Multimodal Unlearning Across Vision, Language, Video, and Audio: Survey of Methods, Datasets, and Benchmarks

Nobin Sarwar, Shubhashis Roy Dipta, Zheyuan Liu, Vaidehi Patil
University of Maryland, Baltimore County
University of Notre Dame UNC Chapel Hill
{sms2, sroydip1}@umbc.edu
zliu29@nd.edu, vaidehi@cs.unc.edu

Abstract

With the growing adoption of VLMs, DMs, LLMs, and AFMs, these multimodal foundation models can inadvertently encode sensitive, copyrighted, biased, or unsafe cross-modal associations that originate from their training data. Retraining after deletion requests or policy updates is often impractical, and targeted forgetting remains difficult because knowledge is distributed across shared representations. Multimodal unlearning addresses this challenge by enabling selective removal across modalities while retaining overall utility. This survey offers a unified, system-oriented view of multimodal unlearning across vision, language, audio, and video, grounded in recent advances, emerging applications, and open problems. Our taxonomy enables systematic comparison across model architectures and modalities, clarifying trade-offs among deletion strength, retention, efficiency, reversibility, and robustness. This survey highlights open problems and practical considerations to support future research and deployment of multimodal unlearning.

1 Introduction

Multimodal foundation models, including Vision Language Models (VLMs), Diffusion Models (DMs), Large Language Models (LLMs) and Audio Foundation Models (AFMs)-based (Ho et al., 2020; Team et al., 2023; Yang et al., 2025; Chu et al., 2023; Huang et al., 2024c) generators, support image, video, and audio understanding and generation at scale. Training on web-scale multimodal data improves generalization, but it can also induce memorization and undesired associations involving sensitive, copyrighted, biased, or unsafe content across modalities. As a result, deployed models may need to forget specific items or concepts, such as a copyrighted artwork, a private face, or a harmful trope, while retaining performance on the remaining data (Fan et al., 2023; Gandikota et al., 2023; Zhang et al., 2024d; Sun et al., 2024; Chen et al., 2025d,b; Facchiano et al., 2025). When deletion requests or policy updates affect only part of the training signal, retraining from scratch is often impractical (Voigt and Von dem Bussche, 2017; Goldman, 2020). Targeted removal is challenging because knowledge is distributed in shared representations, so eliminating one association can disrupt unrelated behavior.

Survey	**Venue & Year**	**System-first**	**Text**	**Image**	**Video**	**Audio**
Si et al., 2023	arXiv'23		✔			
Blanco-Justicia et al., 2025	AIR'24	✔	✔			
Liu et al., 2024f	arXiv'24	✔	✔	✔		
Liu et al., 2025b	NMI'25		✔			
Feng et al., 2025b	arXiv'25	✔	✔	✔		✔
Geng et al., 2025	arXiv'25	✔	✔	✔		
Ours	-	✔	✔	✔	✔	✔

Table 1: Comparison of multimodal unlearning surveys across **modalities** and **system-first taxonomy** coverage.

These challenges have driven growing interest in multimodal unlearning as a mechanism for selective data removal and behavior correction. Early work on machine unlearning formalized the goal of efficiently removing training influence from learned models (Cao and Yang, 2015; Bourtoule et al., 2021). Subsequent studies extend this objective to multimodal and generative systems, including DMs and VLMs, by enabling instance-level or concept-level deletion while preserving overall utility (Kim et al., 2023; Liu and Tan, 2024; Li et al., 2024b; Sun et al., 2024; Zhang et al., 2024a; Golatkar et al., 2024). These efforts make multimodal unlearning a central tool for model governance, supporting targeted forgetting without sacrificing overall utility.

While several surveys discuss multimodal unlearning (Table 1), prior work often emphasizes unimodal settings such as text-only or image-only, or it restricts coverage to a narrow set of text-image systems. Many reviews also adopt algorithm-centric taxonomies organized around optimization objectives, which can obscure the intervention points that matter for deploying unlearning in end-to-end

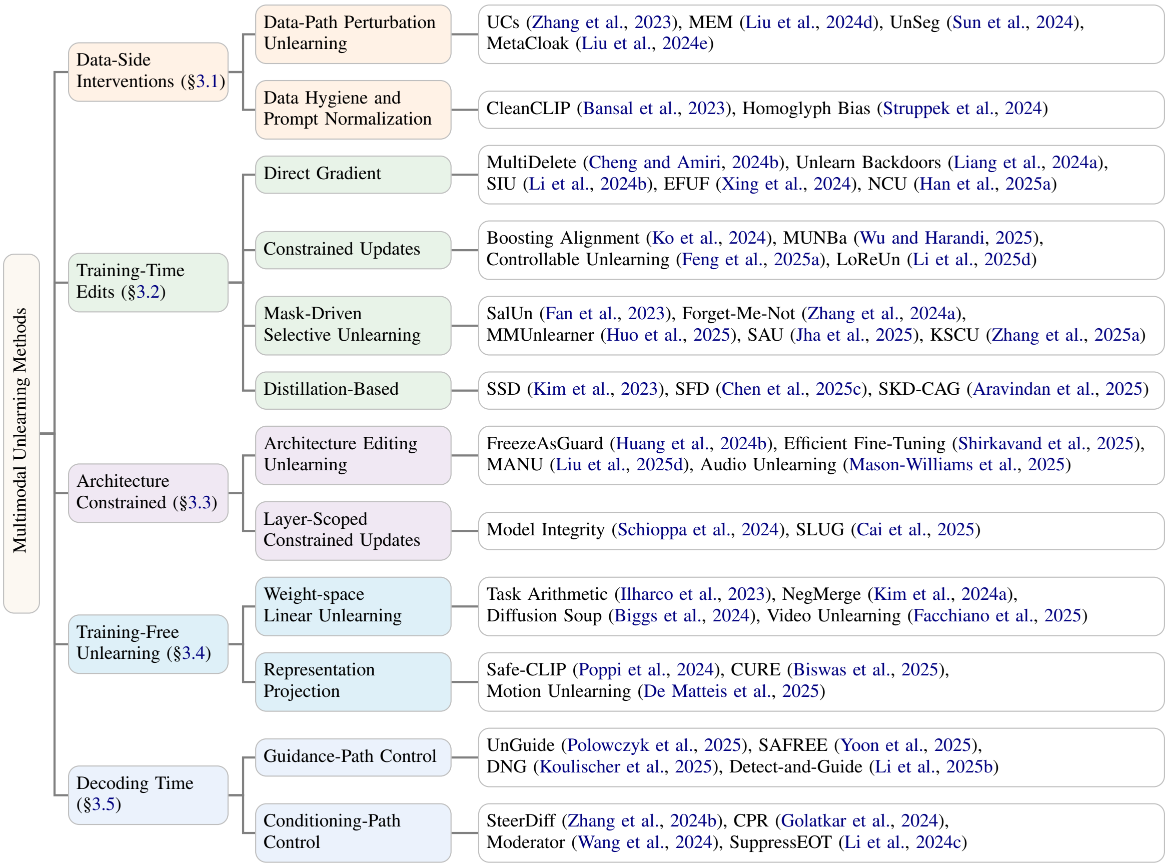
Multimodal Unlearning Methods
Data-Side Interventions (§3.1)
Data-Path Perturbation Unlearning
UCs (Zhang et al., 2023), MEM (Liu et al., 2024d), UnSeg (Sun et al., 2024), MetaCloak (Liu et al., 2024e)
Data Hygiene and Prompt Normalization
CleanCLIP (Bansal et al., 2023), Homoglyph Bias (Struppek et al., 2024)
Training-Time Edits (§3.2)
Direct Gradient
MultiDelete (Cheng and Amiri, 2024b), Unlearn Backdoors (Liang et al., 2024a), SIU (Li et al., 2024b), EFUF (Xing et al., 2024), NCU (Han et al., 2025a)
Constrained Updates
Boosting Alignment (Ko et al., 2024), MUNBa (Wu and Harandi, 2025), Controllable Unlearning (Feng et al., 2025a), LoReUn (Li et al., 2025d)
Mask-Driven Selective Unlearning
SalUn (Fan et al., 2023), Forget-Me-Not (Zhang et al., 2024a), MMUnlearner (Huo et al., 2025), SAU (Jha et al., 2025), KSCU (Zhang et al., 2025a)
Distillation-Based
SSD (Kim et al., 2023), SFD (Chen et al., 2025c), SKD-CAG (Aravindan et al., 2025)
Architecture Constrained (§3.3)
Architecture Editing Unlearning
FreezeAsGuard (Huang et al., 2024b), Efficient Fine-Tuning (Shirkavand et al., 2025), MANU (Liu et al., 2025d), Audio Unlearning (Mason-Williams et al., 2025)
Layer-Scoped Constrained Updates
Model Integrity (Schioppa et al., 2024), SLUG (Cai et al., 2025)
Training-Free Unlearning (§3.4)
Weight-space Linear Unlearning
Task Arithmetic (Ilharco et al., 2023), NegMerge (Kim et al., 2024a), Diffusion Soup (Biggs et al., 2024), Video Unlearning (Facchiano et al., 2025)
Representation Projection
Safe-CLIP (Poppi et al., 2024), CURE (Biswas et al., 2025), Motion Unlearning (De Matteis et al., 2025)
Decoding Time (§3.5)
Guidance-Path Control
UnGuide (Polowczyk et al., 2025), SAFREE (Yoon et al., 2025), DNG (Koulischer et al., 2025), Detect-and-Guide (Li et al., 2025b)
Conditioning-Path Control
SteerDiff (Zhang et al., 2024b), CPR (Golatkar et al., 2024), Moderator (Wang et al., 2024), SuppressEOT (Li et al., 2024c)

Figure 1: Taxonomy of multimodal unlearning methods, organized by intervention stage and control pathway, with representative approaches in each category.

multimodal pipelines. As a result, the literature still lacks a unified exposition that connects mechanisms across vision, language, video, and audio.

Motivated by these gaps, this survey provides a comprehensive overview of multimodal unlearning for foundation models across vision, language, video, and audio. Instead of an algorithm-first taxonomy, we adopt a system-first view that organizes methods by intervention stage and control point, with **forgetting target scope** as the top-level split between **instance-level** and **concept-level** forgetting. This organization provides a stable scaffold for both established and emerging methods, enables cross-modal comparisons through shared control pathways, and clarifies trade-offs among deletion strength, utility retention, efficiency, and reversibility. This survey makes the following contributions to multimodal unlearning in foundation models:

- **Foundational Survey.** This survey synthesizes multimodal unlearning across foundation models for image, text, video, and audio, covering mechanisms, theory, and evaluation in one framework.
- **System-Level Lens.** We propose a system-first taxonomy organized by intervention stage and control pathway, enabling comparison across model classes and optimization families.
- **Emerging Frontiers.** We outline open challenges in evaluation, adversarial robustness, and deployment constraints, highlighting directions for accountable targeted unlearning.

