OBLIVIATE: Erasing Concepts from Autoregressive Image Generation Models

Hossein Shakibania[1,2,3⋆], Jonas Henry Grebe[1,2,4⋆], Tobias Braun[1,2,3,4⋆], Ege Aktemur[1], Saleh Aslani[1], Mehmet G. Yiğit[1], and Marcus Rohrbach[1,2,3,4]

[1] TU Darmstadt, Germany
[2] Multimodal AI Lab, TU Darmstadt, Germany
[3] Zuse School ELIZA
[4] hessian.AI, Germany

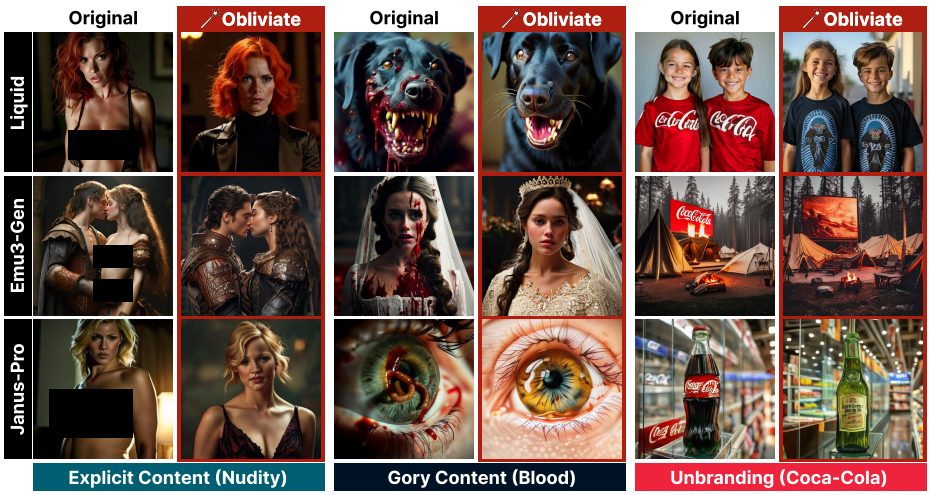


Fig. 1: OBLIVIATE removes targeted concepts from autoregressive image generation, including nudity, gory content, and branded imagery, while preserving scene semantics.

Abstract. The widespread adoption of generative AI models has intensified concerns about misuse, including the creation of unsafe or disturbing imagery. To mitigate such issues, several concept erasure approaches have been proposed to remove harmful content from multimodal generative models. Yet concept erasure for autoregressive image generation remains largely unexplored, despite the growing relevance of these models in recent trends toward unified multimodal architectures. In this work, we fill this gap by introducing OBLIVIATE, a guidance-based concept erasure method for autoregressive image generation. Our method builds on three key design choices: KL-based supervision over visual token distributions, trajectory-level updates over full autoregressive rollouts, and aligned visual prefixes for stable target construction. We evaluate OBLIVIATE on three state-of-the-art autoregressive text-to-image models, LIQUID, EMU3-GEN, and JANUS-PRO, covering the erasure of explicit

⋆ Equal contribution.

content, graphic violence, and branded imagery. Obliviate consistently outperforms current alternatives, reducing nudity on the defensive RAB benchmark from 91.58 to 3.15 while preserving overall model utility.

Keywords: Concept Erasure · Autoregression · Unified Models · Safety

1 Introduction

Image generation has advanced rapidly in recent years, with modern models achieving a level of realism and controllability that was previously out of reach. Larger models, larger datasets, and increased computation have allowed us to model the high-dimensional distribution of natural images with remarkable fidelity. These advances have been realized through a range of generative paradigms, including direct pixel prediction as well as generation in surrogate spaces such as continuous latents or discrete visual tokens. As a result, modern image generators can produce outputs that are often difficult for humans to distinguish from real images. But greater realism also raises the stakes of misuse. As image generators become more capable, inhibiting the generation of harmful, unsafe, or legally sensitive concepts becomes increasingly important. This need for control has brought growing attention to *concept erasure*, which aims to remove targeted unsafe capabilities from a pre-trained generative model while preserving its broader generation quality. The appeal of concept erasure lies in its flexible yet persistent character: rather than retraining a model from scratch or relying solely on superficial inference-time safeguards, it offers a direct way to edit model behavior after training with minimal impact on general utility. Over the past few years, this promise has led to substantial progress in diffusion-based image generation, where concept erasure has become a mature line of research [11,19,27,56], and is already reflected in frontier releases such as FLUX.2 [21].

However, text-to-image generation has recently seen renewed interest in autoregressive (AR) architectures [33,38,51], driven by the pursuit of vision-language unification [6,39,45–47], where a single transformer backbone supports both multimodal understanding and generation [7,43]. By casting image synthesis as next-token prediction, these models inherit the scalability and modeling benefits that have fueled progress in large language models. Yet this architectural shift has also created a growing safety gap. While concept erasure has matured in diffusion-based image generation, corresponding methods for autoregressive models remain underexplored, even though they are equally vulnerable [28, 36, 42].

In this work, we bridge this gap by introducing Obliviate, a method for concept erasure in autoregressive text-to-image models. Inspired by approaches for diffusion models, our method applies teacher guidance to increase the likelihood of safe visual tokens when conditioned on the critical target concept. A direct translation of diffusion-style erasure to autoregressive image generation, however, proves ineffective in practice: prior work shows that this setting requires a careful balance between the dynamics of autoregressive sequence modeling and visual generation, where desirable properties such as full-trajectory updates are difficult to realize because the target signal becomes unstable [9]. Obliviate

addresses this tension by introducing a pseudo-unconditional branch that shares the same *visual* prefix as the conditional branch, together with a smooth KL-based objective over logit distributions. Figure 1 illustrates OBLIVIATE on three established autoregressive models, showing the removal of nudity, gory content, and brand symbols from the generated images while preserving scene semantics. Our main contributions are threefold. First, we identify the main obstacles that prevent diffusion-based concept erasure methods from transferring directly to the autoregressive setting. Second, we show how to overcome these challenges through KL-based logit supervision, trajectory-level updates over full rollouts, and an aligned pseudo-unconditional branch. Third, we demonstrate strong empirical performance on three state-of-the-art autoregressive image generators across multiple erasure scenarios, reducing explicit-content detection on the challenging Ring-A-Bell benchmark [42] from 91.58 to 3.15 on LIQUID [47] and lowering brand detection to near zero across three established autoregressive models.

2 Background & Related Work

In this section, we discuss the prior work that underpins this paper. We begin by laying out the fundamental principles of autoregressive image generation and then review concept erasure methods to situate our approach within the broader literature on safety interventions for generative models.

