden states into one-dimensional vectors, compromising the essential spatiotemporal locality inherent in video representations. Furthermore, standard SAEs lack mechanisms for rigorous semantic isolation across spatial regions, leading to concept leakage between target and non-target content. Our proposed PConvSAE addresses both limitations through a convolutional architecture that preserves full spatiotemporal structure and a partitioned design for high-purity feature disentanglement.

3 Methodology

3.1 Overview of *EraseSAE*

Given a pretrained DiT-based T2V diffusion model and a set of target concepts, *EraseSAE* surgically removes specified semantics from generated videos while preserving unrelated content. The framework follows a *decompose–attribute–erase* pipeline, as illustrated in Fig. 1. In the *decompose* stage (Sec. 3.2), we train a PConvSAE on intermediate activations at an identified intervention layer. Through multi-dimensional convolutions and a strictly partitioned dual-branch architecture, PConvSAE transforms dense spatiotemporal representations into disentangled monosemantic features, separating concept-specific semantics from general scene context. In the *attribute* stage (Sec. 3.3), a one-time contrastive

procedure contrasts activation distributions from paired prompts that include and exclude the target concept. Log-ratio scoring against hard-negative baselines isolates a compact set of concept-specific feature kernels with high semantic purity. In the *erase* stage (Sec. 3.4), the identified kernels produce timestep-resolved spatiotemporal masks that dynamically track the target concept throughout the denoising process. A spatially-modulated classifier-free guidance mechanism uses these masks to confine suppression exclusively to active concept regions.

3.2 Partitioned Convolutional Sparse Autoencoder

In DiT-based video diffusion models, intermediate hidden states encode densely entangled spatiotemporal representations where semantic content is distributed across spatial positions and temporal frames in a tightly coupled manner. Conventional MLP-based SAEs [9] flatten these high-dimensional hidden states into one-dimensional vectors prior to sparse decomposition, destroying the inherent spatial structure and temporal continuity of video features. This degrades reconstruction fidelity and impairs concept localization in both space and time.

To address these challenges, we first conduct hierarchical feature probing across transformer layers to identify the optimal intervention layer that captures rich target semantics while minimally affecting global structural fidelity. At this identified layer, we introduce PConvSAE, which replaces the standard linear projection with a stack of 2D convolutions operating on temporally folded activation tensors. By folding the temporal dimension into the batch axis and applying spatial convolutions directly on the resulting tensor, PConvSAE encodes and reconstructs features on their native topology, preserving the complete spatiotemporal structure without the information loss inherent in flattening operations.

Architecture Design. The central design principle of PConvSAE is the explicit decomposition of entangled visual representations into two complementary and functionally distinct components: concept-agnostic scene context and concept-specific semantics. Accordingly, PConvSAE partitions its latent space into two structurally decoupled computational branches: a *Context Branch* for encoding concept-agnostic scene representations f_{ctx}, and a *Concept Branch* for isolating concept-specific features f_{cpt}. The Context Branch captures global structural priors, background layout, illumination, and temporal dynamics that constitute the scene canvas, while the Concept Branch exclusively accommodates localized features semantically bound to target concepts. Together, they form a complete and non-redundant decomposition: suppressing Concept Branch features removes only target semantics while the Context Branch preserves all remaining visual content. Within the Concept Branch, we further enforce concept partitioning by dividing the channel dimension into C mutually exclusive subspaces (one per target concept, each with N_{cpt} feature kernels), preventing semantic interference during attribution and erasure.

Given the debiased hidden state $X \in \mathbb{R}^{(B\times T)\times D\times H\times W}$ obtained by subtracting the running mean of activations, both branches project X into a high-dimensional latent space via independent spatial convolutional encoders:

$$z_{\text{ctx}} = W_{\text{enc}}^{\text{ctx}} * X + b_{\text{enc}}^{\text{ctx}}, \quad z_{\text{cpt}} = W_{\text{enc}}^{\text{cpt}} * X + b_{\text{enc}}^{\text{cpt}} \tag{1}$$

where W_{enc} and b_{enc} denote the convolutional weights and biases of the respective encoders, and $*$ represents the convolution operation. The two encoders do not share parameters, ensuring entirely independent projection subspaces.

To impose sparsity while respecting the spatial nature of visual concepts, we introduce a Spatial-Aware Local Activation mechanism. Rather than applying a conventional channel-wise Top-K operator, we evaluate each channel c by its peak spatial response $s_c = \max_{h,w} z_{c,h,w}$ and select the index set $\mathcal{I}_{\text{top}}$ of the K most responsive channels. The sparse feature representation is then defined as:

$$f_{c,h,w} = \begin{cases} \text{ReLU}(z_{c,h,w}), & \text{if } c \in \mathcal{I}_{\text{top}} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where c, h, and w index the channel, height, and width, respectively. This formulation activates only the most salient channels while suppressing others across their entire spatial extent, producing spatially coherent sparse codes. The mechanism is applied independently to both branches with separate sparsity budgets, allowing the Context Branch to retain broader scene-level features while the Concept Branch maintains highly selective concept-correlated activations.