2 Formalizing Multimodal Unlearning

The goal of multimodal unlearning is to remove the influence of a designated forget set while preserving utility on retained content across individual modalities and their shared representations (Cao and Yang, 2015; Ginart et al., 2019; Guo et al., 2020; Bourtoule et al., 2021; Li et al., 2024b). Given a learning algorithm A and multimodal training data $D = \{(I_i, T_i)\}_{i=1}^{N}$ consisting of paired images I and texts T, let $M_o = A(D)$ denote the original model. For simplicity, we use image-text pairs; the formulation generalizes to video and audio. For a forget set $D_f \subseteq D$, define the retained data $D_r = D \setminus D_f$ and the retrained reference model $M_r = A(D_r)$. Single image unlearning corresponds to the setting $D_f = \{(I_f, T_f)\}$, where forgetting removes a single image-text association while preserving utility on D_r. Unlearning proceeds by applying U to the original model and data to obtain $M_u = U(M_o, D, D_f)$. The unlearning objective requires the distribution induced by this procedure to be close to that of retraining, where closeness is measured over joint multimodal predictive outputs and model parameters through the induced distributions P_r and P_u:

$$P_r(A(D_r)) \approx P_u(U(M_o, D, D_f)).$$

To formalize approximate retraining equivalence, an (ε, δ) unlearning criterion is adopted to provide theoretical guarantees and to mirror stability notions from Differential Privacy (DP) (Dwork et al., 2006; Sekhari et al., 2021; Neel et al., 2021):

$$P[A(D \setminus D_f) \in R] \le e^{\varepsilon}\, P[U(A(D), D, D_f) \in R] + \delta,$$
$$P[U(A(D), D, D_f) \in R] \le e^{\varepsilon}\, P[A(D \setminus D_f) \in R] + \delta,$$

where R ranges over measurable events in the joint space of model parameters and multimodal predictive outputs. The pair of inequalities defines a symmetric divergence bound, ensuring that retraining and unlearning induce distributions that are mutually close up to (ε, δ). Probabilities $P[\cdot]$ are taken over the randomness of A and U and any evaluation sampling, with $\varepsilon = \delta = 0$ recovering exact retraining equivalence.

Optimization Objective. In multimodal models, forgetting is operationalized through a two term objective that suppresses responses associated with the forget set while preserving utility on the retained set across individual modalities and their fusion mechanisms:

$$\min_{\theta} \; J(\theta) \;=\; F_{\text{forget}}(\theta; D_f) \;+\; \lambda\, F_{\text{retain}}(\theta; D_r),$$

where F_{forget} reduces the influence of multimodal associations in D_f and F_{retain} preserves utility on retained multimodal dataset D_r. This abstraction separates deletion strength from utility preservation and establishes a common target for evaluating multimodal unlearning algorithms.

VLM Unlearning Overview. We define Vision-Language Model unlearning as reducing target cross-modal associations while retaining general utility; details appear in Appendix A.1

Diffusion Model Unlearning Overview. We define diffusion unlearning as diminishing target concept influence in the conditional denoising process while preserving generation quality; details appear in Appendix A.2

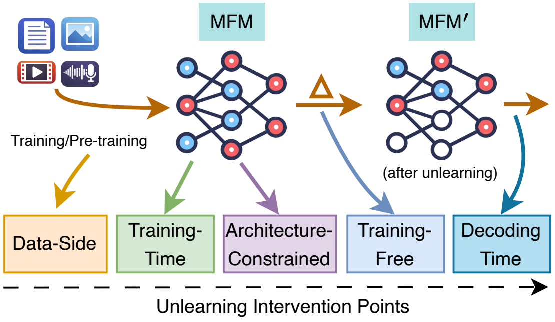

Figure 2: Unlearning intervention points for a Multimodal Foundation Model (MFM). Methods intervene at the data side, during training, via architecture-constrained edits, or at decoding time, producing an updated model (MFM′) with reduced influence from targeted content. Training-free methods use closed-form parameter or representation edits (denoted by Δ) to directly transform the model without retraining.

3 Multimodal Unlearning Methods

We organize multimodal unlearning methods by **forgetting target scope** and, within each scope, by the **intervention stage** and control mechanism in the multimodal pipeline (Figures 1 and 2).

3.1 Data-Side Interventions

Data-Path Perturbation Unlearning. Data-path perturbation unlearning edits inputs, not weights, to reduce the learnability of targeted clusters, pairs, or subjects while preserving utility on the remaining corpus (Zhang et al., 2023; Liu et al., 2024d; Sun et al., 2024; Liu et al., 2024e). Typical instantiations include cluster-wise perturbations, coupled image-text edits, segmentation-disrupting generators, and transformation-robust cloaks for personalization resistance. We view this as constrained perturbation design:

$$\|p_{\text{img}}(x)\| \le \epsilon_{\text{img}}, \quad \|p_{\text{txt}}(t)\| \le \epsilon_{\text{txt}}, \quad p \in \Pi_T,$$

where p perturbs target samples within image/text budgets and enforces robustness to common transforms T.

Data Hygiene and Prompt Normalization. Data hygiene reduces backdoor and trigger effects by curating or down-weighting suspicious image-text pairs, while prompt normalization canonicalizes visually or lexically similar tokens prior to optimization (Bansal et al., 2023; Struppek et al., 2024). As a data-side intervention, this design mitigates spurious correlations and preserves downstream utility without modifying model parameters. We summarize both operations as:

$$w(x,t) \in [0,1], \qquad t \mapsto N(t),$$

where $w(x,t)$ down-weights or removes flagged pairs and $N(\cdot)$ maps look-alike tokens or script variants to canonical forms. This abstraction highlights two complementary levers, corpus curation and prompt normalization, that mitigate spurious associations at the data and input levels.

3.2 Training-Time Edits

Direct Gradient. Direct gradient methods formulate unlearning as targeted risk minimization over a retain set and a forget set. The procedure first identifies behaviors to remove using curated data or token-level signals, then updates parameters so that responses on the forget set degrade while performance on retained data remains stable. In VLMs, this approach includes clean fine-tuning that disrupts poisoned cross-modal associations and objectives that decouple cross-modal structure from unimodal features (Bansal et al., 2023; Cheng and Amiri, 2024b). When only a small number of examples are available, token-localized updates and single-image objectives provide finer control (Liang et al., 2024a; Liu et al., 2024c; Li et al., 2024b). Related variants weaken spurious correspondences or isolate individual class associations through lightweight projections or regularization (Xing et al., 2024; Han et al., 2025a; Yang et al., 2024a; Kravets and Namboodiri, 2025a,b). Text-only unlearning in the language backbone further tests how forgetting transfers to visual inputs (Chakraborty et al., 2024). A generic objective used across contrastive and generative settings is:

$$\begin{aligned} J(\theta) &= \mathbb{E}_{(x_r,y_r)\in R}\, \mathcal{L}_u(f_\theta(x_r), y_r) \\ &\quad + \alpha\, \mathbb{E}_{(x_f,y_f)\in F}\, \mathcal{L}_f(f_\theta(x_f), y_f) \\ &\quad + \beta\, \mathbb{E}_{x_f\in F}\, D_a(f_\theta, x_f; a) \\ &\quad + \gamma\, \Omega(\theta, \theta_0), \end{aligned}$$

where the first term preserves utility on retained data, the second suppresses behavior on the forget set, the optional redirection term steers outputs away from forgotten content, and the regularizer limits deviation from a reference model.

Diffusion models instantiate this template through preference-aligned denoising, anchor redirection, or uncertainty-based objectives, while text-to-video variants apply similar updates to the shared text encoder (Park et al., 2024; Kumari et al., 2023; Li et al., 2024d; Liu and Tan, 2024; Spartalis et al., 2025). Audio and music systems adapt the same principle with task-specific losses that reduce speaker identity evidence, suppress memorized transcripts, or remove licensed content while preserving generation quality (Kim et al., 2025b; Liu, 2025; Pathak et al., 2025; Kim et al., 2025a).

Constrained Updates. Constrained update methods retain the locate-then-unlearn workflow but make the trade-off between forgetting and retention explicit. Instead of relying on unconstrained optimization, these approaches impose bounds that limit residual competence on the forget set and restrict deviation from a reference model while optimizing utility on retained data. The constraints act as tunable controls that specify acceptable levels of remaining harmful behavior and parameter change, which is especially relevant for repeated or deployment-facing unlearning (Schioppa et al., 2024; Wu and Harandi, 2025; Feng et al., 2025a). At a high level, forgetting is framed as constrained risk minimization,

$$\begin{aligned} \min_{\theta} \quad & \underbrace{\mathcal{J}_R(\theta)}_{\text{retain risk}} + \underbrace{\Omega(\theta,\theta_0)}_{\text{stability}} \\ \text{s.t.} \quad & \underbrace{\mathcal{C}_f(\theta)}_{\text{forget efficacy}} \le 0, \qquad \underbrace{\mathcal{C}_i(\theta)}_{\text{integrity}} \le 0 \end{aligned}$$

where the objective preserves performance on retained data through a stability prior, while the constraints enforce forgetting efficacy and model integrity relative to a reference checkpoint.

Existing methods differ primarily in how they instantiate these constraints and balance them during optimization. Joint constrained updates reconcile gradients for forgetting and utility (Wu and Harandi, 2025). Constraint-bounded solvers trace Pareto fronts for image-to-image editing tasks (Feng et al., 2025a). Integrity-aware formulations preserve perceptual similarity or enforce monotonic improvement across objectives (Schioppa et al., 2024; Ko et al., 2024). Related work applies importance-weighted deletion, knowledge tracing that removes fine-grained classes while retaining coarse recognition, or constrained recommendation updates that track divergence under user-level deletions (Alberti et al., 2025; Sinha et al., 2025; Li et al., 2025d).

Mask-Driven Selective Unlearning. Mask-driven methods follow the locate-then-unlearn workflow but constrain updates to a localized support identified through saliency, attention, or architectural structure. By restricting modification to parameters, features, spatial regions, or selected diffusion steps that most strongly encode the forget signal, these methods focus optimization where it matters while limiting collateral effects on retained behavior. Representative approaches include parameter-level masks derived from gradient or Fisher saliency (Fan et al., 2023; Huo et al., 2025), activation or spatial masks that suppress trigger-aligned attention (Zhang et al., 2024a; Jha et al., 2025), and diffusion-time masking schemes that update only a subset of denoising steps to stabilize multi-concept unlearning (Zhang et al., 2025a; Li et al., 2025c).

Distillation-Based Unlearning. Distillation-based unlearning follows the locate-then-unlearn paradigm by transferring behavior through a teacher-student setup, where the student is guided toward a safe target while retaining competence on non-forgotten prompts. Methods mainly differ in how the unlearning target is specified and how supervision is obtained. Existing work includes self-distillation that aligns conditional and unconditional predictions to suppress unsafe concepts (Kim et al., 2023), data-free distillation that relies on lightweight generators to approximate forget and retain distributions (Chen et al., 2025c), and attention-guided distillation that weakens adversarial trigger pathways during knowledge transfer (Aravindan et al., 2025). Across settings, distillation provides a training-time mechanism to redirect model behavior without direct access to original training data, while controlling drift relative to a reference model.

3.3 Architecture-Constrained Unlearning

Architecture Editing Unlearning. Architecture editing methods follow the locate-then-unlearn paradigm by modifying network structure through pruning, freezing, or controlled regrowth. Instead of reshaping the loss, these methods intervene directly in the computation graph to restrict pathways that encode the forget signal while limiting parameter drift elsewhere. Representative approaches include modality-aware pruning with light fine-tuning (Liu et al., 2025d), bilevel pruning coupled with suppression objectives (Shirkavand et al., 2025), freezing adaptation-critical tensors during downstream adaptation (Huang et al., 2024b), and prune-and-regrow strategies in audio models that restore capacity before fine-tuning on retained data (Mason-Williams et al., 2025). By confining updates to localized structural components, architecture editing can better preserve retained behavior than global parameter updates, although its success depends on precise localization of the forget signal and sufficient residual capacity in the remaining network.

Layer-Scoped Constrained Updates. Layer-scoped constrained updates follow locate-then-unlearn by first identifying where the target concept concentrates, then restricting edits to that support to limit collateral damage. SLUG (Cai et al., 2025) localizes the update to a selected layer to achieve targeted removal with minimal parameter drift. Model-integrity-controlled updates (Schioppa et al., 2024) instead constrain the update to preserve base behavior, typically by penalizing deviations from a reference model while enforcing forgetting efficacy.