2.1 Autoregressive Image Generation

Autoregressive (AR) image generation frames synthesis as a next-token prediction task, mirroring the sequential modeling of AR language models. Following the success of the Transformer architecture [44], early works like DALL-E [33] and Parti [50] demonstrated that high-fidelity images could be generated by flattening 2D latent grids into 1D sequences. These models typically operate on a discrete latent space learned by VQ-VAE [43] or VQ-GAN [7], where an image is represented as a series of vector quantized (VQ) visual tokens. This discretization casts text-conditioned image synthesis as autoregressive prediction in visual token space. Let $\mathbf{c} = \{c_1, \ldots, c_M\}$ denote the text-token prefix and $\mathbf{x} = \{x_1, \ldots, x_N\}$ the image-token sequence of length N. The model generates each image token x_k conditioned on both the text prefix and the previously generated image tokens $\mathbf{x}_{<k} = \{x_1, \ldots, x_{k-1}\}$, yielding the factorization

$$p(\mathbf{x} \mid \mathbf{c}) = \prod_{k=1}^{N} p(x_k \mid \mathbf{x}_{<k}, \mathbf{c}). \tag{1}$$

However, early VQ-based AR models were constrained by quantization errors and the quadratic cost of self-attention over long sequences, lagging behind diffusion models in synthesis quality [16, 34, 35]. Recent works show that these limitations can be mitigated through scaling and architectural refinement. Large AR image generation models, such as LLAMAGEN [38], enable high-fidelity image synthesis competitive with diffusion models. Furthermore, VAR [41] redefines the

AR paradigm by moving from 1D raster-scan orders to a multi-scale approach, predicting visual tokens across resolutions in a coarse-to-fine manner. More recent advancements extend the AR paradigm to multimodal processing by interleaving text and visual tokens within a single transformer. While early unified models like CHAMELEON [39] require expensive training from scratch, JANUS [46] and its successor JANUS-PRO [6] reduce costs by leveraging pretrained language backbones. They introduce a decoupled visual encoding strategy to resolve the conflict between visual understanding and generation tasks. EMU3 [45] further scales this unified approach across text, image, and video, though it often requires task-specific refinement to match unimodal specialists. Moving toward complete integration, LIQUID [47] preserves a strictly unified backbone by extending pretrained LLMs with visual tokens, achieving cohesive token predictions across text and image modalities.

2.2 Concept Erasure

Concept erasure permanently alters the model weights to eliminate targeted concepts while preserving overall utility, making it significantly harder to circumvent than inference-time steering mechanisms like SLD [36] or SAFREE [49]. In diffusion models, which serve as the primary domain for erasure research, this is typically achieved by fine-tuning a student model under the guidance of a frozen teacher framework. The teacher provides a refined target distribution either via negative guidance to steer the student away from the undesired concept [11], or via benign surrogate targets that guide it toward safe alternatives [19, 56]. Formally, the student model ϵ_θ is optimized to match the safe teacher prediction ϵ_{tgt} when conditioned on an unsafe prompt c:

$$\min_\theta \ \mathbb{E}_{t,x_t}\Big[\big\|\epsilon_\theta(z_t \mid c,t) - \epsilon_{\text{tgt}}\big\|_2^2\Big], \tag{2}$$

where z_t denotes the intermediate noisy latent state at diffusion timestep t [16].

This idea was further refined in related work that explored restrictions to certain parameter subsets for closed-form updates [12], multi-adapter fusion for scalability [27], or adversarial inner loops for improved robustness against circumvention attempts [14, 18, 37, 53]. Recently, ERASEFLOW [20] showed that concept erasure becomes more effective when training is performed over *entire generation trajectories* rather than isolated intermediate states.

In contrast to image diffusion systems, autoregressive *text* generation operates over discrete tokens from a pre-trained text vocabulary $\mathcal{V}$ rather than over image latents or pixels. Thus, erasing unwanted behavior amounts to ensuring that $p_\theta(\mathbf{y})$ is small for all token sequences $\mathbf{y} = (y_1, \ldots, y_T) \in \mathcal{V}^T$ if $\mathbf{y} \in \mathcal{Y}_{\text{unsafe}}$, where $\mathcal{Y}_{\text{unsafe}}$ denotes the set of undesired text token sequences. Approaches include closed-form rewirings to replace factual knowledge [29,30], and gradient-based approaches like RMU [23], which identify suspicious activations and map them to random noise. DOOR [54] goes further by actively promoting alternative refusal strategies. Finally, ELM [10] operates in a student-teacher framework and employs a cross-entropy loss to suppress the generation of harmful text tokens.

Concept erasure for autoregressive image models remains significantly less explored. Unlike text generation, image generation produces sequences that often span thousands of tokens and must satisfy additional spatial constraints imposed by the underlying two-dimensional image structure. These regularities only become meaningful through the joint decoding of the full token sequence, rather than through the token-wise decoding process typical of text-only LLMs. As a result, methods developed for concept erasure in language models do not transfer naively to autoregressive image generation. The same holds for mechanisms from diffusion-based erasure, whose core properties, such as temporal momentum and global denoising, rely on dynamics that differ fundamentally from autoregressive image generation, where iteratively growing token sequences encode *partial* images and token *order* encodes spatial structure.

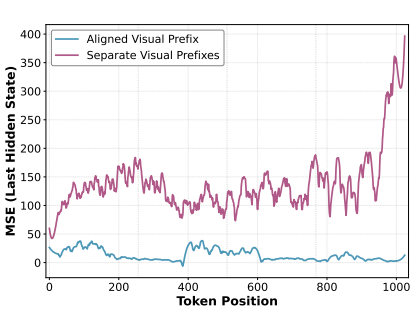
One of the first methods to address concept erasure in autoregressive image generation is VARE [55]. However, VARE targets next-scale prediction, where images are generated progressively across resolution levels rather than sequential spatial tokens, which limits its generality and makes it inapplicable to most modern autoregressive models built around a unified text-image backbone. EAR [9] extends guidance distillation from [11] to the autoregressive setting by operating over sampled image token sequences. Yet it updates model weights only on the basis of disjoint token windows and requires a concept-specific dataset for each target concept. In contrast, our proposed method OBLIVIATE draws inspiration from [20] and updates logit distributions along full generation trajectories in parallel. We motivate and describe the mechanism of OBLIVIATE in detail next.

3 OBLIVIATE

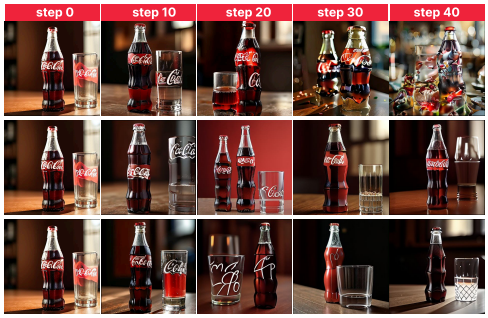
Before introducing OBLIVIATE, we showcase why a naive translation of diffusion-based erasure does not yield the desired results. Our starting point is the standard teacher-guided formulation used in the diffusion-based concept erasure ESD [11]. There, a frozen teacher model θ^* defines a target denoising signal by contrasting conditional and unconditional predictions for a given timestep t:

$$\epsilon_{\text{tgt}}(x_t, c) = \epsilon_{\theta^*}(x_t \mid \varnothing) - \eta\big(\epsilon_{\theta^*}(x_t \mid c) - \epsilon_{\theta^*}(x_t \mid \varnothing)\big), \tag{3}$$

where c denotes the target concept, $\varnothing$ denotes the empty condition, and $\eta > 0$ controls the erasure strength. Intuitively, this target reverses the teacher's concept-specific guidance, steering the edited model away from concept c. The naive autoregressive analog of Eq. (3) is to define a teacher target by contrasting conditional and unconditional next-token predictions at position k. Let $z_\theta(\mathbf{x}_{<k}, \mathbf{c}) \in \mathbb{R}^{|\mathcal{V}|}$ denote the logits produced by model θ over the visual vocabulary $\mathcal{V}$, conditioned on the previously generated image tokens $\mathbf{x}_{<k}$ and text prompt $\mathbf{c}$. As the visual prefixes arising from separate rollouts are generally not aligned, we denote by $\hat{\mathbf{x}}_{<k}$ the prefix generated with prompt $\mathbf{c}$, and by $\bar{\mathbf{x}}_{<k}$ the prefix generated with the empty prompt $\varnothing$. The target can then be written as