The activated features are projected back via respective convolutional decoders, and the overall reconstruction is obtained by additive superposition:

$$X_{\text{recon}} = W_{\text{dec}}^{\text{ctx}} * f_{\text{ctx}} + W_{\text{dec}}^{\text{cpt}} * f_{\text{cpt}} + b_{\text{dec}} \tag{3}$$

This additive formulation embodies the complementary principle: at erasure time, removing the Concept Branch contribution while retaining the Context Branch naturally yields a clean scene with the target concept excised.

Joint Optimization. Training PConvSAE requires guiding the two branches toward their designated complementary roles. We formulate a joint optimization objective supervised by spatial masks derived from the diffusion model's own cross-attention maps, eliminating the need for external annotations. Specifically, we extract cross-attention maps between visual hidden states and textual embeddings of the target concept [4,14], average and normalize them across selected layers, and apply dynamic quantile-based thresholding to yield a binarized target region mask $M_{\text{tgt}} \in \{0, 1\}$, with the background mask defined as its complement $M_{\text{bg}} = 1 - M_{\text{tgt}}$. Guided by this pair of mutually exclusive spatial priors, the overall optimization objective comprises five components.

The Context Reconstruction Loss ($\mathcal{L}_{\text{ctx}}$) compels the Context Branch to faithfully reconstruct the background region while being suppressed within the target spatial extent, steering it toward encoding only concept-agnostic content:

$$\mathcal{L}_{\text{ctx}} = \lambda_{\text{bg}} \frac{\|M_{\text{bg}} \odot (X_{\text{ctx}} - X)\|_2^2}{\sum M_{\text{bg}}} + \lambda_{\text{tgt}} \frac{\|M_{\text{tgt}} \odot X_{\text{ctx}}\|_2^2}{\sum M_{\text{tgt}}} \tag{4}$$

where $\odot$ denotes the Hadamard product and $\|\cdot\|_2^2$ represents the squared L_2 norm. The hyperparameters λ_{bg} and λ_{tgt} balance background reconstruction fidelity against target region suppression, and the denominators normalize each term by the effective area of its corresponding spatial mask.

The *Concept Reconstruction Loss* ($\mathcal{L}_{\text{cpt}}$) ensures the Concept Branch bears sole responsibility for encoding target region semantics. Since the Context Branch is suppressed within M_{tgt}, the combined output of both branches is constrained to recover the original hidden state within this region:

$$\mathcal{L}_{\text{cpt}} = \lambda_{\text{cpt}} \frac{\|M_{\text{tgt}} \odot (X_{\text{ctx}} + X_{\text{cpt}} - X)\|_2^2}{\sum M_{\text{tgt}}} \tag{5}$$

where λ_{cpt} weights the reconstruction fidelity within the target region. Together, $\mathcal{L}_{\text{ctx}}$ and $\mathcal{L}_{\text{cpt}}$ establish a complementary reconstruction protocol: the Context Branch covers the background while the Concept Branch covers the target region, and their union faithfully recovers the complete hidden state.

The *Identity Leakage Penalty* ($\mathcal{L}_{\text{leak}}$) enforces semantic purity across the concept dictionary. We construct a binary penalty tensor $\mathbf{P}$ that equals 0 for a concept's assigned kernels within M_{tgt} and 1 elsewhere:

$$\mathcal{L}_{\text{leak}} = \lambda_{\text{leak}} \frac{\|\mathbf{P} \odot f_{\text{cpt}}\|_1}{\sum \mathbf{P}} \tag{6}$$

where $\|\cdot\|_1$ denotes the L_1 norm, and λ_{leak} controls the penalty strength. This strict spatial-semantic cross-regularization ensures that concept-specific feature kernels are strictly confined to their designated scope, eliminating false activations on background pixels or unassociated concept samples.

The *Temporal Consistency Loss* ($\mathcal{L}_{\text{temp}}$) penalizes abrupt fluctuations in Context Branch activations across consecutive frames, since the scene canvas encoded by this branch must remain temporally stable to prevent flickering:

$$\mathcal{L}_{\text{temp}} = \lambda_{\text{temp}} \frac{\|f_{\text{ctx}}^{i} - f_{\text{ctx}}^{i-1}\|_2^2}{Z} \tag{7}$$

where Z is a constant equal to the total feature volume, i is the frame index.

The *Auxiliary Loss* ($\mathcal{L}_{\text{aux}}$) mitigates the dead latent problem [3] by encouraging inactive feature kernels across both branches to approximate the residual reconstruction error, ensuring high dictionary utilization. The overall training objective for PConvSAE is formulated as the sum of all five components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ctx}} + \mathcal{L}_{\text{cpt}} + \mathcal{L}_{\text{leak}} + \mathcal{L}_{\text{temp}} + \mathcal{L}_{\text{aux}} \tag{8}$$

Through this joint optimization, PConvSAE decomposes entangled spatiotemporal representations into a structured, interpretable feature dictionary where concept-agnostic scene context and concept-specific semantics are cleanly separated into complementary branches. This decomposition provides the foundation for precise concept attribution and targeted erasure in the subsequent stages.