3.4 Training-Free Unlearning

Weight-Space Linear Unlearning. Weight-space Linear Unlearning (WLU) follows the locate-then-unlearn paradigm but replaces iterative optimization with closed-form edits in parameter space. Instead of retraining, these methods modify a reference checkpoint through linear operations that suppress unwanted behavior while largely preserving retained utility. Representative instances include task-vector subtraction or negation (Ilharco et al., 2023), sign-consistent aggregation and weight negation (Kim et al., 2024a), low-rank suppression updates derived from safe and unsafe activations (Facchiano et al., 2025), and checkpoint averaging schemes that exclude shards associated with the forget data (Biggs et al., 2024).

Formally, WLU constructs an edited model θ' as a linear transformation of a reference model θ_0, where the direction and magnitude of the update encode the target behavior to remove. These edits remain training-free, composable across tasks, and easy to reverse, which makes WLU attractive when retraining is infeasible or when rapid post hoc control is required.

Representation Projection Unlearning. Representation Projection Unlearning (RPU) follows the locate-then-unlearn paradigm but replaces iterative optimization with closed-form edits in representation space. Instead of updating model parameters, these methods suppress target concepts by projecting internal activations or attention outputs away from a learned subspace associated with the forget signal. This strategy localizes change, limits collateral effects, and preserves overall model structure. Representative examples include CURE (Biswas et al., 2025), which projects joint embeddings to re-

Modality	Dataset	Size	Used in
	Identity Unlearning		
Image	CelebA (Liu et al., 2015)	202,599 images	Dai and Gifford, 2023; Dontsov et al., 2024; Huang et al., 2024a; Cai et al., 2025; Zhang et al., 2024c; Liu et al., 2025c
	CelebA-HQ (Karras et al., 2018)	30K high-quality images from CelebA	Huang et al., 2024a; Alberti et al., 2025; Nagasubramaniam et al., 2025
	Flickr-Faces HQ (Karras et al., 2019)	70K face images	Nagasubramaniam et al., 2025
	CASIA-WebFace (Yi et al., 2014)	494K face images	Dontsov et al., 2024
	FairFace (Karkkainen and Joo, 2021)	108,501 face images	Alabdulmohsin et al., 2024
	MillionCelebs (Zhang et al., 2020)	18.8M images of 636K identities	Dontsov et al., 2024
	VGGFace2 (Cao et al., 2018)	3.3M face images	Liu et al., 2024e; Li et al., 2025a
	PinsFaces (Burak, 2020)	17.5K cropped face photos	Kravets and Namboodiri, 2025a,c
Audio	VoxCeleb1 (Nagrani et al., 2017)	150K utterances from 1.3k speakers	Cheng and Amiri, 2025
	Affect and Video Unlearning		
Image	EmoSet (Yang et al., 2023)	3.3M images, 118K human-labeled with emotion and attributes.	Zhou et al., 2024b
	UnBiasedEmo (Panda et al., 2018)	3K affective images (6 emotion classes)	Zhou et al., 2024b
Video	UCF101 (Soomro et al., 2012)	13K videos across 101 action classes	Cheng and Amiri, 2024a
	Web-Scale Data Hygiene via Unlearning		
Image-Text	LAION-400M (Schuhmann et al., 2021)	400M CLIP-filtered image-text pairs	Poppi et al., 2024; Cai et al., 2025
	CC3M (Sharma et al., 2018)	3.3M web-harvested image-caption pairs	Bansal et al., 2023; Liang et al., 2024b; Han et al., 2025a
	Flickr30K (Young et al., 2014)	31K images with 158K captions	Alabdulmohsin et al., 2024; Liu et al., 2024d; Han et al., 2025a

Table 2: Key datasets commonly used in multimodal unlearning. Datasets are grouped by unlearning setting (**identity unlearning; affect and video unlearning; Web-Scale Data Hygiene via Unlearning**) and modality, with their sizes and representative studies.

move visual concepts, and related projection-based methods that operate on multimodal representation spaces (Poppi et al., 2024; De Matteis et al., 2025). The core operation applies an orthogonal projection that removes components aligned with the forget subspace:

$$h' = (I - UU^\top)\, h, \qquad W' = W\,(I - UU^\top),$$

where h denotes an intermediate representation, W an attention or projection matrix, and U a column-orthonormal basis spanning the forget subspace. The operator $I - UU^\top$ filters out directions linked to the target concept, yielding edited representations or projections without retraining. The effectiveness of RPU depends on how accurately the forget subspace is identified. Existing methods estimate U by factorizing attention features or by analyzing joint embedding statistics, which enables targeted suppression while keeping unrelated representations intact.

3.5 Decoding Time Unlearning

Guidance-Path Control. Guidance-path control performs locate-then-unlearn at decoding time by modifying the sampler rather than the model parameters. Instead of updating weights, these methods reshape the score used during generation to suppress target concepts while preserving visual quality and stylistic coherence. The base checkpoint remains fixed, enabling prompt-time selectivity and compatibility with standard sampling procedures, as in Dynamic Negative Guidance (Koulischer et al., 2025), UnGuide (Polowczyk et al., 2025), and Steering Guidance (Park et al., 2025), as well as detection-driven variants that combine concept identification with localized guidance to restrict unsafe content during generation (Li et al., 2025b; Yoon et al., 2025). A common formulation adjusts the predicted score at each denoising step:

$$\hat{\epsilon}_t = \epsilon_\theta(x_t, c) + a_t\big[\epsilon_{\text{alt}}(x_t, c) - \epsilon_\theta(x_t, c)\big] - b_t\, M_t\, d_t,$$

where x_t denotes the latent at step t, c the conditioning signal, and ϵ_θ the base predictor. The remaining terms introduce time-dependent steering, optional alternative guidance, and localized suppression through masks and direction vectors.

Benchmark	Modality	Unlearning Target	Task Type	Key Statistics	Evaluation Objective
Unified Benchmark Suites					
MU-Bench (Cheng and Amiri, 2024a)	Multimodal	Mixed (instances, datasets, modalities)	Multi-task	9 datasets, 20 architectures	Unified unlearning evaluation (efficacy, utility, efficiency)
MLLMU-Bench (Liu et al., 2025c)	VLM	Private data (fictitious & real identities)	Multi-task QA	500 fictitious and 153 public celebrities, 20.7K QA pairs	Privacy unlearning across efficacy, generalization, utility
PEBench (Xu et al., 2025b)	VLM	Synthetic identities & events	Multi-task	200 identities, 8K images, 16K QA pairs	Privacy and event unlearning with controlled scope and audits
UMU-Bench (Wang et al., 2025a)	VLM	knowledge instances	Multi-task	500 fictitious, 153 real	Modality-aligned unlearning completeness and utility
Identity and Privacy Unlearning					
CLEAR (Dontsov et al., 2024)	VLM	Identity	VQA	200 synthetic IDs, 3.7K images, 4K QA pairs	Identity leakage reduction with VQA accuracy retention
FIUBench (Ma et al., 2024b)	VLM	Identity	VQA	400 synthetic IDs, 8K QA pairs	Right-to-be-forgotten under privacy constraints
UnSLU-BENCH (Koudounas et al., 2025)	Audio	Speaker	Intent classification	Multi-speaker data, 4 languages	Speaker erasure with intent accuracy retention
Content and Knowledge Unlearning					
CPDM (Ma et al., 2024a)	DM	Styles/portraits	Generation	2.1K anchors, 18.9K generated images	Copyright similarity reduction with quality retention
UnlearnCanvas (Zhang et al., 2024d)	DM	Artistic styles	Generation	60 styles, 20 objects, high-res stylized images	Style forgetting with retention and generation fidelity/diversity
Holistic Unlearning (Moon et al., 2025)	DM	Mixed concepts	Generation	33 target concepts, 16k prompts per concept	Faithfulness, alignment, robustness, efficiency
Six-CD (Ren et al., 2025)	DM	Concept removal	Generation	Six concept categories, dual-version prompts	Cross category concept suppression with retainability checks
MMUBench (Li et al., 2024b)	VLM	Concept-level visual recognition	VQA	20 concepts, 50 images per concept	Concept-level visual unlearning with multimodal utility retention
UnLOK-VQA (Patil et al., 2024)	VLM	Targeted pretrained multimodal knowledge	VQA	500 samples with rephrase and neighborhood data	Privacy leakage reduction under attack-and-defense evaluation
SafeEraser (Chen et al., 2025a)	VLM	Harmful knowledge	VQA	3K images, 28.8K QA pairs	Harmful response reduction while preserving VQA utility

Table 3: Representative multimodal unlearning benchmarks grouped by unlearning target, reporting modality, task type, scale, and evaluation objective. Multimodal refers to image, text, audio, and video; VLM, Vision-Language Model; DM, Diffusion Model.

Conditioning-Path Control. Conditioning-path control performs locate-then-unlearn by modifying the conditioning signal that guides generation, while leaving model parameters unchanged. The sampler therefore operates under a weakened or safer condition for the target concept, which preserves inference latency and supports reversible control (Zhang et al., 2024b; Li et al., 2024c; Wang et al., 2024; Golatkar et al., 2024; Bui et al., 2025).

Let c denote the original conditioning input, such as a text embedding or a retrieval-augmented vector, and let s_θ be the conditional score used during sampling. Conditioning-path control constructs a transformed condition

$$c' \;=\; (1-\alpha)\,c \;+\; \alpha\,T(c, R, \text{policy}),$$

and then applies $s_\theta(x_t \mid c')$ at each denoising step. The scalar $\alpha \in [0, 1]$ controls the strength of intervention, R denotes an optional retrieval store, and T specifies the control mechanism.

Representative instantiations include projection toward a safe subspace in SteerDiff (Zhang et al., 2024b), policy-aware prompt rewriting and coordination in Moderator (Wang et al., 2024), hidden-key conditioning that gates concept activation (Bui et al., 2025), and retrieval mixing with selective deletion in CPR (Golatkar et al., 2024). These approaches share a common structure that alters conditioning pathways to suppress targeted concepts without retraining.

4 Datasets for Multimodal Unlearning

We organize datasets for multimodal unlearning by application setting and modality, and summarize them across four tables. Table 2 covers identity, affect, and video unlearning benchmarks, including face, emotion, and action datasets, and web-scale data hygiene removes noisy or sensitive alignments from large pretraining corpora. Table 4 focuses on personalization and copyright unlearning, capturing subject-specific and licensed content removal in generative models. Table 5 presents speech and safety robustness datasets used to study speaker, content, and jailbreak unlearning. Finally, Table 6 reports class-level unlearning benchmarks spanning image classification and segmentation settings

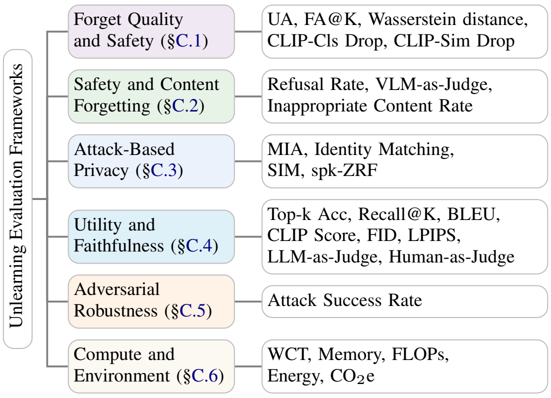
Unlearning Evaluation Frameworks
Forget Quality and Safety (§C.1)
UA, FA@K, Wasserstein distance, CLIP-Cls Drop, CLIP-Sim Drop
Safety and Content Forgetting (§C.2)
Refusal Rate, VLM-as-Judge, Inappropriate Content Rate
Attack-Based Privacy (§C.3)
MIA, Identity Matching, SIM, spk-ZRF
Utility and Faithfulness (§C.4)
Top-k Acc, Recall@K, BLEU, CLIP Score, FID, LPIPS, LLM-as-Judge, Human-as-Judge
Adversarial Robustness (§C.5)
Attack Success Rate
Compute and Environment (§C.6)
WCT, Memory, FLOPs, Energy, CO₂e

Figure 3: Evaluation dimensions and representative metrics for multimodal unlearning.