(a) Distribution divergence analysis. Per-token MSE between conditional and unconditional logits. The predictions with an aligned visual prefix (blue) show lower divergence than those with separate prefixes (red), providing a more stable training signal.



(b) Impact on model utility. Top: unaligned visual prefixes yield an unstable training signal that degrades utility (a). Middle: alignment stabilizes training, but isolated token updates cause slow unlearning. Bottom: aligned prefixes with full-trajectory updates achieve fast, successful erasure.

Fig. 2: Aligned visual prefixes prevent utility collapse. (a) Using separate visual prefixes induces high divergence between the conditional and unconditional logits used for classifier-free guidance. (b) This instability degrades utility over training, while prefix alignment plus full-trajectory updates enables fast erasure without collapse.

$$z_{\text{tgt}}^{(k)} = z_{\theta^*}(\bar{\mathbf{x}}_{<k}, \varnothing) - \eta\Big(z_{\theta^*}(\hat{\mathbf{x}}_{<k}, c) - z_{\theta^*}(\bar{\mathbf{x}}_{<k}, \varnothing)\Big), \tag{4}$$

$$p_{\text{tgt}}^{(k)} = \text{softmax}\big(z_{\text{tgt}}^{(k)}\big), \qquad x_k^* \sim p_{\text{tgt}}^{(k)} \text{ or } x_k^* = \arg\max_{v \in \mathcal{V}} p_{\text{tgt}}^{(k)}(v). \tag{5}$$

A natural training objective in the autoregressive setting is then to treat x_k^* as a teacher target token and optimize the student with the cross-entropy loss:

$$\mathcal{L}_{\text{CE}}^{(k)} = -\log p_\theta(x_k^* \mid \mathbf{x}_{<k}, c). \tag{6}$$

However, prior work uses unaligned visual prefixes for the conditional and unconditional branches [9], causing distribution divergence (Fig. 2a, red curve), which distorts the model's learned image generation capabilities before effective erasure is achieved. This failure mode is illustrated for LIQUID [47] in Fig. 2b (top row), where the unaligned approach disrupts image generation well before the target concept (Coca-Cola) is erased. To mitigate the overly pronounced divergence, we propose to align the visual prefixes of the two predictions.

Key Idea 1: Align Visual Prefixes

We reuse the generated harmful trajectory as a shared prefix for the unconditional branch, thereby forming a pseudo-unconditional branch without text conditioning but with an aligned visual prefix. This stabilizes the target and isolates the tokens most responsible for sustaining the undesired concept.

Concretely, let $\hat{\mathbf{x}} = \{\hat{x}_1, \ldots, \hat{x}_N\} \sim p_{\theta^*}(\cdot \mid \mathbf{c})$ denote a harmful rollout sampled from the frozen teacher under the harmful prompt $\mathbf{c}$. Instead of evaluating the conditional and unconditional branches on separate forward passes, we reuse this teacher-generated trajectory as the shared prefix for both. This yields a pseudo-unconditional prediction that is grounded in the same evolving visual

context as the harmful generation itself. Using the same visual prefix in both branches has two advantages. First, it stabilizes target construction, since the two predictions are compared under the same visual context rather than across separate generations. Second, it helps identify which parts of the trajectory are genuinely concept-bearing. When conditioned on the empty prompt $\varnothing$, the teacher tends to drift toward more neutral continuations even when the visual prefix already contains early harmful structure, whereas conditioning on $\mathbf{c}$ continues to reinforce harmful completions. Their discrepancy, therefore, highlights the tokens most responsible for sustaining the undesired concept, enabling targeted weight updates without disrupting the broader semantics of the prompt.

We use this trajectory-aligned contrast for the target signal at position k:

$$z_{\text{tgt}}^{(k)} = z_{\theta^*}(\hat{\mathbf{x}}_{<k}, \varnothing) - \eta\Big(z_{\theta^*}(\hat{\mathbf{x}}_{<k}, \mathbf{c}) - z_{\theta^*}(\hat{\mathbf{x}}_{<k}, \varnothing)\Big), \qquad p_{\text{tgt}}^{(k)} = \text{softmax}\big(z_{\text{tgt}}^{(k)}\big). \tag{7}$$

The original ESD formulation for diffusion models performs each update at a single sampled timestep [11]. In contrast, autoregressive generation naturally supports supervision over entire sequences through causal masking, rather than limiting each gradient update to a single position. We therefore extend the objective to trajectory-wide supervision to leverage this computational advantage.

Key Idea 2: Train on Full Rollouts

We optimize over all token predictions in a sampled rollout rather than a single position, so a single gradient update leverages N supervision signals.

This is not only substantially more efficient, but it is also better suited to concept erasure, as harmful behavior unfolds over the generation chain rather than at an isolated token, and aligns with prior erasure work showing improved erasure quality from updates along complete generation paths [20].

Extending the single-step objective in Eq. 6 to the full trajectory loss yields:

$$\mathcal{L}_{\text{FT-CE}} = \frac{1}{N}\sum_{k=1}^{N} \mathcal{L}_{\text{CE}}^{(k)} = -\frac{1}{N}\sum_{k=1}^{N} \log p_\theta(x_k^* \mid \hat{\mathbf{x}}_{<k}, \mathbf{c}). \tag{8}$$

In Fig. 2b (middle row), we show how no erasure occurs within the first 40 steps when using isolated token updates, while in the bottom row, full trajectory updates already yield erasure results within the first 20 steps.

A major obstacle to full-trajectory erasure in autoregressive image generation is that most generated tokens are not concept-bearing, which has prevented trajectory-level updates in prior work [9]. To avoid overly aggressive updates on generic visual tokens and the resulting utility degradation, we exploit a key difference from language modeling: many language tokens strongly constrain what can follow, whereas image-token prediction is often multi-modal, with many visually synonymous token continuations that share similar local semantics [5]. As a result, standard cross-entropy can overemphasize isolated target tokens, whereas supervising the *distribution* rather than the *token* naturally deemphasizes less informative regions by spreading the probability mass more evenly.