3.3 Contrastive Attribution Mechanism

The structured latent space established by PConvSAE provides a foundation for precise concept identification. The remaining challenge is to determine *which*

kernels within the Concept Branch are intrinsically bound to each target concept. Performing this identification dynamically at every inference step would introduce prohibitive computational overhead. We therefore decompose the process into a one-time offline attribution phase that locks concept-specific kernels prior to deployment, followed by an online mask-guided erasure stage in Sec. 3.4.

The attribution phase aims to isolate, from the Concept Branch, the minimal subset of feature kernels that respond exclusively and consistently to a designated target concept c_i. We first collect a set of positive samples generated from prompts containing c_i and compute the spatial mean activation of each feature kernel strictly within the corresponding target region masks, yielding a per-kernel activation profile $\mu_{c_i} \in \mathbb{R}^{N_{\text{cpt}}}$. To ensure that the identified kernels are selective for c_i rather than responsive to background structures or semantically adjacent but distinct concepts, we construct a hard-negative baseline μ_{base} that aggregates all competing activation sources. Concretely, we compute μ_{bg}, the global mean activation across pure background samples, together with μ_{c_j} for every non-target concept c_j ($j \neq i$). The hard-negative baseline is then defined as the element-wise maximum over all competing signals:

$$\mu_{\text{base}} = \max\Big(\mu_{\text{bg}},\ \max_{j \neq i} \mu_{c_j}\Big) \tag{9}$$

This formulation enforces a *one-versus-max* exclusion principle: a kernel qualifies as concept-specific only if its response to c_i exceeds both the background activation level and the strongest response elicited by any alternative concept.

We quantify the concept specificity of each kernel through a contrastive log-ratio attribution score $\mathcal{S}_{\text{attr}}$:

$$\mathcal{S}_{\text{attr}} = \mu_{c_i} \odot \max\Big(0,\ \log \frac{\mu_{c_i} + \epsilon}{\mu_{\text{base}} + \epsilon}\Big) \odot M_{\text{stable}} \tag{10}$$

where ϵ is a small constant for numerical stability. The multiplicative weighting by μ_{c_i} ensures that kernels with higher absolute activation receive proportionally greater scores, favoring strongly responsive features over those that are selective yet weakly active. The binary mask M_{stable} suppresses transient activations by retaining only temporally robust kernels. To construct this mask, we compute the activation consistency of each kernel, defined as the fraction of positive samples in which the kernel successfully fires within the target region. Kernels whose consistency falls below a predefined robustness threshold τ are discarded.

The top-K kernels ranked by $\mathcal{S}_{\text{attr}}$ form an initial candidate set. To preserve the architectural semantic isolation enforced during PConvSAE training, we intersect this candidate set with the pre-allocated index partition $\mathcal{I}_{\text{tgt}}$ assigned to concept c_i in the channel partitioning scheme of the Concept Branch (Sec. 3.2). The resulting set $\mathcal{K}_{\text{tgt}}$ constitutes the definitively locked concept kernels. Because this offline attribution procedure is executed only once per target concept, it produces a compact kernel dictionary of high semantic purity while entirely eliminating the need for dynamic feature search during inference.

3.4 Concept Erasure with Dynamic Spatiotemporal Masks

Diffusion features progressively transition from coarse global structure to fine-grained local detail across denoising timesteps. This progressive evolution causes the spatial extent of a target concept to shift continuously throughout the reverse process. Applying a static mask derived from a single timestep would therefore introduce severe pixel-level misalignments and boundary artifacts as features drift. To maintain tight spatial correspondence between the mask and the evolving concept representation, we propose a dynamic mask generation and intervention mechanism that adaptively updates the erasure region at each timestep t.

Dynamic Spatiotemporal Mask Generation. Within a designated critical intervention interval, we pass the intermediate hidden states into the frozen PConvSAE at each timestep for online feature probing. From the Context Branch, we extract the aggregated activation heatmap and compute its spatial complement to obtain a dynamic foreground mask M_{fg}. Because the Context Branch is trained to encode concept-agnostic scene content (Eq. 4), its complement naturally highlights regions that fall outside the background canvas, providing a reliable foreground prior. Simultaneously, using the offline-locked kernel set $\mathcal{K}_{\text{tgt}}$, we extract the corresponding activation heatmap from the Concept Branch to produce the target concept mask M_{cpt}. The final intervention mask M_t is obtained as the Hadamard product of these two normalized components:

$$M_t = \text{Norm}(M_{\text{fg}}) \odot \text{Norm}(M_{\text{cpt}}) \tag{11}$$

where $\text{Norm}(\cdot)$ denotes min-max normalization. A dynamic threshold is subsequently applied to yield a binarized mask. The intersection of M_{fg} and M_{cpt} provides complementary spatial constraints: M_{fg} confines the mask to foreground regions, preventing spurious suppression of background content, while M_{cpt} further restricts it to locations where the target concept is semantically active. Because this mask is recomputed at every timestep, it adaptively tracks the spatial trajectory of the target concept across both frames and denoising stages.