(Tables 4, 5, and 6 are in Appendix B).

5 Multimodal Unlearning Benchmarks

Multimodal unlearning has become central to addressing privacy, copyright, and safety concerns in vision-language and generative models. We review recent benchmarks that evaluate multimodal unlearning across diverse targets, modalities, and tasks. As summarized in Table 3, existing benchmarks range from unified suites spanning multiple datasets and architectures to task-specific evaluations of identity, privacy, content, and safety unlearning. These benchmarks support standardized comparisons and provide complementary evidence for unlearning efficacy, utility retention, robustness, and efficiency across vision, language, audio, and generative settings.

6 Evaluation Metrics Overview

Evaluation of multimodal unlearning relies on metric suites that jointly characterize forgetting, utility retention, robustness, and efficiency, as summarized in Figure 3. Prior work measures forgetting using targeted performance drops and concept-suppression signals, and complements these with safety and privacy audits that probe refusal behavior and membership or identity leakage. Retained capability is then verified on non-forgotten data using task and generation quality metrics, while robustness and practicality are assessed via adversarial stress tests and compute or environmental budgets. We defer metric definitions and protocols to Appendix C, which consolidates formulations and validation procedures across vision, language, audio, and generative settings.

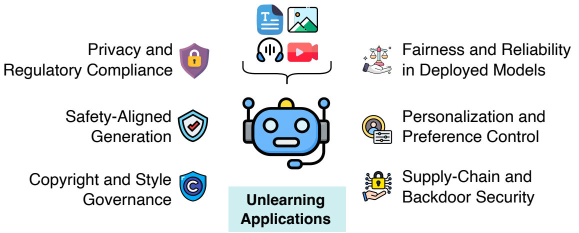
Privacy and Regulatory Compliance
Safety-Aligned Generation
Copyright and Style Governance
Unlearning Applications
Fairness and Reliability in Deployed Models
Personalization and Preference Control
Supply-Chain and Backdoor Security

Figure 4: Core application scenarios of multimodal unlearning across privacy, safety, governance, personalization, and security.

7 Multimodal Unlearning Applications

Multimodal unlearning supports deployed settings that require selective removal of learned information without full retraining. Figure 4 summarizes the primary application scenarios. Although application settings differ in targets, constraints, and evaluation priorities, they share a common objective: remove specific identities, attributes, concepts, or behaviors while preserving general capability and stability. We defer detailed use cases and representative studies to Appendix F.

8 Challenges and Future Directions

Multimodal unlearning still lacks certified deletion and stable evaluation. Open issues include cross-modal generalization, robustness to adaptive reactivation, and utility preservation under practical compute budgets. We note the main points here and defer detailed failure modes and future directions to Appendix G.

9 Conclusion

This survey presents a systematic review of multimodal unlearning as a core capability for accountable Multimodal Foundation Models (MFMs), with an emphasis on selective removal while preserving utility. By reviewing existing methods, highlighting emerging trends, and discussing open challenges, we adopt a system-oriented perspective that organizes unlearning mechanisms by intervention stage and control pathway, enabling comparison across vision, language, video, and audio models. Our synthesis highlights key gaps in evaluation reliability, robustness to adversarial reactivation, and deployment-facing constraints. Finally, we outline research directions toward unified benchmarks, stronger robustness guarantees, and tighter integration between unlearning mechanisms and deployment pipelines.

Limitations

This survey aims to provide broad coverage of multimodal unlearning for foundation models, but several limitations remain. First, despite systematic efforts to include relevant studies published before submission, some recent or less visible works may be omitted due to the rapid pace of progress in this area. Second, the analysis prioritizes system-level perspectives, such as intervention stages and control pathways, rather than method-centric or algorithm optimization-oriented perspectives. For detailed algorithmic design and optimization procedures, we encourage readers to refer to the original papers. Third, space constraints limit the depth of discussion for specific topics, including fine-grained taxonomies and modality-specific nuances, some of which appear in the appendix. In addition, the rapid evolution of methods, datasets, and evaluation protocols makes it challenging to maintain an entirely up-to-date and stable taxonomy. We hope this survey contributes to the ongoing development of multimodal unlearning and supports its relevance and utility in both academic and industrial settings. Finally, the lack of universally accepted benchmarks across modalities limits direct comparison and underscores the need for continued refinement of evaluation standards.

Acknowledgement

We thank Prof. Sijia Liu for helpful feedback and constructive suggestions.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and 1 others. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Ibrahim Alabdulmohsin, Xiao Wang, Andreas Peter Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. 2024. Clip the bias: How useful is balancing data in multimodal learning? In *The Twelfth International Conference on Learning Representations*.

Silas Alberti, Kenan Hasanaliyev, Manav Shah, and Stefano Ermon. 2025. Data unlearning in diffusion models. In *The Thirteenth International Conference on Learning Representations*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*.

Ashwath Vaithinathan Aravindan, Abha Jha, Matthew Salaway, Atharva Sandeep Bhide, and Duygu Nur Yaldiz. 2025. Sealing the backdoor: Unlearning adversarial text triggers in diffusion models using knowledge distillation. *arXiv preprint arXiv:2508.18235*.

Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. 2023. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. 2024. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*.

Praneeth Bedapudi. 2019. Nudenet: Neural nets for nudity classification, detection and selective censoring. https://github.com/bedapudi6788/NudeNet. GitHub repository.

Benjamin Biggs, Arjun Seshadri, Yang Zou, Achin Jain, Aditya Golatkar, Yusheng Xie, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Diffusion soup: Model merging for text-to-image diffusion models. In *ECCV (63)*.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying mmd gans. In *International Conference on Learning Representations*.

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

Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares-Salor, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. 2025. Digital forgetting in large language models: A survey of unlearning methods. *Artificial Intelligence Review*.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *European conference on computer vision*.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*.

Anh Tuan Bui, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. 2025. Hiding and recovering knowledge in text-to-image diffusion models via learnable prompts. In *ICLR Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*.

Burak. 2020. Pins face recognition. https://www.kaggle.com/datasets/hereisburak/pins-face-recognition. Kaggle dataset.

Zikui Cai, Yaoteng Tan, and M Salman Asif. 2025. Targeted unlearning with single layer unlearning gradient. In *Forty-second International Conference on Machine Learning*.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*.

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*.

Trishna Chakraborty, Erfan Shayegani, Zikui Cai, Nael Abu-Ghazaleh, M Salman Asif, Yue Dong, Amit Roy-Chowdhury, and Chengyu Song. 2024. Can textual unlearning solve cross-modality safety alignment? In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Junkai Chen, Zhijie Deng, Kening Zheng, Yibo Yan, Shuliang Liu, PeiJun Wu, Peijie Jiang, Jia Liu, and Xuming Hu. 2025a. Safeeraser: Enhancing safety in multimodal large language models through multimodal machine unlearning. *arXiv preprint arXiv:2502.12520*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pretraining for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

Siyi Chen, Yimeng Zhang, Sijia Liu, and Qing Qu. 2025b. The dual power of interpretable token embeddings: Jailbreaking attacks and defenses for diffusion model unlearning. *arXiv preprint arXiv:2504.21307*.

Tianqi Chen, Shujian Zhang, and Mingyuan Zhou. 2025c. Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models. In *The Thirteenth International Conference on Learning Representations*.

Yiwei Chen, Yuguang Yao, Yihua Zhang, Bingquan Shen, Gaowen Liu, and Sijia Liu. 2025d. Safety mirage: How spurious correlations undermine vlm safety fine-tuning. *arXiv preprint arXiv:2503.11832*.

Jiali Cheng and Hadi Amiri. 2024a. Mu-bench: A multitask multimodal benchmark for machine unlearning. *arXiv preprint arXiv:2406.14796*.

Jiali Cheng and Hadi Amiri. 2024b. Multidelete for multimodal machine unlearning. In *European Conference on Computer Vision*.

Jiali Cheng and Hadi Amiri. 2025. Speech unlearning. *arXiv preprint arXiv:2506.00848*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Bartosz Cywiński and Kamil Deja. 2025. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Forty-second International Conference on Machine Learning*.

Juntao Dai, Tianle Chen, Xuyao Wang, Ziran Yang, Taiye Chen, Jiaming Ji, and Yaodong Yang. 2024. Safesora: Towards safety alignment of text2video generation via a human preference dataset. *Advances in Neural Information Processing Systems*.

Zheng Dai and David K Gifford. 2023. Training data attribution for diffusion models. *arXiv preprint arXiv:2306.02174*.

Pucheng Dang, Xing Hu, Dong Li, Rui Zhang, Qi Guo, and Kaidi Xu. 2025a. Diffzoo: A purely query-based black-box attack for red-teaming text-to-image generative model via zeroth order optimization. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Yizhou Dang, Yuting Liu, Enneng Yang, Guibing Guo, Linying Jiang, Jianzhe Zhao, and Xingwei Wang. 2025b. Efficient and adaptive recommendation unlearning: A guided filtering framework to erase outdated preferences. *ACM Transactions on Information Systems*.

Edoardo De Matteis, Matteo Migliarini, Alessio Sampieri, Indro Spinelli, and Fabio Galasso. 2025. Human motion unlearning. *arXiv preprint arXiv:2503.18674*.

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

Yi Ding, Lijun Li, Bing Cao, and Jing Shao. 2025. Rethinking bottlenecks in safety fine-tuning of vision language models. *arXiv preprint arXiv:2501.18533*.

Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A Smith, Nicole DeCario, and Will Buchanan. 2022. Measuring the carbon intensity of ai in cloud instances. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Y Rogov, Ivan Oseledets, and Elena Tutubalina. 2024. Clear: Character unlearning in textual and visual modalities. *arXiv preprint arXiv:2410.18057*.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*.

Simone Facchiano, Stefano Saravalle, Matteo Migliarini, Edoardo De Matteis, Alessio Sampieri, Andrea Pilzer, Emanuele Rodolà, Indro Spinelli, Luca Franco, and Fabio Galasso. 2025. Video unlearning via low-rank refusal vector. *arXiv preprint arXiv:2506.07891*.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. 2023. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations*.

XiaoHua Feng, Yuyuan Li, Chaochao Chen, Li Zhang, Longfei Li, JUN ZHOU, and Xiaolin Zheng. 2025a. Controllable Unlearning for Image-to-Image Generative Models via ε-Constrained Optimization. In *The Thirteenth International Conference on Learning Representations*.

Xiaohua Feng, Jiaming Zhang, Fengyuan Yu, Chengye Wang, Li Zhang, Kaixiang Li, Yuyuan Li, Chaochao Chen, and Jianwei Yin. 2025b. A survey on generative model unlearning: Fundamentals, taxonomy, evaluation, and future direction. *arXiv preprint arXiv:2507.19894*.

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

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

Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. 2024a. On large language model continual unlearning. *arXiv preprint arXiv:2407.10223*.

Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. 2024b. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. *arXiv preprint arXiv:2410.12777*.

Jiahui Geng, Qing Li, Herbert Woisetschlaeger, Zongxiong Chen, Fengyu Cai, Yuxia Wang, Preslav Nakov, Hans-Arno Jacobsen, and Fakhri Karray. 2025. A comprehensive survey of machine unlearning techniques for large language models. *arXiv preprint arXiv:2503.01854*.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. 2019. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*.