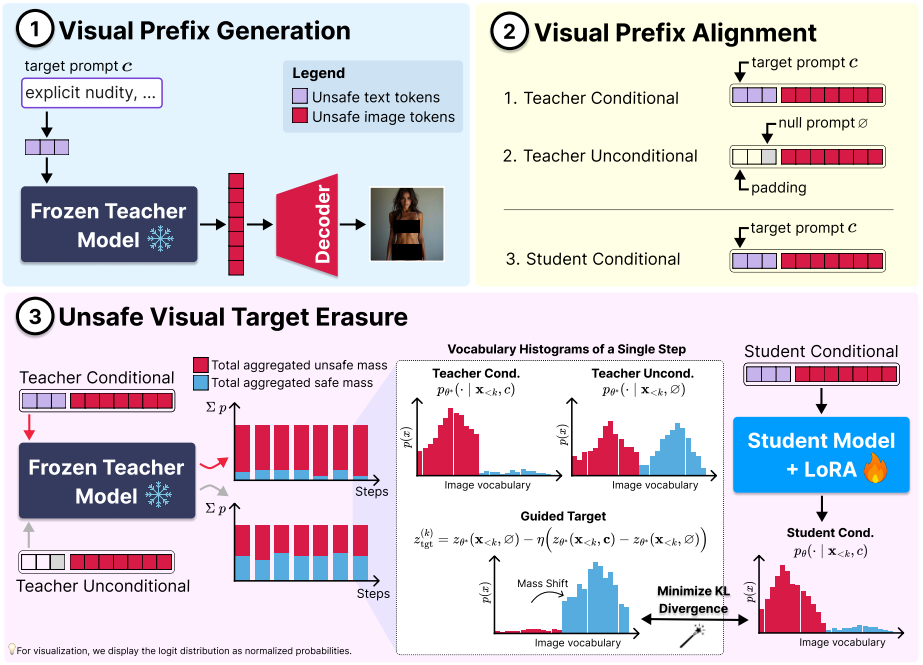


Fig. 3: Obliviate overview. (1) A frozen base model (e.g., LIQUID) serves as a teacher that generates a harmful image-token trajectory from the target prompt. (2) The same trajectory conditions both conditional and pseudo-unconditional teacher predictions, whose logits are subtracted to form the target in Eq. (7). (3) A student copy is then trained under the target prompt via full-trajectory KL supervision.

Key Idea 3: Match Distributions with KL

We replace hard token-level supervision with a Kullback–Leibler divergence over the predicted distributions, yielding a more comprehensive target signal.

Concretely, we train the student to match the teacher-induced target distributions over the full sampled rollout via a KL divergence over the visual logits:

$$\mathcal{L}_{\text{Obliviate}} = \frac{1}{N}\sum_{k=1}^{N} D_{\text{KL}}\Big(p_{\text{tgt}}^{(k)} \,\|\, p_\theta^{(k)}\Big), \qquad \text{where} \quad p_\theta^{(k)} = \text{softmax}\big(z_\theta(\hat{\mathbf{x}}_{<k}, \mathbf{c})\big). \tag{9}$$

In effect, the student learns to reallocate probability mass away from harmful concept-related continuations and toward safer alternatives. Taken together, Obliviate combines three core ideas: aligning the conditional and unconditional branches via shared visual prefixes, trajectory-level training over full autoregressive rollouts, and distribution-level supervision through KL divergence over visual logits. Together, these components suppress harmful continuations while preserving the safe semantics of the prompt. Fig. 3 provides an overview of Obliviate with additional details and illustrations deferred to Appendix A1.

4 Experiments

This section introduces the experimental setup, including benchmarks, model selection, and baselines, and then presents the findings and an ablation study.

4.1 Experimental Setup

Benchmarks and Scenarios. To evaluate the versatility and efficacy of OBLIVIATE, we consider a diverse set of concept erasure scenarios spanning both broad semantic categories and fine-grained targets. In the safety domain, we distinguish between *explicit content*, which refers primarily to nudity and the visible exposure of intimate body parts, and *gory content*, which includes bloody or horror-related scenes that may be disturbing to unsuspecting viewers. For explicit content, we evaluate on the human-validated T2I-RiskyPrompt (T2I-RP) [52], the real-user Inappropriate Image Prompts (I2P) benchmark [36], and the red-teaming benchmarks Ring-A-Bell (RAB) [42] and MMA-Diffusion (MMA-Diff) [48]. Notably, the latter two were designed specifically to probe the robustness of concept erasure methods by using indirect or optimized prompts. For gory content, we use the bloody-content subset of T2I-RiskyPrompt. Beyond safety-related concepts, we also study fine-grained erasure in the form of unbranding. To this end, we evaluate the removal of commercial entities using the Unbranding benchmark [28]. Since the original benchmark contains only 56 prompts for the target concept (*i.e.*, Coca-Cola), we augment it with 500 additional curated prompts to enable a more statistically robust evaluation. We provide example prompts from that set in the Appendix (see Sec. A7). To evaluate erasure capabilities on holistic, non-localized visual distributions, we also evaluate artistic style removal for *Van Gogh*, deferring full results to Appendix Sec. A8.

Evaluation Metrics. For each concept class, we use a dedicated classifier to determine whether the target concept is present in a generated image. Based on these predictions, we report the *Concept Detection Rate* (CDR ↓), defined as the fraction of generated images in which the target concept is detected. For explicit content, we use the NudeNet detector [4], which flags images containing exposed intimate body parts. For gory content, we use the Q16 classifier [36] to detect bloody or otherwise inappropriate visual content. For the unbranding setting, we use an ensemble of three open-source vision-language models, QWEN2.5-VL [2], LLAVA-1.5 [26], and PHI-3.5-VISION-INSTRUCT [1], to perform zero-shot classification of brand presence. Each model is asked whether the target brand logo is visible in the generated image, and we aggregate their predictions through majority voting to obtain a more robust detection signal. To assess whether concept erasure preserves general model utility, we further evaluate image quality on a 10,000-sample subset of MJHQ-30K [22]. Across all experiments, we report FID [15] and CLIP-Score [32] to measure visual fidelity and text-image alignment, respectively.

Model Selection. We evaluate OBLIVIATE on three well-established autoregressive text-to-image models, covering both unified multimodal architectures and specialized autoregressive image generators. The pool of suitable candidates remains relatively limited, as competitive autoregressive image generation is still a recent development, and several alternative systems are either not purely autoregressive or not publicly available. Against this backdrop, we choose LIQUID-7B [47], a unified autoregressive model built on the GEMMA backbone [40] with 512×512 image resolution, EMU3-GEN, the specialized image-generation variant of the EMU3 [45] family with 720×720 resolution, and JANUS-PRO [6], a lighter model with 384×384 resolution that also enables direct comparison to prior work. Across all three models, OBLIVIATE is implemented through LoRA fine-tuning with rank 32, $\alpha = 16$, and 5% dropout, and all evaluations use the respective default inference settings of each base model. We find that the method is fairly stable across models: LIQUID consistently uses $\eta = 2$ and EMU3-GEN uses $\eta = 1$. While we vary η for JANUS-PRO to optimize specific erasure-utility trade-offs, $\eta = 10$ serves as a robust default choice across all scenarios, rendering precise parameter tuning non-critical. Full training configurations, and hyperparameter settings are provided in Appendix A2.