Spatially-Modulated Classifier-Free Guidance. Equipped with the dynamic mask M_t, we integrate concept erasure directly into the classifier-free guidance (CFG) mechanism. Standard CFG applies guidance uniformly across all spatial positions, which inadvertently perturbs global illumination, background structure, and temporal dynamics when repurposed for concept suppression. We instead propose *spatially-modulated CFG*, which restricts negative guidance exclusively to the regions delineated by the mask:

$$\hat{\epsilon}_\theta(x_t) = \epsilon_\theta(x_t, c_{\text{pos}}) - s \cdot \big(\epsilon_\theta(x_t, c_{\text{pos}}) - \epsilon_\theta(x_t, c_{\text{neg}})\big) \odot M_t \tag{12}$$

Here $\epsilon_\theta(x_t, c_{\text{pos}})$ and $\epsilon_\theta(x_t, c_{\text{neg}})$ denote the noise predictions conditioned on the positive prompt containing the target concept and the negative prompt providing a safe semantic alternative, respectively, and s controls the erasure guidance scale. Within regions where $M_t \approx 1$, the guidance term steers the denoising trajectory away from the target semantics, imposing localized concept suppression. In regions where $M_t \approx 0$, the original positive-conditioned prediction is preserved without modification, maintaining the integrity of background content.

Table 1: Quantitative comparison of nudity erasure methods on CogVideoX and HunyuanVideo. Best and second best results are **bolded** and underlined.

Method	Nudity Rate (Gen) (↓)		Object Class (↑)		Subject Consistency (↑)		SSIM (↑)		Nudity Rate (Ring-A-Bell) (↓)						Inference Time (s/frame)	
									K16		K38		K77			
	CogX	Hunyuan	CogX	Hunyuan	CogX	Hunyuan	CogX	Hunyuan	CogX	Hunyuan	CogX	Hunyuan	CogX	Hunyuan	CogX	Hunyuan
Original	28.25	68.13	**79.91**	83.88	95.12	**96.12**	-	-	14.47	21.97	20.26	30.53	29.34	23.55	**3.61**	**3.55**
Neg Prompt	27.75	57.25	75.68	**87.59**	95.87	95.48	47.02	47.06	10.79	22.37	34.34	29.34	38.03	31.97	3.66	6.95
SAFREE [38]	5.75	46.38	45.19	71.47	94.49	95.33	42.39	41.78	15.13	36.05	16.58	31.71	15.66	19.74	3.72	3.59
VideoEraser [34]	18.88	-	73.86	-	**96.47**	-	44.36	-	7.11	-	18.03	-	23.68	-	4.03	-
T2VUnlearning [37]	2.88	10.89	51.98	77.33	93.55	95.96	42.39	60.67	5.30	5.13	7.60	7.82	8.65	9.95	3.67	3.60
Ours	**2.62**	**7.13**	77.80	80.61	94.30	96.04	**74.99**	**65.69**	**2.63**	**4.47**	**5.26**	**6.84**	**4.74**	**5.13**	3.66	6.97

4 Experiments

We evaluate *EraseSAE* across two concept erasure tasks, nudity erasure and celebrity identity erasure, on two DiT-based T2V models: CogVideoX-5B [36] and HunyuanVideo [21]. To further validate the generality of our framework beyond the video domain, we conduct additional nudity erasure experiments on the widely adopted T2I model Flux.1 [dev] [23]. Unless stated otherwise, PConvSAE is trained with the Adam optimizer [20] at an initial learning rate of 1×10^{-4} for 30 epochs on 8 NVIDIA A100 GPUs. More implementation details are provided in the supplementary material.

4.1 Nudity Erasure

Experimental Settings. Following [37], we construct a dataset of 500 nudity-related prompts, allocating 400 for training and 100 for testing, denoted as **Gen**. To evaluate defensive robustness against adversarial prompt manipulation, we additionally employ the adversarial prompt benchmark **Ring-A-Bell** [15]. For T2I, we conduct on Flux.1 [dev] using the **I2P** [30] dataset. All output videos are standardized to 32 frames at 8 fps. For quantitative evaluation, we employ **NudeNet** [1] to perform frame-by-frame detection and compute the target nudity exposure rate. To assess non-destructive fidelity, we adopt the Structural Similarity (**SSIM**) [33] to quantify pixel-level consistency against unaltered reference videos. We further incorporate the **VBench** [18] Object Class and Subject Consistency metrics to verify the coherent preservation of non-target semantics.