Aditya Golatkar, Alessandro Achille, Luca Zancato, Yu-Xiang Wang, Ashwin Swaminathan, and Stefano Soatto. 2024. Cpr: Retrieval augmented generation for copyright protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Eric Goldman. 2020. An introduction to the california consumer privacy act (ccpa). *Santa Clara Univ. Legal Studies Research Paper*.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. 2025. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. 2020. Certified data removal from machine learning models. In *International Conference on Machine Learning*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Haochen Han, Alex Jinpeng Wang, Peijun Ye, and Fangming Liu. 2025a. Unlearning the noisy correspondence makes clip more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Xiaoxuan Han, Songlin Yang, Wei Wang, Yang Li, and Jing Dong. 2024. Probing unlearned diffusion models: A transferable adversarial attack perspective. *arXiv preprint arXiv:2404.19382*.

Xiaoxuan Han, Songlin Yang, Wei Wang, Yang Li, and Jing Dong. 2025b. Adaptive median smoothing: Adversarial defense for unlearned text-to-image diffusion models at inference time. In *Forty-second International Conference on Machine Learning*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Jeremy Howard. 2019. Imagenette: A smaller subset of 10 easily classified classes from imagenet. https://github.com/fastai/imagenette. GitHub repository.

Xuhao Hu, Dongrui Liu, Hao Li, Xuan-Jing Huang, and Jing Shao. 2025. Vlsbench: Unveiling visual leakage in multimodal safety. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Huaxi Huang, Xin Yuan, Qiyu Liao, Dadong Wang, and Tongliang Liu. 2024a. Enhancing user-centric privacy protection: An interactive framework through diffusion models and machine unlearning. *arXiv preprint arXiv:2409.03326*.

Kai Huang, Haoming Wang, and Wei Gao. 2024b. Freezeasguard: Mitigating illegal adaptation of diffusion models via selective tensor freezing. *arXiv preprint arXiv:2405.17472*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, and 1 others. 2024c. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, and Xuming Hu. 2025. Mmunlearner: Reformulating multimodal machine unlearning in the era of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Abha Jha, Ashwath Vaithinathan Aravindan, Matthew Salaway, Atharva Sandeep Bhide, and Duygu Nur Yaldiz. 2025. Backdoor defense in diffusion models via spatial attention unlearning. *arXiv preprint arXiv:2504.18563*.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. 2023. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*.

Er Jin, Yang Zhang, Yongli Mou, Yanfei Dong, Stefan Decker, Kenji Kawaguchi, and Johannes Stegmaier. 2025. Unconsciously forget: Mitigating memorization; without knowing what is being memorized. *arXiv preprint arXiv:2512.09687*.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Tatsuki Kawakami, Kazuki Egashira, Atsuyuki Miyai, Go Irie, and Kiyoharu Aizawa. 2025. Pulse: Practical evaluation scenarios for large multimodal model unlearning. *arXiv preprint arXiv:2507.01271*.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*.

Hyoseo Kim, Dongyoon Han, and Junsuk Choe. 2024a. Negmerge: Consensual weight negation for strong machine unlearning. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.

Jinju Kim, Taehan Kim, Abdul Waheed, and Rita Singh. 2025a. No encore: Unlearning as opt-out in music generation. *arXiv preprint arXiv:2509.06277*.

Minseon Kim, Hyomin Lee, Boqing Gong, Huishuai Zhang, and Sung Ju Hwang. 2024b. Automatic jailbreaking of the text-to-image generative ai systems. In *ICML 2024 Next Generation of AI Safety Workshop*.

Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. 2023. Towards safe self-distillation of internet-scale text-to-image diffusion models. *arXiv preprint arXiv:2307.05977*.

Taesoo Kim, Jinju Kim, Dong Chan Kim, Jong Hwan Ko, and Gyeong-Moon Park. 2025b. Do not mimic my voice: Speaker identity unlearning for zero-shot text-to-speech. In *ICML 2025 Workshop on Machine Unlearning for Generative AI*.

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

Myeongseob Ko, Henry Li, Zhun Wang, Jonathan Patsenker, Jiachen Tianhao Wang, Qinbin Li, Ming Jin, Dawn Song, and Ruoxi Jia. 2024. Boosting alignment for post-unlearning text-to-image generative models. *Advances in Neural Information Processing Systems*.

Alkis Koudounas, Moreno La Quatra, Lorenzo Vaiani, Luca Colomba, Giuseppe Attanasio, Eliana Pastor, Luca Cagliero, and Elena Baralis. 2023. Italic: An italian intent classification dataset. In *INTERSPEECH*.

Alkis Koudounas, Claudio Savelli, Flavio Giobergia, and Elena Baralis. 2025. "Alexa, can you forget me?" machine unlearning benchmark in spoken language understanding. *arXiv preprint arXiv:2505.15700*.

Felix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester, and Luca Ambrogioni. 2025. Dynamic negative guidance of diffusion models. In *The Thirteenth International Conference on Learning Representations*.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*.

Alexey Kravets and Vinay P Namboodiri. 2025a. Zero-shot class unlearning in clip with synthetic samples. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Alexey Kravets and Vinay P Namboodiri. 2025b. Zero-shot clip class forgetting via text-image space adaptation. *Transactions on Machine Learning Research*.

Alexey Kravets and Vinay P Namboodiri. 2025c. Zero-shot clip class forgetting via text-image space adaptation. *Transactions on Machine Learning Research*.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. Technical report, Department of Computer Science, University of Toronto. Technical Report.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. 2023. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 2002. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.

Boheng Li, Renjie Gu, Junjie Wang, Leyi Qi, Yiming Li, Run Wang, Zhan Qin, and Tianwei Zhang. 2025a. Towards resilient safety-driven unlearning for diffusion models against downstream fine-tuning. *arXiv preprint arXiv:2507.16302*.

Feifei Li, Mi Zhang, Yiming Sun, and Min Yang. 2025b. Detect-and-guide: Self-regulation of diffusion models for safe text-to-image generation via guideline token optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Gen Li, Yang Xiao, Jie Ji, Kaiyuan Deng, Bo Hui, Linke Guo, and Xiaolong Ma. 2025c. Sculpting memory: Multi-concept forgetting in diffusion models via dynamic mask and concept-aware optimization. *arXiv preprint arXiv:2504.09039*.

Guihong Li, Hsiang Hsu, Chun-Fu Chen, and Radu Marculescu. 2024a. Machine unlearning for image-to-image generative models. *arXiv preprint arXiv:2402.00351*.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. 2024b. Single image unlearning: Efficient machine unlearning in multimodal large language models. *Advances in Neural Information Processing Systems*.

Senmao Li, Joost van de Weijer, Fahad Khan, Qibin Hou, Yaxing Wang, and 1 others. 2024c. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*.

Xiang Li, Qianli Shen, Haonan Wang, and Kenji Kawaguchi. 2025d. Loreun: Data itself implicitly provides cues to improve machine unlearning. *arXiv preprint arXiv:2507.22499*.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. 2024d. Safegen: Mitigating unsafe content generation in text-to-image models. *CoRR*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Siyuan Liang, Kuanrong Liu, Jiajun Gong, Jiawei Liang, Yuan Xun, Ee-Chien Chang, and Xiaochun Cao. 2024a. Unlearning backdoor threats: Enhancing backdoor defense in multimodal contrastive learning via local token unlearning. *arXiv preprint arXiv:2403.16257*.

Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. 2024b. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

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

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*.

Chris Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024a. Large language model unlearning via embedding-corrupted prompts. *Advances in Neural Information Processing Systems*.

Hongbin Liu, Wenjie Qu, Jinyuan Jia, and Neil Zhenqiang Gong. 2024b. Pre-trained encoders in self-supervised learning improve secure and privacy-preserving supervised learning. In *2024 IEEE Security and Privacy Workshops (SPW)*.

Kuanrong Liu, Siyuan Liang, Jiawei Liang, Pengwen Dai, and Xiaochun Cao. 2024c. Efficient backdoor defense in multimodal contrastive learning: A token-level unlearning method for mitigating threats. *arXiv preprint arXiv:2409.19526*.

Renyang Liu, Guanlin Li, Tianwei Zhang, and See-Kiong Ng. 2025a. Image can bring your memory back: A novel multi-modal guided attack against image generation model unlearning. *arXiv preprint arXiv:2507.07139*.

Shiqi Liu and Yihua Tan. 2024. Unlearning concepts from text-to-video diffusion models. *arXiv preprint arXiv:2407.14209*.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, and 1 others. 2025b. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*.

Xinwei Liu, Xiaojun Jia, Yuan Xun, Siyuan Liang, and Xiaochun Cao. 2024d. Multimodal unlearnable examples: Protecting data against multimodal contrastive learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. 2024e. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhe Liu. 2025. Unlearning llm-based speech recognition models. In *Proc. Interspeech 2025*.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. 2025c. Protecting privacy in multimodal large language models with mllmu-bench. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024f. Machine unlearning in generative ai: A survey. *arXiv preprint arXiv:2407.20516*.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. 2025d. Modality-aware neuron pruning for unlearning in multimodal large language models. *arXiv preprint arXiv:2502.15910*.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*.

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

Rui Ma, Qiang Zhou, Bangjun Xiao, Yizhu Jin, Daquan Zhou, Xiuyu Li, Aishani Singh, Yi Qu, Kurt Keutzer, Xiaodong Xie, and 1 others. 2024a. A dataset and benchmark for copyright protection from text-to-image diffusion models. *arXiv preprint arXiv:2403.12052*.

Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Jinsheng Pan, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, and 1 others. 2024b. Benchmarking vision language model unlearning via fictitious facial identity dataset. *arXiv preprint arXiv:2411.03554*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*.

Israel Mason-Williams, Jing Han, Helen Yannakoudakis, and Cecilia Mascolo. 2025. Machine unlearning in audio: Bridging the modality gap via the prune and regrow paradigm.

Saemi Moon, Minjong Lee, Sangdon Park, and Dongwoo Kim. 2025. Holistic unlearning benchmark: A multi-faceted evaluation for text-to-image diffusion model unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Piyush Nagasubramaniam, Neeraj Karamchandani, Chen Wu, and Sencun Zhu. 2025. Prompting forgetting: Unlearning in gans via textual guidance. *arXiv preprint arXiv:2504.01218*.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2021. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, and 1 others. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*.

Jiadong Pan, Hongcheng Gao, Zongyu Wu, Taihang Hu, Li Su, Qingming Huang, and Liang Li. 2024. Leveraging catastrophic forgetting to develop safe diffusion models against malicious finetuning. *Advances in Neural Information Processing Systems*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*.

Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. 2018. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Sunghyun Park, Seokeon Choi, Hyoungwoo Park, and Sungrack Yun. 2025. Steering guidance for personalized text-to-image diffusion models. *arXiv preprint arXiv:2508.00319*.

Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. 2024. Direct unlearning optimization for robust and safe text-to-image models. *Advances in Neural Information Processing Systems*.

Shreyansh Pathak, Sonu Shreshtha, Richa Singh, and Mayank Vatsa. 2025. Quantum-inspired audio unlearning: Towards privacy-preserving voice biometrics. *arXiv preprint arXiv:2507.22208*.

Vaidehi Patil, Yi-Lin Sung, Peter Hase, Jie Peng, Tianlong Chen, and Mohit Bansal. 2024. Unlearning sensitive information in multimodal llms: Benchmark and attack-defense evaluation. *Transactions on Machine Learning Research*.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. 2025. Dreambench++: A human-aligned benchmark for personalized image generation. In *The Thirteenth International Conference on Learning Representations*.

Agnieszka Polowczyk, Alicja Polowczyk, Dawid Malarz, Artur Kasymov, Marcin Mazur, Jacek Tabor, PrzemysĹ Spurek, and 1 others. 2025. Unguide: Learning to forget with lora-guided diffusion models. *arXiv preprint arXiv:2508.05755*.

Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Safe-clip: Removing nsfw concepts from vision-and-language models. In *Computer Vision – ECCV 2024*.

Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang Tang, and Lingjuan Lyu. 2025. Six-cd: Benchmarking concept removals for text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Matan Rusanovsky, Shimon Malnick, Amir Jevnisek, Ohad Fried, and Shai Avidan. 2025. Memories of forgotten concepts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Babak Saleh and Ahmed Elgammal. 2015. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*.

Andrea Schioppa, Emiel Hoogeboom, and Jonathan Heek. 2024. Model integrity when unlearning with t2i diffusion models. *arXiv preprint arXiv:2411.02068*.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. 2021. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*.

Marcin Sendera, Łukasz Struski, Kamil Książek, Kryspin Musiol, Jacek Tabor, and Dawid Rymarczyk. 2025. Semu: Singular value decomposition for efficient machine unlearning. *arXiv preprint arXiv:2502.07587*.

Aakash Sen Sharma, Niladri Sarkar, Vikram Chundawat, Ankur A Mali, and Murari Mandal. 2024. Unlearning or concealment? a critical analysis and evaluation metrics for unlearning in diffusion models. *arXiv preprint arXiv:2409.05668*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Reza Shirkavand, Peiran Yu, Shangqian Gao, Gowthami Somepalli, Tom Goldstein, and Heng Huang. 2025. Efficient fine-tuning and concept suppression for pruned diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*.

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*.

Yash Sinha, Murari Mandal, and Mohan Kankanhalli. 2025. Multi-modal recommendation unlearning for legal, licensing, and modality constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. 2014. Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics (ToG)*.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Diana Sousa, André Lamúrias, and Francisco M Couto. 2019. A silver standard corpus of human phenotype-gene relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Christoforos N Spartalis, Theodoros Semertzidis, Petros Daras, and Stratis Gavves. 2025. Unleashing uncertainty: Efficient machine unleanring for generative ai. In *ICML 2025 Workshop on Machine Unlearning for Generative AI*.

Lukas Struppek, Dominik Hintersdorf, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. 2024. Exploiting cultural biases via homoglyphs intext-to-image synthesis. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Ye Sun, Hao Zhang, Tiehua Zhang, Xingjun Ma, and Yu-Gang Jiang. 2024. Unseg: one universal unlearnable example generator is enough against all image segmentation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*.

Vinith M Suriyakumar, Rohan Alur, Ayush Sekhari, Manish Raghavan, and Ashia C Wilson. 2024. Unstable unlearning: The hidden risk of concept resurgence in diffusion models. *arXiv preprint arXiv:2410.08074*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Alexander Y Tong, Guillaume Huguet, Amine Natik, Kincaid MacDonald, Manik Kuchroo, Ronald Coifman, Guy Wolf, and Smita Krishnaswamy. 2021. Diffusion earth mover's distance and distribution embeddings. In *International Conference on Machine Learning*.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. Fvd: A new metric for video generation. *ICLR 2019 Workshop DeepGenStruct*.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*.

Chengye Wang, Yuyuan Li, XiaoHua Feng, Chaochao Chen, Xiaolin Zheng, and Jianwei Yin. 2025a. Umubench: Closing the modality gap in multimodal unlearning evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Peiran Wang, Qiyu Li, Longxuan Yu, Ziyao Wang, Ang Li, and Haojian Jin. 2024. Moderator: Moderating text-to-image diffusion models through fine-grained context-based policies. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*.

Yihan Wang, Yiwei Lu, Guojun Zhang, Franziska Boenisch, Adam Dziedzic, Yaoliang Yu, and Xiao-Shan Gao. 2025b. Muc: Machine unlearning for contrastive learning with black-box evaluation. *Transactions on Machine Learning Research*.

Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. 2025c. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhiqi Wang, Chengyu Zhang, Yuetian Chen, Nathalie Baracaldo, Swanand Kadhe, and Lei Yu. 2025d. Membership inference attacks as privacy tools: Reliability, disparity and ensemble. *arXiv preprint arXiv:2506.13972*.

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.

Fengli Wu, Vaidehi Patil, Jaehong Yoon, Yue Zhang, and Mohit Bansal. 2025a. Medforget: Hierarchy-aware multimodal unlearning testbed for medical ai. *arXiv preprint arXiv:2512.09867*.

Jing Wu and Mehrtash Harandi. 2025. Munba: Machine unlearning via nash bargaining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. 2024. Erasediff: Erasing data influence in diffusion models. *arXiv preprint arXiv:2401.05779*.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. 2025b. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. 2016. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*.

Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2024. Efuf: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Naen Xu, Jinghuai Zhang, Changjiang Li, Zhi Chen, Chunyi Zhou, Qingming Li, Tianyu Du, and Shouling Ji. 2025a. Videoeraser: Concept erasure in text-to-video diffusion models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Zhaopan Xu, Pengfei Zhou, Weidong Tang, Jiaxin Ai, Wangbo Zhao, Kai Wang, Xiaojiang Peng, Wenqi Shao, Hongxun Yao, and Kaipeng Zhang. 2025b. Pebench: A fictitious dataset to benchmark machine unlearning for multimodal large language models. *arXiv preprint arXiv:2503.12545*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. 2023. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Tianyu Yang, Lisen Dai, Xiangqi Wang, Minhao Cheng, Yapeng Tian, and Xiangliang Zhang. 2024a. Cliperase: Efficient unlearning of visual-textual associations in clip. *arXiv preprint arXiv:2410.23330*.

Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. 2024b. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE symposium on security and privacy (SP)*.

Yelp Inc. 2023. Yelp open dataset. https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset. Accessed: 2026-01-01.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*.

Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. 2025. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *The Thirteenth International Conference on Learning Representations*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*.

Chaoshuo Zhang, Chenhao Lin, Zhengyu Zhao, Le Yang, Qian Wang, and Chao Shen. 2025a. Concept unlearning by modeling key steps of diffusion process. *arXiv preprint arXiv:2507.06526*.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 2024a. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Hongxiang Zhang, Yifeng He, and Hao Chen. 2024b. Steerdiff: Steering towards safe text-to-image diffusion models. *arXiv preprint arXiv:2410.02710*.

Jiaming Zhang, Xingjun Ma, Qi Yi, Jitao Sang, Yu-Gang Jiang, Yaowei Wang, and Changsheng Xu. 2023. Unlearnable clusters: Towards label-agnostic unlearnable examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jihai Zhang, Xiang Lan, Xiaoye Qu, Yu Cheng, Mengling Feng, and Bryan Hooi. 2024c. Learning the unlearned: Mitigating feature suppression in contrastive learning. In *European Conference on Computer Vision*.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Xianren Zhang, Hui Liu, Delvin Ce Zhang, Xianfeng Tang, Qi He, Dongwon Lee, and Suhang Wang. 2025b. Does multimodal large language model truly unlearn? stealthy mllm unlearning attack. *arXiv preprint arXiv:2506.17265*.

Xianren Zhang, Hui Liu, Delvin Ce Zhang, Xianfeng Tang, Qi He, Dongwon Lee, and Suhang Wang. 2025c. Sua: Stealthy multimodal large language model unlearning attack. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Yaobin Zhang, Weihong Deng, Mei Wang, Jiani Hu, Xian Li, Dongyue Zhao, and Dongchao Wen. 2020. Global-local gcn: Large-scale label noise cleansing for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yihua Zhang, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Xiaoming Liu, and Sijia Liu. 2024d. Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models. *arXiv preprint arXiv:2402.11846*.

Mengnan Zhao, Lihe Zhang, Tianhang Zheng, Yuqiu Kong, and Baocai Yin. 2024. Separable multi-concept erasure from diffusion models. *arXiv preprint arXiv:2402.05947*.

Shiji Zhou, Lianzhe Wang, Jiangnan Ye, Yongliang Wu, and Heng Chang. 2024a. On the limitations and prospects of machine unlearning for generative ai. *arXiv preprint arXiv:2408.00376*.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. 2024b. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics ACL 2024*.

A Model-Specific Unlearning Formulations

A.1 Formulation of VLM Unlearning

VLM unlearning targets the components that bind vision and language, supporting both instance-level and concept-level removal, while keeping unimodal competence intact. Let a VLM comprise a vision encoder f_v, a text encoder f_t, and a fusion head F. Given forget pairs $D_f = \{(x, c_f)\}$ that align an image x with a forget concept prompt c_f and retain pairs D_r, a compact objective balances suppression and utility:

$$\min_{\theta \in \{\theta_v, \theta_{\text{fusion}}\}} \; L_{\text{retain}}(D_r;\theta) \;+\; \lambda\, L_{\text{forget}}(D_f;\theta) \\ +\, \mu\, \Omega(\theta).$$

A concept hinge decouples semantics by penalizing violations of $S_\theta(x, c_f) \le m$ for $(x, c_f) \in D_f$, where m is a similarity threshold that sets the target upper bound on forget-pair similarity, while a consistency or caption term preserves performance on D_r (Li et al., 2024b). Selective updates use a saliency mask S so that

$$\Delta\theta \;=\; -\eta\, S \odot \nabla_\theta \big(L_{\text{forget}} + \lambda\, L_{\text{retain}}\big)\,,$$

which concentrates edits in visual or cross-attention paths and leaves the text encoder frozen (Huo et al., 2025). Instance-level removal uses the same form with c_f tied to specific pairs, whereas concept-level removal suppresses all realizations of c_f and projects away its anchor in the joint space (Kravets and Namboodiri, 2025a). Empirically, pair decoupling narrows the utility gap to retraining while achieving strong reductions in cross-modal affinity for deleted content.

A.2 Formulation of DM Unlearning

Diffusion Model unlearning focuses on the conditional denoising path tied to a target concept. Let $\epsilon_\theta(x_t, c, t)$ denote the denoiser with conditioning c. A teacher guided loss attenuates the target channel,

$$L_{\text{forget}} = \mathbb{E}\Big[\|\epsilon_\theta(x_t, c_f, t) - \tilde{\epsilon}(x_t, t)\|_2^2\Big],$$
$$L_{\text{retain}} = \mathbb{E}\Big[\|\epsilon(x_t, t) - \epsilon_\theta(x_t, c_r, t)\|_2^2\Big],$$

so ϵ_θ aligns with an unconditional or safe teacher on c_f while generation quality on D_r remains stable (Gandikota et al., 2023; Zhang et al., 2024a). Representation editing complements loss shaping by modifying cross-attention: keys and values associated with c_f are mapped to neutral surrogates, implemented as low rank or sparse updates $W_{\text{attn}} \leftarrow W_{\text{attn}} - \alpha\, \Pi_{c_f}$ across timesteps (Kumari et al., 2023; Gandikota et al., 2024). Sampling time steering reduces classifier-free guidance s or injects negative prompts to deflect c_f without weight changes (Zhang et al., 2024a). Multi-concept settings apply separable edits that localize interference (Zhao et al., 2024). Recent analyses show concept revival under benign fine-tuning, which motivates explicit anti revival regularizers and robustness checks in the evaluation loop (Suriyakumar et al., 2024; Wu et al., 2024). These ingredients define the diffusion toolkit used by current unlearning methods.

B Additional Dataset Details

Several specialized unlearning settings rely on targeted datasets to evaluate concept-level or domain-specific forgetting. Table 4 covers personalization setup and copyright unlearning, as well as knowledge QA and instruction probes for factual or behavioral erasure in vision-language tasks, segmentation and image-to-image unlearning for pixel-level concepts or stylistic attributes, and recommender unlearning for user-item interactions. Table 5 presents speech unlearning, which targets speaker traits and linguistic content, and safety robustness unlearning, which evaluates resistance to jailbreak prompts and refusal consistency. Finally, Table 6 reports class unlearning benchmarks that evaluate the removal of entire semantic categories in classifiers using standard image datasets.