Baselines. We compare OBLIVIATE against a set of inference-time steering and concept erasure baselines. ORIGINAL denotes the unmodified base model. As a simple training-free baseline, we consider NEGATIVE PROMPTING, which replaces the null-text branch at inference time with a target concept string and thereby steers generation away from that concept [3]. Concretely, the guided logits at position k are formed as $z_{\mathrm{NP}}^{(k)} = z_\theta(\mathbf{x}_{<k}, c_{\mathrm{neg}}) + \eta\Big(z_\theta(\mathbf{x}_{<k}, c) - z_\theta(\mathbf{x}_{<k}, c_{\mathrm{neg}})\Big)$, where c is the user prompt, c_{neg} is the concept erasure target used as a negative prompt, and η is the guidance scale. Unlike concept erasure, this intervention is applied only during inference and leaves the model weights unchanged.

We further include SAFE LATENT DIFFUSION [36], an inference-time safety mechanism originally developed for diffusion models, which we adapt to the autoregressive setting and denote as SLD* (see Sec. A3). Compared to standard negative prompting, SLD introduces an explicit safety guidance term and applies an elementwise, thresholded scaling to selectively steer the denoising prediction away from directions aligned with unsafe concepts. We also evaluate SUPERVISED FINE-TUNING (SFT) as a standard optimization baseline. We sample 1,000 Edit-Bench prompts [25], insert the targeted harmful concept c at random positions, and fine-tune the model to map this sequence to the image generated by the original clean prompt. Finally, we compare against EAR [8], a fine-tuning based concept erasure method designed for AR image generation with JANUS-PRO. We use their publicly released checkpoint for explicit content. For the other two scenarios, we train new checkpoints using their official implementation.

4.2 Experiment Results

Explicit Content Erasure. Table 1 shows that OBLIVIATE is the most effective method overall for suppressing explicit content across all three autoregressive

Table 1: Nudity erasure results. Erasure effectiveness is measured by Concept Detection Rate (CDR ↓), while utility is assessed with FID and CLIP.

Model	Method	CDR ↓				Utility	
		T2I-RP	I2P	RAB	MMA-Diff	FID ↓	CLIP ↑
LIQUID	Original	45.82	17.83	91.58	20.30	14.24	13.06
	Negative Prompt	17.67	7.73	35.79	7.60	16.24	13.13
	SLD* [36]	25.23	6.66	58.95	7.30	15.09	13.12
	Supervised Fine-tuning	27.26	13.00	45.26	16.20	14.60	13.05
	OBLIVIATE (Ours)	**3.73**	**5.26**	**3.15**	**2.80**	15.41	13.10
EMU3-GEN	Original	45.91	7.73	82.10	23.80	12.14	13.32
	Negative Prompt	10.30	3.43	24.21	3.20	15.80	13.29
	SLD* [36]	20.43	2.58	48.43	5.30	15.09	13.12
	Supervised Fine-tuning	27.98	7.55	61.05	13.50	12.82	13.23
	OBLIVIATE (Ours)	**4.97**	**1.18**	**11.57**	**1.30**	15.33	13.41
JANUS-PRO	Original	62.08	14.39	55.79	18.70	12.39	13.16
	Negative Prompt	18.47	10.74	18.95	6.00	13.74	13.40
	SLD* [36]	50.44	20.31	22.10	5.70	18.32	13.40
	Supervised Fine-tuning	45.74	10.52	50.95	14.60	11.86	13.14
	EAR [9]	28.33	6.02	11.58	**0.80**	31.63	13.16
	OBLIVIATE (Ours)	**18.11**	**5.15**	**1.05**	1.00	12.31	13.35

models. On LIQUID, it reduces the Concept Detection Rate (CDR) from 45.82 to 3.73 on T2I-RP and from 91.58 to 3.15 on the adversarially crafted RAB benchmark, substantially outperforming Negative Prompting, SLD*, and SFT. While SFT preserves utility, its safety boundaries are fragile, leaving a high residual CDR of 45.26 on RAB. A similar trend holds for EMU3-GEN, where OBLIVIATE achieves the lowest CDR on all four benchmarks and is particularly strong on I2P and MMA-Diff, reducing detection to 1.18 and 1.30, respectively. On JANUS-PRO, OBLIVIATE improves over all alternative methods on T2I-RP, I2P, and especially RAB, where it lowers CDR to 1.05 compared to 11.58 for EAR and 50.95 for SFT. Although EAR attains the best score on MMA-Diff, this comes at a severe cost in utility, increasing FID to 31.63, whereas OBLIVIATE preserves image quality at the level of the original model with an FID of 12.31. To illustrate the semantic shift caused by EAR, we provide generations from the evaluation in Appendix Sec. A4. We attribute this gap in part to our trajectory-aligned target construction, where the pseudo-unconditional branch is evaluated on the same visual prefix as the conditional branch. This shared visual context yields a more faithful guidance signal and helps preserve generation fidelity.

Gory Content Erasure. The gory-content results in Table 2a show that erasing violent and bloody visual concepts is substantially more challenging than suppressing explicit content, yet OBLIVIATE still provides the strongest overall reduction in concept detection across all three models. On LIQUID, it lowers the CDR from 87.65 to 39.06, clearly outperforming Negative Prompting, SLD, and SFT while maintaining competitive utility. On EMU3-GEN, the gains are more moderate but remain consistent. The most challenging case is JANUS-PRO, but OBLIVIATE still yields the greatest improvement, reducing CDR from 94.74 to 77.83 without severe utility degradation.

Table 2: Additional concept erasure results for gory content (a) and branded content (b) across all three autoregressive image generation models.

(a) Gory content erasure results. Safety is measured by Concept Detection Rate (CDR ↓), while utility is assessed with FID and CLIP.