Quantitative Results. Tab. 1 presents comprehensive comparisons with state-of-the-art methods on the video nudity erasure task. Our method achieves the lowest nudity exposure rates on both CogVideoX-5B and HunyuanVideo, reducing the Gen-set exposure to 2.62 and 7.13, respectively. This consistent superiority over the strongest baseline, T2VUnlearning [37], validates the effectiveness of operating in the monosemantic feature space: because PConvSAE decomposes entangled activations into concept-specific sparse features, the resulting concept kernels capture the target semantics with high purity, enabling thorough suppression that weight-editing approaches struggle to achieve.

The advantage of *EraseSAE* is particularly pronounced under adversarial evaluation. On the Ring-A-Bell benchmark across all three difficulty levels (K16,

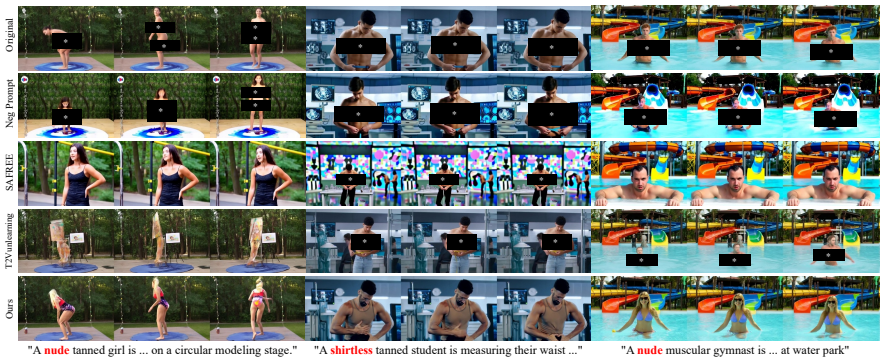

Fig. 2: Qualitative comparison of nudity erasure methods on HunyuanVideo.

K38, K77), our method maintains consistently low detection rates. We attribute this robustness to the contrastive attribution mechanism, which locks concept kernels through one-versus-max exclusion against hard negatives and thereby captures intrinsic concept semantics rather than surface-level prompt patterns. Consequently, adversarial prompt reformulations fail to circumvent the erasure because the underlying monosemantic features remain suppressed regardless of input phrasing. As shown in Tab. 3, *EraseSAE* further extends its efficacy to the T2I model Flux.1 [dev], achieving the lowest total nudity count on the challenging I2P benchmark and demonstrating the architectural generality.

Crucially, this strong erasure performance incurs minimal collateral damage. Our method achieves SSIM of 74.99 and 65.69 on CogVideoX-5B and HunyuanVideo, surpassing the next best method by 27.97 and 5.02 points. This gap reflects the spatiotemporal locality afforded by the dynamic mask mechanism, which confines interventions to active concept regions and preserves the pixel-level structure that global editing inevitably degrades. The VBench Object Class and Subject Consistency scores remain competitive with the unmodified models, confirming that PConvSAE's partitioned architecture disentangles concept-specific features from scene context. *EraseSAE* further introduces negligible inference overhead. On CogVideoX-5B its latency reaches 3.66 s/frame against 3.61 for the original model, while the marginal cost on HunyuanVideo originates solely from the additional negative-conditioned forward pass and stays on par with standard negative prompting. Because concept attribution is conducted entirely offline, no per-step feature search is required during inference.

Qualitative Results. Fig. 2 provides visual comparisons between *EraseSAE* and prior methods. Training-free approaches such as Neg Prompt and SAFREE [38] fail to completely suppress sensitive content. T2VUnlearning [37] achieves stronger erasure through weight editing, yet introduces visible artifacts in background regions and distorts non-target objects due to global parameter modifications. In contrast, *EraseSAE* cleanly removes the target concept while faithfully preserving unrelated visual content, demonstrating the practical benefit of spatially-modulated erasure guided by monosemantic feature maps.

Table 2: Quantitative results of celebrity erasure on CogVideoX and HunyuanVideo.