C Detailed Unlearning Evaluation Frameworks

C.1 Forget Quality and Safety

Unlearning Accuracy. Unlearning Accuracy (UA) measures forgetting efficacy as the complement of predictive accuracy on the forget set (Wu and Harandi, 2025; Schioppa et al., 2024; Sendera et al., 2025):

$$\text{UA} = 100\% - \text{Accuracy}(D_f),$$

where D_f denotes the subset designated for removal. Related forgetting-oriented metrics include Forget Accuracy, which reports post-unlearning accuracy on the forbidden class (Pathak et al., 2025), and Removal Accuracy, which measures the fraction of attack triggers that no longer elicit the undesired behavior (Aravindan et al., 2025; Jha et al., 2025).

Zero-Shot Forget Accuracy (FA@k). For VLMs with zero-shot prediction, FA@k measures whether the true label of a forget example appears among the top-k model predictions. Given a forget set D_f and model scores $f(x)$,

$$\text{FA@}k = \frac{1}{|D_f|} \sum_{(x,y)\in D_f} \mathbf{1}\big\{\, y \in \text{Top-k}\big(f(x)\big) \,\big\}.$$

This metric is commonly reported for $k \in \{1, 5\}$ in zero-shot VLM evaluations (Cai et al., 2025).

Degree of Unlearning. Distributional change in concept scores before and after unlearning can be quantified using the 1-Wasserstein distance. Let B denote the pre-unlearning score distribution, A the post-unlearning distribution, and R a reference distribution. The degree of unlearning is defined as

$$\gamma = \frac{W_1(A, B)}{W_1(B, R)},$$

where $W_1(\cdot,\cdot)$ denotes the 1-Wasserstein distance (Solomon et al., 2014; Tong et al., 2021).

CLIP Classification Drop. Concept erasure in image generation can be verified through classification performance on generated samples. Let a generator produce n images for a concept prompt before unlearning, $\{x_i^{\text{pre}}\}_{i=1}^n$, and after unlearning, $\{x_i^{\text{post}}\}_{i=1}^n$. Using a zero-shot CLIP classifier or a specialized detector $c(\cdot) \in \{0, 1\}$, the classification drop is computed as

$$\Delta_{\text{cls}} = \frac{1}{n}\sum_{i=1}^{n} c(x_i^{\text{pre}}) - \frac{1}{n}\sum_{i=1}^{n} c\Big(x_i^{\text{post}}\Big).$$

A higher Δ_{cls} indicates greater removal of the target concept from generated outputs. CLIP-based classification accuracy serves as a standard erasure indicator such as ESD (Gandikota et al., 2023), MACE (Lu et al., 2024).

CLIP Similarity Drop. CLIP image-text similarity provides a continuous signal of residual concept alignment. Using the same image sets and the concept text t, let $f_{\text{img}}, f_{\text{text}}$ be CLIP encoders and let $\cos(\cdot,\cdot)$ denote cosine similarity. Define average similarities

$$s_{\text{pre}} = \frac{1}{n}\sum_{i=1}^{n} \cos\big(f_{\text{img}}(x_i^{\text{pre}}),\ f_{\text{text}}(t)\big)$$

$$s_{\text{post}} = \frac{1}{n}\sum_{i=1}^{n} \cos\big(f_{\text{img}}(x_i^{\text{post}}),\ f_{\text{text}}(t)\big)$$

and the similarity drop $\Delta_{\text{sim}} = s_{\text{pre}} - s_{\text{post}}$. When Δ_{sim} increases, alignment with the concept decreases. Empirical reports show that classifier confidence can collapse while CLIP similarity falls only slightly, so reporting both measures is helpful for diagnosing residual representations (Gandikota et al., 2023; Rusanovsky et al., 2025; Wang et al., 2025c).

C.2 Safety & Content Forgetting

Refusal Rate on Forbidden Prompts. Also referred to as rejection rate, Refusal Rate (RR) measures how often the model refuses harmful queries after unlearning (Chen et al., 2025d). Let D be the evaluation set of harmful text-image inputs and R_i the model response to the i-th prompt. Define the refusal indicator $I_{\text{ref}}(R_i) = 1$ if the response contains refusal content (per a predefined policy template) and 0 otherwise. The metric is

$$RR = \frac{1}{|D|}\sum_{i=1}^{|D|} I_R(R_i)\,,$$

so higher RR indicates more consistent rejection of harmful requests.

Inappropriate Content Rate. This metric measures how often a model produces unsafe content under sensitive prompts. In image generation, a standard protocol samples outputs and reports the fraction flagged by external NSFW detectors (e.g., Q16 or NudeNet), where lower post-unlearning rates indicate safer behavior (Schramowski et al., 2023). Let $Y_{\text{pre}} = \{y_i^{\text{pre}}\}_{i=1}^{n}$ and $Y_{\text{post}} = \{y_i^{\text{post}}\}_{i=1}^{n}$ denote outputs before and after unlearning for the same prompt set, and let IR_{pre} and IR_{post} be the corresponding flagged fractions under a binary detector $d(\cdot) \in \{0,1\}$. The improvement is summarized by the drop $\Delta IR = IR_{\text{pre}} - IR_{\text{post}}$. Several works also estimate harm with an LLM-based judge (optionally via image captions) and aggregate scores by thresholding or averaging (Wang et al., 2024).

VLM-Based Judgments. Pretrained VLMs can serve as external judges for presence of a forbidden concept. Let a VQA-style judge output a binary decision $g(y) \in \{0,1\}$ for concept presence, or a matching score $s(y,t) \in [0,1]$ for image y and concept text t. Define the yes-rate drop and similarity drop as

$$\Delta_{\text{VQA}} = \frac{1}{n}\sum_{i=1}^{n} g(y_i^{\text{pre}}) - \frac{1}{n}\sum_{i=1}^{n} g(y_i^{\text{post}}),$$

$$\Delta_s = \frac{1}{n}\sum_{i=1}^{n} s(y_i^{\text{pre}}, t) - \frac{1}{n}\sum_{i=1}^{n} s(y_i^{\text{post}}, t).$$

VLMs used for g or s include VQA heads such as CLIP-FlanT5-based VQAScore and ITM scores from BLIP-2; these are standard tools for judging whether generated content still expresses the concept (Lin et al., 2024). Larger Δ_{VQA} or Δ_s indicates more effective forgetting.

C.3 Attack-Based Privacy

Membership Inference Attack and Enhanced Variants. Membership Inference Attacks (MIA) are a standard privacy test for evaluating whether an unlearned model still leaks information about forgotten data. MIA estimates how easily an adversary can infer whether a sample was part of the original training set. For a forget set D_f, following established formulations (Shokri et al., 2017; Carlini et al., 2022; Jia et al., 2023; Wang et al., 2025d), MIA efficacy is defined as

$$\text{MIA} = \frac{1}{|D_f|}\sum_{x_i \in D_f} \mathbf{1}\big[\,A(F_T, x_i) \in \{0,1\}\,\big],$$

where F_T denotes the evaluated target model and A the membership inference attacker, which predicts membership as 1 if $x_i \in D_{\text{train}}$ and 0 otherwise. Higher MIA efficacy indicates that the unlearned model behaves closer to a model retrained without the forgotten data. Beyond the basic setting, prior work proposes enhanced MIA variants that audit specific components or compare unlearned models against retrained references, providing stronger privacy guarantees (Dontsov et al., 2024; Wang et al., 2025b; Koudounas et al., 2025).

Identity Matching. Identity leakage metrics assess whether model outputs still reveal a forgotten identity after unlearning. In vision settings, evaluation typically relies on recognition accuracy or embedding similarity between generated outputs and reference images. Forgetting is considered successful when recognition accuracy for the erased identity drops to chance level and embedding similarity exhibits a substantial decline (Biswas et al., 2025; Cai et al., 2025; Nagasubramaniam et al., 2025). Common embedding-based measures include Identity Matching Score (IMS) (Liu et al., 2024e) and Identity Score Matching (ISM) (Wu et al., 2025b). In text and multimodal evaluations, identity leakage is monitored through identity mentions in generated captions or VQA responses, where effective erasure drives correct mention rates toward zero (Dontsov et al., 2024; Ma et al., 2024b).

Voice Privacy. In speech unlearning, privacy evaluation assesses whether a model can still recognize or reproduce a forgotten speaker after unlearning. A common signal is speaker similarity (SIM), which measures the alignment between embeddings of generated and reference utterances; effective unlearning reduces SIM for forgotten speakers while preserving similarity for retained ones (Chen et al., 2022).

Complementary to similarity, speaker Zero-Retrain Forgetting (spk-ZRF) (Kim et al., 2025b) evaluates whether speaker identity becomes uncorrelated with prompting after unlearning. It computes the Jensen-Shannon divergence between speaker identity distributions obtained with and without speaker prompts,

$$JSD_i = \frac{1}{2}\left[D_{\mathrm{KL}}(p_i \,\|\, m_i) + D_{\mathrm{KL}}(q_i \,\|\, m_i)\right]$$

$$\text{spk-ZRF} = 1 - \frac{1}{n_f}\sum_{i=1}^{n_f} JSD_i,$$

where higher spk-ZRF values indicate that generated speech no longer preserves the forgotten speaker identity.

C.4 Model Utility and Faithfulness

Classification Accuracy. Retained utility on non-forgotten data is commonly measured by Top-k classification accuracy on remaining classes (Bansal et al., 2023; Struppek et al., 2024; Han et al., 2025a; Biswas et al., 2025):

$$\text{Top-k Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[\, y_i \in \text{Top-k}\big(\hat{\mathbf{p}}_i\big) \,\big]\,,$$

where y_i denotes the ground-truth label and $\hat{\mathbf{p}}_i$ the predicted class scores.

Cross-Modal Retrieval Utility. For multimodal models, utility retention is evaluated using retrieval metrics such as Recall@K and R-Precision on held-out benchmarks (Yang et al., 2024a; Sinha et al., 2025):

$$\text{Recall@}K = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[\, R(q_i) \cap \text{TopK}(q_i) \neq \varnothing \,\big]\,.$$

Language and QA Metrics. Retained capability on non-forgotten data is tracked with standard NLP scores. For VLMs that perform question answering or caption generation, language quality on non-forgotten examples is assessed with BLEU (Zhang et al., 2025b), ROUGE-L (Dontsov et al., 2024), and METEOR (Liu et al., 2024a). Stable BLEU/ROUGE-L/METEOR on unrelated VQA or captioning items indicates preserved language utility. In addition, CLIP Score (Hessel et al., 2021) is widely used to assess image-text alignment, with consistent scores on non-target prompts suggesting that multimodal semantic alignment remains intact following unlearning (Yang et al., 2024a; Cheng and Amiri, 2024b).

Generative Output Quality. To ensure image generation quality is retained, vision metrics like Fréchet Inception Distance (FID) (Heusel et al., 2017), Fréchet Video Distance (FVD) (Unterthiner et al., 2019; Facchiano et al., 2025), Kernel Inception Distance (KID) (Bińkowski et al., 2018) and inverted FID (IFID) (Li et al., 2024c) are commonly reported. These metrics compare the distribution of generated images to that of real images using feature statistics. FID computes the distance between the means (μ) and covariances (Σ) of Inception features for generated (g) and real (r) samples:

$$\begin{aligned}\text{FID}(r,g) &= \|\mu_r - \mu_g\|_2^2 \\ &\quad + \text{Tr}\big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\big).\end{aligned}$$

Lower FID and stable KID values on retain-set prompts indicate that unlearning preserves fidelity and diversity of generated images (Fan et al., 2023; Zhang et al., 2024d; Chen et al., 2025c).