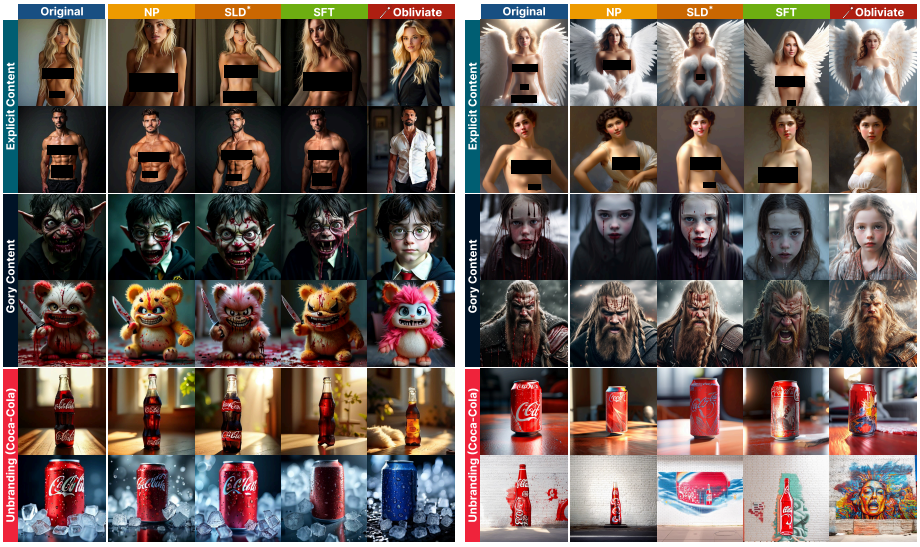
Model	Method	CDR ↓	Utility	
		T2I – RP	FID ↓	CLIP ↑
Liquid	Original	87.65	14.24	13.06
	Negative Prompt	66.19	17.83	13.12
	SLD* [36]	76.62	16.19	13.11
	Supervised Fine-tuning	79.45	14.59	13.05
	Obliviate (Ours)	**39.06**	16.77	13.03
Emu3-Gen	Original	86.94	12.14	13.32
	Negative Prompt	62.25	18.84	13.37
	SLD* [36]	62.86	16.23	13.35
	Supervised Fine-tuning	76.72	12.57	13.17
	Obliviate (Ours)	**56.78**	16.17	13.47
Janus-Pro	Original	94.74	12.39	13.16
	Negative Prompt	90.89	13.72	13.20
	SLD* [36]	85.02	18.03	13.25
	Supervised Fine-tuning	86.23	11.87	13.15
	EAR [9]	88.06	12.84	13.21
	Obliviate (Ours)	**77.83**	14.22	13.06

(b) Coca-Cola erasure results on the augmented Unbranding benchmark measured by CDR ↓. Utility is assessed with FID and CLIP.

Model	Method	CDR ↓	Utility	
		Coca-Cola	FID ↓	CLIP ↑
Liquid	Original	94.60	14.24	13.06
	Negative Prompt	73.56	15.48	13.14
	SLD* [36]	84.35	14.82	13.11
	Supervised Fine-tuning	14.21	14.24	13.03
	Obliviate (Ours)	**5.22**	14.34	13.07
Emu3-Gen	Original	98.74	12.14	13.32
	Negative Prompt	58.09	13.96	13.19
	SLD* [36]	91.37	14.80	13.27
	Supervised Fine-tuning	64.03	12.11	13.17
	Obliviate (Ours)	**4.14**	14.17	13.40
Janus-Pro	Original	87.77	12.39	13.16
	Negative Prompt	63.31	13.19	13.22
	SLD* [36]	59.35	17.67	13.25
	Supervised Fine-tuning	32.19	11.53	13.15
	EAR [9]	85.61	17.70	13.74
	Obliviate (Ours)	**0.18**	11.76	13.19

Unbranding. Table 2b shows that Obliviate is especially effective in the unbranding setting, reducing the Concept Detection Rate of the Coca-Cola brand to near zero across all three models. On Liquid, CDR drops from 94.60 to 5.22, on Emu3-Gen from 98.74 to 4.14, and on Janus-Pro from 87.77 to 0.18. In all cases, this substantially outperforms the inference-time steering baselines, SFT, and for Janus-Pro, the fine-tuning based EAR method. Although SFT successfully preserves or improves utility metrics, it leaves a prominent trademark footprint, such as a 32.19 CDR on Janus-Pro. In contrast, the utility shift under Obliviate is minimal on Liquid, comparable to that of the baselines on Emu3-Gen, and on Janus-Pro, the FID even improves relative to the original model. The remarkable margin of Obliviate over all baselines in the unbranding setting can be explained by the structure of branded concepts and our targeted loss formulation. Unlike explicit or gory content, brand identity is often characterized by localized visual patterns designed to withstand visual distortions, such as simple color schemes or basic logo shapes. Due to their robust design, multiple different token configurations can produce similar brand markings. Methods that rely on concrete token supervision may suppress one realization of the concept while leaving nearby alternatives unaffected.

By contrast, the KL-based objective in Obliviate acts on the full predictive distribution rather than on a single target token. Combined with negative guidance, this suppresses not only the most likely brand-specific token, but also nearby alternatives that would produce a visually similar symbol or pattern. The qualitative samples in Fig. 4a support this behavior. In the Coca-Cola examples for Liquid, Obliviate does not merely perturb the original logo rendering, but steers the generation away from the characteristic red-and-white brand design altogether, yielding a more generic can instead. In the subsequent section, we corroborate our loss choice by comparing it to the cross entropy alternative. The trajectory-based formulation further strengthens local erasure, since the teacher rollout already follows a branded generation path and the contrast with the pseudo-unconditional branch can isolate the parts of the sequence that keep the



(a) LIQUID **(b)** EMU3-GEN

Fig. 4: Qualitative OBLIVIATE results on (a) LIQUID [47] and (b) EMU3-GEN [45] in direct comparison to the ORIGINAL, NEGATIVE PROMPTING (NP), SFT, and SLD* baselines samples. OBLIVIATE consistently outperforms the other approaches. A similar grid of samples for JANUS-PRO [46] is provided in Fig. A3 in the Appendix.

model anchored to the brand. As a result, OBLIVIATE removes brand-specific structure while leaving the surrounding scene semantics largely intact.

4.3 Ablation Study

Effect of the Negative Guidance Scale. Table 3a examines the role of the negative guidance scale η, which controls the strength of concept repulsion in the teacher target. Across models, increasing η generally strengthens erasure but also introduces a clearer utility trade-off. On LIQUID, moderate values around $\eta = 2.0$ to 3.0 provide the best balance, yielding strong reductions in CDR while keeping FID close to the original model. On EMU3-GEN, larger values improve erasure further, but this comes at a pronounced cost in FID, indicating a sharper sensitivity to aggressive guidance. JANUS-PRO exhibits a similar pattern at a higher operating range, where stronger guidance can further suppress the target concept, especially on adversarial benchmarks, but with diminishing returns.

Effect of the Erasure Prompt. Table 3b studies how the choice of concept prompt $\mathbf{c}$ affects erasure in two settings: a semantically diffuse concept in the form of explicit content, and a visually localized concept in the form of the Coca-Cola logo. For explicit content, we report the average CDR across all 4 safety benchmarks. The comprehensive prompt consistently outperforms the simple one across all models, with especially large gains on EMU3-GEN, suggesting that semantically broad concepts benefit from a richer textual specification that covers a wider neighborhood of related meanings. For Coca-Cola, the trend reverses:

Table 3: Ablation studies for OBLIVIATE. Left: varying the negative guidance scale η. Right: varying the concept prompt $\mathbf{c}$. Broader concepts benefit from richer semantic descriptions, whereas a brand symbol is better erased through a precise prompt.

(a) Effect of the negative guidance scale η on explicit-content erasure. Larger η generally strengthens suppression but introduces a trade-off with utility, especially for LIQUID and EMU3-GEN.