Method	Metric	Donald Trump		Barack Obama		Elon Musk		Queen Elizabeth		Taylor Swift		Average		Preserve ↑	
		CogX	Hunyuan	CogX	Hunyuan	CogX	Hunyuan	CogX	Hunyuan	CogX	Hunyuan	CogX	Hunyuan	CogX	Hunyuan
Original	Erase (↓)	77.50	87.50	72.50	75.00	49.50	53.50	68.50	73.00	49.00	46.50	63.40	67.10	-	-
Neg Prompt	Erase (↓)	63.50	59.00	69.00	53.00	24.50	32.50	29.50	50.00	26.00	41.50	42.50	47.20	**59.80**	60.20
	SSIM (↑)	31.30	35.31	32.58	35.31	33.57	43.76	27.44	33.02	32.58	40.13	31.49	37.51		
SAFREE [38]	Erase (↓)	62.50	**9.50**	67.00	13.50	17.50	53.00	21.50	68.00	19.50	50.50	37.60	47.20	48.40	47.00
	SSIM (↑)	30.74	31.40	30.37	30.17	34.08	39.54	27.26	32.36	31.87	32.26	30.86	33.14		
VideoEraser [34]	Erase (↓)	**3.00**	-	20.00	-	19.00	-	41.00	-	42.00	-	25.00	-	52.10	-
	SSIM (↑)	29.32	-	28.11	-	30.73	-	22.85	-	25.73	-	27.35	-		
Ours	Erase (↓)	12.00	26.50	**1.50**	**11.50**	**6.50**	**11.50**	**16.50**	**26.00**	**3.50**	**20.50**	**8.00**	**19.20**	58.60	**61.50**
	SSIM (↑)	**63.12**	**79.53**	**61.5**	**62.12**	**66.35**	**85.54**	**61.14**	**75.89**	**64.81**	**84.27**	**63.38**	**77.47**		

4.2 Celebrity Erasure

Experimental Settings. To evaluate identity protection capabilities, we conduct erasure experiments targeting five prominent public figures: *Donald Trump, Barack Obama, Elon Musk, Queen Elizabeth, and Taylor Swift*. Each category comprises 400 prompts, equally split between training and testing. Our evaluation encompasses three complementary assessments. The primary assessment evaluates standard identity erasure efficacy by measuring the residual detection accuracy of the target individual. The secondary assessment quantifies cross-identity interference: we apply the erasure intervention for a single target celebrity while concurrently generating videos of the remaining four celebrities, and compute the average detection accuracy of these unedited identities to measure their retention rate. The tertiary assessment employs SSIM to quantify the preservation of non-target backgrounds and general context. Target identity presence is measured using the Giphy celebrity detection algorithm [34].

Quantitative Results. Tab. 2 presents the comprehensive evaluation of celebrity identity erasure. Our framework reduces the average target detection accuracy to 8.00 and 19.20 on CogVideoX-5B and HunyuanVideo, respectively, demonstrating effective identity suppression across architectures and diverse facial characteristics. While the target identity is thoroughly suppressed, *EraseSAE* maintains average preservation accuracies of 58.60 and 61.50 for non-target identities on the two models, closely matching or exceeding the preservation scores of all baselines. This outcome validates the strict semantic exclusivity enforced by the one-versus-max attribution scoring (Eq. 10) and the channel partitioning scheme within the Concept Branch: the feature kernels locked for one identity reside in a mutually exclusive subspace from those encoding other identities, preventing cross-concept interference during erasure. In contrast, VideoEraser [34], despite achieving the lowest erasure rate for Donald Trump (3.00), attains only 52.10 preservation accuracy, indicating that its text-embedding manipulation inadvertently corrupts shared facial representations. The SSIM scores further underscore the non-destructive nature of our approach.

To validate the monosemantic property of the co-learned concept kernels, we present a cross-identity interference heatmap in Fig. 3. For each concept's dedicated kernel set, we compute the raw activation magnitude in response to prompts targeting each of the five celebrity identities and normalize the values to

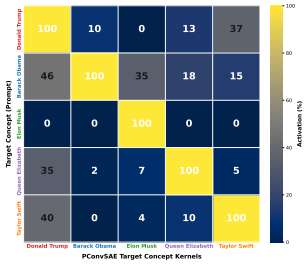

Fig. 3: Cross-identity interference heatmap of 5 celebrities.

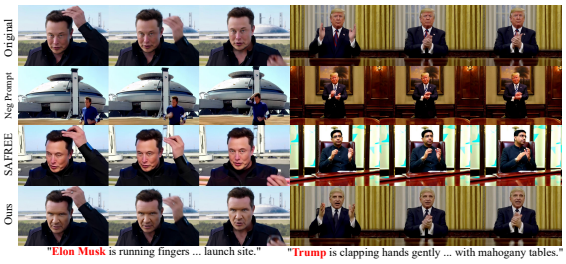

Fig. 4: Qualitative comparison of celebrity erasure methods on HunyuanVideo.

a percentage scale. The heatmap exhibits strong diagonal dominance, confirming that each kernel set responds selectively and exclusively to its designated target identity, with minimal cross-activation to other identities. Beyond cross-identity interference, we directly quantify the semantic purity of locked kernels [32]. We define the *purity* of class c as $P_c = \frac{1}{|\mathcal{K}_{\text{tgt}}^{(c)}|} \sum_{k \in \mathcal{K}_{\text{tgt}}^{(c)}} \text{Act}_c(k) / \sum_{c'=1}^{C} \text{Act}_{c'}(k)$, where $\text{Act}_c(k)$ is the spatiotemporal mean of the post-Top-K activation of kernel k over 100 samples from class c. Averaged over the 5 identities, PConvSAE attains $\bar{P} = 0.76$, more than twice the 0.37 of a linear SAE, confirming high monosemanticity.