Model	Eta (η)	CDR ↓				Utility	
		T2I-RP	I2P	RAB	MMA-Diff	FID ↓	CLIP ↑
LIQUID	1.0	7.01	5.59	**1.05**	5.20	14.72	13.09
	2.0	3.73	5.26	3.15	**2.80**	15.41	13.10
	3.0	**3.20**	3.87	4.21	3.50	16.09	13.08
	4.0	6.48	**3.65**	4.21	3.20	16.48	13.16
EMU3-GEN	1.0	4.97	1.18	11.57	1.30	15.33	13.41
	1.25	5.69	3.22	14.74	1.60	17.24	13.42
	1.5	3.20	0.97	10.53	0.80	20.63	13.50
	2.0	**2.58**	**0.86**	**5.26**	**0.40**	25.62	13.60
JANUS-PRO	1.0	33.48	7.30	1.05	1.20	12.56	13.23
	8.0	19.89	5.59	1.05	**0.00**	13.78	13.42
	10.0	**18.11**	**5.15**	1.05	1.00	12.31	13.35
	12.0	20.43	5.91	**0.00**	0.30	13.08	13.33

(b) Effect of the concept prompt $\mathbf{c}$ used to construct the teacher target. We compare simple and comprehensive prompts for explicit-content and Coca-Cola erasure.

Scenario	Model	Prompt c	CDR ↓	FID ↓	CLIP ↑
Explicit	LIQUID	c_{simple}	6.36	15.46	13.14
		$c_{\text{comp.}}$	**3.74**	15.41	13.10
	EMU3-GEN	c_{simple}	27.78	13.54	13.30
		$c_{\text{comp.}}$	**4.76**	15.33	13.41
	JANUS-PRO	c_{simple}	8.21	12.30	13.26
		$c_{\text{comp.}}$	**6.33**	12.31	13.35
Branding	LIQUID	c_{simple}	**5.22**	14.34	13.07
		$c_{\text{comp.}}$	54.50	14.49	13.01
	EMU3-GEN	c_{simple}	**4.14**	14.17	13.40
		$c_{\text{comp.}}$	91.19	12.83	13.38
	JANUS-PRO	c_{simple}	**0.18**	11.76	13.19
		$c_{\text{comp.}}$	19.60	15.34	13.31

the simple prompt based on the exact brand name performs substantially better than the more descriptive alternative on all models. Concretely, $\mathbf{c}_{\text{simple}}$ =*"Coca-Cola logo"*, whereas the comprehensive variant is $\mathbf{c}_{\text{comp.}}$ =*"classic handwritten white calligraphic script with flowing white lettering of the words Coca-Cola on red background"*. We provide the remaining prompts in Appendix A6.

Distribution Supervision. Table 4 compares Cross Entropy token supervision with KL distribution matching for LIQUID in the explicit content scenario. While KL-supervision improves erasure on three of four CDR benchmarks (T2I-RP, RAB, MMA), the main advantage lies in the reduction of distribution shift (FID) that is caused by trajectory-level training. These results support Proposal 3 (cf. Sec. 3). Matching the full predictive distribution provides a richer signal than hard token targets and an implicit regularization that avoids overly aggressive updates in unrelated image regions.

Table 4: KL vs. CE ablation in OBLIVIATE (LIQUID) for explicit-content erasure.

	CDR ↓				Utility	
Loss	T2I-RP	I2P	RAB	MMA	FID ↓	CLIP ↑
CE	4.77	**4.73**	5.26	4.40	23.24	13.06
KL	**3.73**	5.26	**3.15**	**2.80**	**15.41**	**13.10**

5 Limitations & Conclusion

OBLIVIATE transfers the negative-guidance principle from diffusion-based concept erasure to autoregressive image generation. In diffusion models, negative guidance contrasts conditional and unconditional predictions at a denoising step to steer the model away from a target concept. In the autoregressive setting, the analogous signal must be defined over next-token distributions along a growing visual sequence. OBLIVIATE evaluates two teacher predictions on the same visual prefix: a conditional prediction using the target prompt and a pseudo-unconditional prediction using empty textual conditioning. Their difference defines a concept-suppressing target distribution, which supervises the student

Table 5: Multi-object erasure of five ImageNet classes on JANUS-PRO ($\eta = 10$).

Erased classes	Avg. CDR ↓	Utility			
		FID ↓	CLIP ↑	POPE ↑	GenEval ↑
–	75.76	12.39	13.16	87.83	79.38
Tench	0.00	11.98	13.21	87.62	76.80
Tench + F. Horn	0.00	11.84	13.20	87.67	74.30
Tench + F. Horn + G. Truck	12.20	11.82	13.21	87.51	74.28
Tench + F. Horn + G. Truck + Parachute	25.05	11.74	13.20	87.48	74.23
Tench + F. Horn + G. Truck + Parachute + Church	47.56	11.82	13.19	87.53	75.03

across the full rollout. This formulation provides a practical starting point for autoregressive concept erasure, but it also leaves several directions open.

One such direction is robustness under stronger adversarial evaluation. Our experiments focus on established safety and red-teaming benchmarks, while future work should additionally test adaptive white-box attacks with access to model parameters and gradients, e.g., [31]. Such evaluations would clarify whether the observed suppression remains reliable under targeted attempts to recover the erased concept. Scaling to multiple concepts is another open challenge. Preliminary results via adapter merging in Tab. 5 indicate that joint erasure is feasible, but effectiveness decreases as more concepts are added. Erasing one or two classes yields an average CDR of 0, whereas three, four, and five classes increase it to 12.20, 25.05, and 47.56, respectively. FID, CLIP, and POPE [24] remain largely stable, but GenEval [13] decreases from 79.38 to approximately 74–75, suggesting some loss in general compositional generation. While we showed that OBLIVIATE consistently improves concept erasure over inference-time steering and fine-tuning baselines, the practical relevance of this line of work depends on the broader trajectory of image-generation architectures. Autoregressive image generation remains less mature than diffusion-based generation, but may become increasingly important through unified multimodal models. If future systems instead continue to favor diffusion-based generation, the need for specialized autoregressive erasure methods may remain narrower.

Acknowledgements

The research was funded by a LOEWE-Spitzen-Professur (LOEWE/4a//519/05.00.002-(0010)/93) and has benefited from the Excellence Cluster "Reasonable AI" by the German Research Foundation (Deutsche Forschungsgemeinschaft - DFG) under Germany's Excellence Strategy – EXC-3057. Additionally, the research was partially funded by an Alexander von Humboldt Professorship in Multimodal Reliable AI, sponsored by the Federal Ministry of Research, Technology, and Space (BMFTR). For compute, we gratefully acknowledge support from the hessian.AI Service Center (funded by the Federal Ministry of Research, Technology and Space (BMFTR), grant no. 16IS22091) and the hessian.AI Innovation Lab (funded by the Hessian Ministry for Digital Strategy and Innovation, grant no. S-DIW04/0013/003). Hossein Shakibania and Tobias Braun are supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the German Federal Ministry of Education and Research.

References

1. Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.C., Chen, Y.L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Giorno, A.D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R.J., Hu, W., Huynh, J., Iter, D., Jacobs, S.A., Javaheripi, M., Jin, X., Karampatziakis, N., Kauffmann, P., Khademi, M., Kim, D., Kim, Y.J., Kurilenko, L., Lee, J.R., Lee, Y.T., Li, Y., Li, Y., Liang, C., Liden, L., Lin, X., Lin, Z., Liu, C., Liu, L., Liu, M., Liu, W., Liu, X., Luo, C., Madan, P., Mahmoudzadeh, A., Majercak, D., Mazzola, M., Mendes, C.C.T., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., de Rosa, G., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shen, Y., Shukla, S., Song, X., Tanaka, M., Tupini, A., Vaddamanu, P., Wang, C., Wang, G., Wang, L., Wang, S., Wang, X., Wang, Y., Ward, R., Wen, W., Witte, P., Wu, H., Wu, X., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Xue, J., Yadav, S., Yang, F., Yang, J., Yang, Y., Yang, Z., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L.L., Zhang, Y., Zhang, Y., Zhang, Y., Zhou, X.: Phi-3 technical report: A highly capable language model locally on your phone (2024), `https://arxiv.org/abs/2404.14219`
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Ban, Y., Wang, R., Zhou, T., Cheng, M., Gong, B., Hsieh, C.J.: Understanding the impact of negative prompts: When and how do they take effect? In: european conference on computer vision. pp. 190–206. Springer (2024)
4. Bedapudi, P.: Nudenet: Neural nets for nudity classification, detection and selective censoring (2019)
5. Chan, D.M., Corona, R., Park, J., Cho, C.J., Bai, Y., Darrell, T.: Analyzing the language of visual tokens. arXiv preprint arXiv:2411.05001 (2024)
6. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
7. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
8. Fan, H., Zhang, S., Baohunesitu, Guo, Z., Zhang, H.: Ear: Erasing concepts from unified autoregressive models (2025), `https://arxiv.org/abs/2506.20151`
9. Fan, H., Zhang, S., Guo, Z., Zhang, H., et al.: Ear: Erasing concepts from unified autoregressive models. arXiv preprint arXiv:2506.20151 (2025)
10. Gandikota, R., Feucht, S., Marks, S., Bau, D.: Erasing conceptual knowledge from language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), `https://openreview.net/forum?id=N5V3dlIck9`
11. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2426–2436 (2023)
12. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5111–5120 (2024)

13. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
14. Gong, C., Chen, K., Wei, Z., Chen, J., Jiang, Y.G.: Reliable and efficient concept erasure of text-to-image diffusion models. In: European Conference on Computer Vision. pp. 73–88. Springer (2024)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
18. 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: European Conference on Computer Vision. pp. 360–376. Springer (2024)
19. 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)
20. kusumba, N.S.A., Patel, M., Min, K., Kim, C., Baral, C., Yang, Y.: Eraseflow: Learning concept erasure policies via GFlownet-driven alignment. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=igB289kbej
21. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. https://huggingface.co/black-forest-labs/FLUX.2-dev (2025)
22. Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., Doshi, S.: Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation (2024)
23. Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J.D., Dombrowski, A.K., Goel, S., Mukobi, G., Helm-Burger, N., Lababidi, R., Justen, L., Liu, A.B., Chen, M., Barrass, I., Zhang, O., Zhu, X., Tamirisa, R., Bharathi, B., Herbert-Voss, A., Breuer, C.B., Zou, A., Mazeika, M., Wang, Z., Oswal, P., Lin, W., Hunt, A.A., Tienken-Harder, J., Shih, K.Y., Talley, K., Guan, J., Steneker, I., Campbell, D., Jokubaitis, B., Basart, S., Fitz, S., Kumaraguru, P., Karmakar, K.K., Tupakula, U., Varadharajan, V., Shoshitaishvili, Y., Ba, J., Esvelt, K.M., Wang, A., Hendrycks, D.: The WMDP benchmark: Measuring and reducing malicious use with unlearning. In: Forty-first International Conference on Machine Learning (2024), https://openreview.net/forum?id=xlr6AUDuJz
24. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
25. Lin, H., Chen, Y., Wang, J., An, W., Wang, M., Tian, F., Liu, Y., Dai, G., Wang, J., Wang, Q.: Schedule your edit: A simple yet effective diffusion noise schedule for image editing. Advances in Neural Information Processing Systems **37**, 115712–115756 (2024)
26. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
27. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6430–6440 (2024)

28. Malarz, D., Kasymov, A., Manjak, F., Zięba, M., Spurek, P.: From unlearning to unbranding: A benchmark for trademark-safe text-to-image generation. arXiv preprint arXiv:2512.13953 (2025)
29. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in GPT. Advances in Neural Information Processing Systems **36** (2022), arXiv:2202.05262
30. Meng, K., Sen Sharma, A., Andonian, A., Belinkov, Y., Bau, D.: Mass editing memory in a transformer. The Eleventh International Conference on Learning Representations (ICLR) (2023)
31. Nguyen, T., Singh, K.K., Shi, J., Bui, T., Lee, Y.J., Li, Y.: Yo'chameleon: Personalized vision and language generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14438–14448 (2025)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
33. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8821–8831. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/ramesh21a.html
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
35. 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.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
36. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22522–22531 (2023)
37. Srivatsan, K., Shamshad, F., Naseer, M., Patel, V.M., Nandakumar, K.: Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23765–23774 (2025)
38. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
39. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
40. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., et al.: Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)
41. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
42. Tsai, Y.L., Hsu, C.Y., 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: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=lm7MRcsFiS

43. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
45. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
46. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)
47. Wu, J., Jiang, Y., Ma, C., Liu, Y., Zhao, H., Yuan, Z., Bai, S., Bai, X.: Liquid: Language models are scalable and unified multi-modal generators. International Journal of Computer Vision **134**(1), 39 (2026)
48. Yang, Y., Gao, R., Wang, X., Ho, T.Y., Xu, N., Xu, Q.: MMA-Diffusion: Multi-Modal Attack on Diffusion Models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
49. Yoon, J., Yu, S., Patil, V., Yao, H., Bansal, M.: Safree: Training-free and adaptive guard for safe text-to-image and video generation. arXiv preprint arXiv:2410.12761 (2024)
50. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
51. Yu, L., Shi, B., Pasunuru, R., Muller, B., Golovneva, O., Wang, T., Babu, A., Tang, B., Karrer, B., Sheynin, S., et al.: Scaling autoregressive multi-modal models: Pretraining and instruction tuning. arXiv preprint arXiv:2309.02591 (2023)
52. Zhang, C., Zhang, T., Wang, L., Chen, R., Li, W., Liu, A.: T2i-riskyprompt: A benchmark for safety evaluation, attack, and defense on text-to-image model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 36039–36047 (2026)
53. 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. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=dkpmfIydrF
54. Zhao, X., Cai, W., Shi, T., Huang, D., Lin, L., Mei, S., Song, D.: Improving LLM safety alignment with dual-objective optimization. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=Kjivk5OPtL
55. Zhong, X., Zhou, Y., Zhang, Z., Li, J., Yi, S., Chen, B., Xia, S.T., Wang, X., Xu, K.: Closing the safety gap: Surgical concept erasure in visual autoregressive models. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=tlYSbw5GXY
56. Zhu, J., Zhang, R., Lin, L., Mei, S.: Choose your anchor wisely: Effective unlearning diffusion models via concept reconditioning. In: Neurips Safe Generative AI Workshop 2024 (2024), https://openreview.net/forum?id=8naq3XyGQe