Qualitative Results. Qualitative results in Fig. 4 corroborate the quantitative findings. Neg Prompt and SAFREE exhibit inconsistent erasure, with recognizable facial features of the target identity persisting across multiple frames. In contrast, *EraseSAE* completely suppresses the target celebrity's distinguishing features while faithfully preserving the fidelity of complex background.

4.3 Ablation Studies

All ablation studies are conducted on HunyuanVideo for the nudity erasure task.

Loss Components. Tab. 4 reports a cumulative ablation of the PConvSAE objectives. With only $\mathcal{L}_{\text{ctx}}$, the Concept Branch lacks explicit guidance for isolating target semantics, yielding a high nudity rate of 24.5. Sequentially introducing $\mathcal{L}_{\text{cpt}}$ and $\mathcal{L}_{\text{leak}}$ strengthens semantic separation and concept purity, reducing the rate to 8.09. Adding $\mathcal{L}_{\text{temp}}$ further raises SSIM by 18.32 points to 65.69, confirming its role in stabilizing temporal coherence.

Architectural Design. Tab. 5 evaluates three variants: (1) *Linear SAE* [9], the traditional linear sparse autoencoder extended to T2V activations; (2) *PConvSAE (single-branch)*, our convolutional encoder-decoder with a single undifferentiated branch; and (3) *PConvSAE (dual-branch)*, the full architecture with separate Context and Concept Branches. The Linear SAE flattens spatiotemporal activations into one-dimensional vectors, destroying local structure needed for precise localization and yielding the weakest results (12.76 nudity rate, 56.97 SSIM). The single-branch variant preserves spatial topology and improves both metrics, yet residual concept-context entanglement persists without explicit par-

Table 3: Quantity of explicit content (↓) with Flux.1 on I2P dataset.

Method	Common	Female	Male	Total
Original	406	161	38	605
UCE [11]	122	39	12	173
MACE [26]	173	55	28	256
EAP [2]	287	86	13	386
EraseAnything [12]	129	48	**22**	199
Ours	**102**	**31**	27	**160**

Table 4: Ablation study on the impact of each loss function design.

$\mathcal{L}_{\text{ctx}}$	$\mathcal{L}_{\text{cpt}}$	$\mathcal{L}_{\text{leak}}$	$\mathcal{L}_{\text{temp}}$	Nude Rate ↓	SSIM ↑
✓				24.5	40.12
✓	✓			10.21	43.33
✓	✓	✓		8.09	47.37
✓	✓	✓	✓	**7.13**	**65.69**

Table 5: Ablation study on different SAE architectural variants.

Method	Nude Rate ↓	SSIM ↑
Linear SAE [9]	12.76	56.97
PConvSAE (single-branch)	9.06	61.37
PConvSAE (dual-branch)	**7.13**	**65.69**

Table 6: Ablation study on different inference strategies.

Method	Nude Rate ↓	SSIM ↑
SAE-Sub	51.30	62.31
SAE-Mask	**6.56**	53.51
SM-CFG	7.13	**65.69**

titioning. The dual-branch PConvSAE attains the best performance on all metrics, demonstrating that separating concept-specific and context-agnostic representations is essential for both thorough erasure and high-fidelity preservation.

Inference Strategy. Tab. 6 compares three inference-time interventions. SAE-Sub directly subtracts target feature activations and suppresses concepts insufficiently (51.30 nudity rate). SAE-Mask zeros out all activations within the masked regions across layers; this aggressive operation lowers the nudity rate to 6.56 but corrupts the scene context encoded therein, degrading SSIM to 53.51. Our spatially-modulated classifier-free guidance (SM-CFG) attains the optimal trade-off between erasure thoroughness and structural preservation.

5 Conclusion

In this paper, we introduce *EraseSAE*, a pioneering framework for surgical concept erasure in T2V diffusion models. By shifting the intervention paradigm from entangled polysemantic neurons to disentangled monosemantic features, we resolve the granularity mismatch plaguing prior approaches. Our proposed PConvSAE effectively decomposes dense spatiotemporal activations into interpretable units while preserving essential structural coherence. Coupled with a novel contrastive attribution mechanism, our *decompose-attribute-erase* pipeline precisely isolates and suppresses target concepts exclusively within their active regions. Extensive evaluations demonstrate that *EraseSAE* achieves state-of-the-art erasure efficacy with minimal degradation to overall generation quality.

Acknowledgments

This work was supported by the Key Science & Technology Project of Anhui Province No. 202523o09050002, National Natural Science Foundation of China No. 62472396, Anhui Provincial Natural Science Foundation No. 2508085QF212, and Fundamental Research Funds for the Central Universities No. WK2102026003.

References

1. Bedapudi, P.: Nudenet: Neural nets for nudity classification, detection and selective censoring (2019)
2. Bui, A., Vuong, L., Doan, K., Le, T., Montague, P., Abraham, T., Phung, D.: Erasing undesirable concepts in diffusion models with adversarial preservation. NeurIPS (2024)
3. Bussmann, B., Leask, P., Nanda, N.: Batchtopk sparse autoencoders. arXiv preprint arXiv:2412.06410 (2024)
4. Cai, M., Cun, X., Li, X., Liu, W., Zhang, Z., Zhang, Y., Shan, Y., Yue, X.: Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In: CVPR (2025)
5. Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., et al.: Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705 (2025)
6. Cai, Q., Chen, J., Gao, C., Gong, Z., Li, Y., Pan, Y., Peng, Y., Qiu, Z., Yu, K., Zhang, Y., et al.: Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. arXiv preprint arXiv:2605.11061 (2026)
7. Chin, Z.Y., Jiang, C.M., Huang, C.C., Chen, P.Y., Chiu, W.C.: Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In: ICML (2024)
8. Cunningham, H., Ewart, A., Riggs, L., Huben, R., Sharkey, L.: Sparse autoencoders find highly interpretable features in language models. In: ICLR (2024)
9. Cywiński, B., Deja, K.: Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In: ICML (2025)
10. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: ICCV (2023)
11. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: WACV (2024)
12. Gao, D., Lu, S., Zhou, W., Chu, J., Zhang, J., Jia, M., Zhang, B., Fan, Z., Zhang, W.: Eraseanything: Enabling concept erasure in rectified flow transformers. In: ICML (2025)
13. Geng, J., Li, Q.: Sauce: Selective concept unlearning in vision-language models with sparse autoencoders. In: ICCV (2025)
14. Gong, C., Li, D., Pan, Y., Chen, J., Yao, T., Mei, T.: Freeinpaint: Tuning-free prompt alignment and visual rationality enhancement in image inpainting. In: AAAI (2026)
15. Hsu, C.Y., Tsai, Y.L., Xie, C., Lin, C.H., Chen, J.Y., Li, B., Chen, P.Y., Yu, C.M., Huang, C.Y.: Ring-a-bell! how reliable are concept removal methods for diffusion models? In: ICLR (2024)
16. Huang, A., Cai, Z., Xiong, Z.: A survey of machine unlearning in generative ai models: Methods, applications, security, and challenges. IoT-J (2025)
17. Huang, C.P., Chang, K.P., Tsai, C.T., Lai, Y.H., Yang, F.E., Wang, Y.C.F.: Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In: ECCV (2024)
18. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR (2024)

19. Jiang, H.H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J., Gebru, T.: Ai art and its impact on artists. In: AIES (2023)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
22. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: ICCV (2023)
23. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
24. Lee, B.H., Lim, S., Chun, S.Y.: Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In: CVPR (2025)
25. Li, S., van de Weijer, J., Hu, T., Khan, F.S., Hou, Q., Wang, Y., Yang, J.: Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In: ICLR (2024)
26. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: CVPR (2024)
27. Na, B., Kang, M., Kwak, J., Park, M., Shin, J., Jun, S., Lee, G., Kim, J.H., Moon, I.C.: Training-free safe text embedding guidance for text-to-image diffusion models. NeurIPS (2025)
28. Olshausen, B.A., Field, D.J.: Sparse coding with an overcomplete basis set: A strategy employed by v1? Vision research **37**(23), 3311–3325 (1997)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
30. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: CVPR (2023)
31. Shi, W., Li, S., Liang, T., Wan, M., Ma, G., Wang, X., He, X.: Route sparse autoencoder to interpret large language models. In: EMNLP (2025)
32. Templeton, A.: Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Anthropic (2024)
33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing (2004)
34. Xu, N., Zhang, J., Li, C., Chen, Z., Zhou, C., Li, Q., Du, T., Ji, S.: Videoeraser: Concept erasure in text-to-video diffusion models. In: EMNLP (2025)
35. Yang, T., Cao, J., Xu, C.: Pruning for robust concept erasing in diffusion models. arXiv preprint arXiv:2405.16534 (2024)
36. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR (2025)
37. Ye, X., Cheng, S., Wang, Y., Xiong, Y., Li, Y.: T2vunlearning: A concept erasing method for text-to-video diffusion models. arXiv preprint arXiv:2505.17550 (2025)
38. Yoon, J., Yu, S., Patil, V., Yao, H., Bansal, M.: Safree: Training-free and adaptive guard for safe text-to-image and video generation. In: ICLR (2025)
39. Zhang, Y., Jin, E., Dong, Y., Wu, Y., Torr, P., Khakzar, A., Stegmaier, J., Kawaguchi, K.: Minimalist concept erasure in generative models. In: ICML (2025)
40. Zhang, Y., Chen, X., Jia, J., Zhang, Y., Fan, C., Liu, J., Hong, M., Ding, K., Liu, S.: Defensive unlearning with adversarial training for robust concept erasure in diffusion models. NeurIPS (2024)