Beyond distributional similarity, perceptual and faithfulness metrics provide complementary signals. PickScore (Kirstain et al., 2023) and Aesthetic Score (AES) (Schuhmann et al., 2022) evaluate semantic alignment and visual appeal, while Polling-based Object Probing Evaluation

(POPE) (Li et al., 2023) measures object hallucination in VLM outputs; stable scores suggest that unlearning does not degrade perceptual quality or semantic correctness (Ma et al., 2024b; Li et al., 2024b).

Perceptual Similarity. Perceptual similarity metrics assess whether unlearning alters model outputs on benign inputs by comparing generations from the unlearned model to those of the original model. The Learned Perceptual Image Patch Similarity (LPIPS) score (Zhang et al., 2018) measures perceptual distance between two images in a deep feature space. Lower LPIPS values on retain prompts indicate higher integrity, meaning that outputs remain perceptually close on non-target inputs after unlearning. Mean LPIPS on benign prompts is therefore commonly reported to verify that unlearning preserves visual details, style, and overall generation quality (Dai and Gifford, 2023; Park et al., 2024; Biswas et al., 2025).

LLM-as-Judge Evaluation. Several multimodal unlearning studies use large language models as semantic evaluators to score model outputs. These approaches prompt an LLM with task-specific rubrics and interpret its responses as scores for safety, factuality, or answer quality. Recent multimodal benchmarks adopt GPT-Eval-style setups to rate generated outputs along these semantic dimensions (Ma et al., 2024b; Park et al., 2024; Xu et al., 2025b; Liu et al., 2025c). Such evaluations provide a semantics-aware assessment of unlearning behavior that complements surface-level automatic metrics.

Human-Centered Evaluation. While most unlearning work relies on automatic metrics, several multimodal studies incorporate human judgment to assess perceived safety and fidelity. In safety-oriented evaluations, annotators label model outputs from different training or unlearning conditions for harmfulness, and aggregated judgments with high inter-annotator agreement reveal changes in harmful output rates after unlearning (Chakraborty et al., 2024). In diffusion unlearning, human studies compare generated images against reference subjects to assess whether unlearning suppresses identity- or style-specific resemblance while preserving benign generations (Huang et al., 2024b). These evaluations provide complementary evidence that unlearning reduces harmful or identifiable content beyond what automated metrics capture.

C.5 Adversarial Perturbation Robustness

Attack Success Rate (ASR) quantifies how often adversarially perturbed inputs still elicit forbidden content from an unlearned model. Let D be the evaluation set of harmful text-image pairs and $R_i = f(x_i^{\text{adv}})$ the response to the i-th adversarial input; a response is unsafe if it contains forbidden content. The ASR is defined as

$$\text{ASR} = \frac{1}{|D|} \sum_{i=1}^{|D|} I_A(R_i)\,,$$

where $I_A(\cdot)$ is an indicator that returns 1 when the response contains harmful knowledge and 0 otherwise (Chen et al., 2025a). A higher ASR indicates that forgotten content remains vulnerable to adversarial reactivation, suggesting incomplete unlearning. Prior work reports ASR under both white-box and black-box attack settings to assess robustness of unlearning against adaptive adversaries (Bansal et al., 2023; Zhang et al., 2024b; Biswas et al., 2025).

C.6 Compute and Environmental Budget

Run-time and Memory Usage. Compute footprint anchors the edit budget for unlearning methods. Studies now report wall-clock runtime (often denoted WCT) and peak memory as first-class metrics under Run-Time Efficiency (RTE, typically measured in minutes), alongside peak GPU memory consumption (in GB), to certify that forgetting is practical at scale. Beyond elapsed time, some work also quantifies training cost using total floating-point operations (TFLOPs) and effective throughput (TFLOPS), and characterises inference cost via a relative complexity ratio with respect to a backbone model (Zhang et al., 2024c). Across image classification, diffusion, and contrastive settings, recent work consistently reports WCT, memory usage, and FLOP-based measures, showing modest additional compute compared to full retraining and making unlearning overheads comparable across architectures and hardware platforms (Fan et al., 2023; Li et al., 2025d; Dang et al., 2025a; Cywiński and Deja, 2025; Wang et al., 2025b; Spartalis et al., 2025).

Environmental Cost. Beyond accuracy and robustness, multimodal unlearning also introduces an environmental cost. Recent work estimates emissions by logging GPU energy in kilowatt-hours and multiplying by an assumed grid carbon intensity of about 0.4 kgCO$_2$e per kWh (Dodge et al., 2022;

Chakraborty et al., 2024). These measurements show that multimodal unlearning consumes substantially more energy than text-only unlearning on the same GPU, so reporting energy use and derived CO$_2$e for each setting helps evaluations of unlearning account for environmental impact alongside safety and privacy.

D Unlearning Robustness

Adversarial Reactivation Attacks. Adversarial reactivation attacks evaluate unlearning robustness by optimizing prompts or guidance that recover a forgotten concept without modifying model weights. These attacks exploit residual conditioning, safety, or cross-modal pathways and operate at decoding or prompting time using gradient-based, surrogate, or zeroth-order search (Kim et al., 2024b; Dang et al., 2025a; Zhang et al., 2025b).

$$\begin{aligned} &\max_{p,z}\ S_\theta(p,z;c)\ -\ \lambda_1\,\Delta(p,p_0)\ -\ \lambda_2\,R(z)\\ &\text{s.t.}\quad \text{queries} \le Q_{\max},\quad C(p)\in\mathcal{B}. \end{aligned}$$

Here θ denotes fixed model parameters; p is a discrete prompt and z an optional conditioning latent or embedding; $S_\theta(p,z;c)$ scores concept c (for example CLIP similarity, an NSFW detector logit, or a task success score); Δ bounds prompt edits from a seed p_0; R regularizes latents; $\mathcal{B}$ enforces benign surface form and $Q_{\max}$ limits black-box queries. Transfer terms or surrogate models can be included by adding $\alpha\,\mathbb{E}_\phi\,S_\phi(p,z;c)$ to encourage cross-model success (Han et al., 2024; Liu et al., 2025a).

Methods differ in how they optimize this objective. AutoJailbreaking (Kim et al., 2024b), which performs LLM-driven prompt search to evade filters and reveal residual unsafe behavior; Diff-ZOO (Dang et al., 2025a), which uses query efficient zeroth order ascent in the discrete token space to elicit the target under strict black box budgets; and Stealthy MLLM (Zhang et al., 2025b), which designs distribution shifted or dual purpose prompts that pass standard checks yet recover forgotten answers, exposing evaluation blind spots.

Inference-time Defenses. Inference-time defenses mitigate residual failures after unlearning by intervening during sampling rather than modifying parameters. They operate on the conditioning stream to suppress adversarial signals while preserving responses to benign prompts, commonly through subspace projection of adversarial token directions or adaptive smoothing of token activations (Chen et al., 2025b; Han et al., 2025b).

$$s_t^{\text{def}}(x_t,E) = s_\theta\big(x_t \mid S_t(\Pi_\perp E)\big), \Pi_\perp = I - UU^\top,$$

Here x_t denotes the latent at timestep t, E the matrix of text token embeddings, and s_θ the conditional score function. The matrix U spans an estimated adversarial subspace, and $\Pi_\perp$ projects embeddings orthogonally to that subspace. The operator S_t applies token-wise smoothing, such as median filtering, before scoring. Setting S_t to the identity recovers pure projection, while setting U to zero recovers adaptive smoothing.

E Unlearning-Adjacent Controls

Bias and Privacy Safeguards. Bias and privacy safeguards intervene on the data path. They constrain what the model sees and how prompts are encoded before any weight update, so optimization proceeds on balanced evidence with reduced attribute leakage (Alabdulmohsin et al., 2024; Huang et al., 2024a; Liu et al., 2024b).

$$\begin{aligned} \min_\theta\ &\mathbb{E}_{(x,y)\sim D}\big[w_{\text{bal}}(y)\,L(f_\theta(x),y)\big]\\ &+\lambda\,R_{\text{priv}}(f_\theta;A,g), \end{aligned}$$

where f_θ is the model, L the task loss, $w_{\text{bal}}(y)$ denotes class- or attribute-level reweighting for bias control, A indexes sensitive attributes, g is a privacy editing operator such as differentially private image sanitization, and R_{priv} penalizes residual attribute leakage.

In-Context Mitigation. In-context mitigation steers a frozen VLM at prompt time by inserting a small set of curated multimodal demonstrations and summaries, so that decoding conditions on safer evidence rather than on harmful patterns (Zhou et al., 2024b). Because it operates entirely through the input channel, it avoids retraining and remains reversible, but its effectiveness depends on demonstration quality, retrieval coverage, and the available context budget.

F Comprehensive Application Scenarios

Privacy and Regulatory Compliance. Unlearning for privacy and regulatory compliance addresses deletion requests, right-to-be-forgotten (RTBF) enforcement, and license-driven removals across multimodal systems. In vision-language pipelines, unlearning is used to erase specific identities, sensitive attributes, or marked image-text pairs while

Modality	Dataset	Size	Used in
Personalization Setup			
Image	DreamBooth (Ruiz et al., 2023)	30 subjects, 4-6 images each	Liu et al., 2024e; Li et al., 2025a
Image-Text	DiffusionDB (Wang et al., 2023)	14M images, 1.8M prompts	Pan et al., 2024; Li et al., 2025a
	DreamBench++ (Peng et al., 2025)	150 images with 1,350 prompts	Li et al., 2025a
Copyright Unlearning			
Image	CPDM (Ma et al., 2024a)	2.1K anchors and 18.9K paired generated images	Moon et al., 2025; Liu et al., 2025b; Jin et al., 2025; Ren et al., 2025
	VioT (Kim et al., 2024b)	100 images total across 5 copyrighted categories	Kim et al., 2024b
Audio	MusicCaps (Agostinelli et al., 2023)	5.5K captioned clips	Kim et al., 2025a
Knowledge QA and Instruction Probes			
Image-Text	VQA (Antol et al., 2015)	255K images, 764K questions, 10M human answers	Ma et al., 2024b; Dontsov et al., 2024; Chen et al., 2025d
	VQAv2 (Goyal et al., 2017)	265K images with 1.1M questions	Li et al., 2024b; Chakraborty et al., 2024; Chen et al., 2025d
	NLVR2 (Suhr et al., 2019)	107K caption-image pairs, 29.7K unique sentences	Cheng and Amiri, 2024a,b
	ScienceQA (Lu et al., 2022)	21.2K multimodal multiple-choice science questions	Gao et al., 2024a; Chen et al., 2025d
	GQA (Hudson and Manning, 2019)	113K images with 22.7M compositional visual questions	Li et al., 2024b; Xing et al., 2024; Li et al., 2024c
	UnLOK-VQA (Patil et al., 2024)	500 visual QA samples (OK-VQA (Marino et al., 2019) extension)	Patil et al., 2024; Wu et al., 2025a
	VizWiz (Gurari et al., 2018)	31K real-world visual questions from blind users	Li et al., 2024b; Chen et al., 2025d,a
	POPE (Li et al., 2023)	18K object-image queries for VLM hallucination evaluation	Xing et al., 2024; Li et al., 2024b; Ma et al., 2024b; Xu et al., 2025b
Text	PGR (Sousa et al., 2019)	1.7K PubMed abstracts annotated with 4.2K phenotype-gene relations	Cheng and Amiri, 2024b
Segmentation and I2I Unlearning			
Image	MS-COCO (Lin et al., 2014)	2.5M labeled instances in 328K images (80 classes)	Park et al., 2024; Xing et al., 2024; Cywiński and Deja, 2025; Polowczyk et al., 2025
	UnlearnCanvas (Zhang et al., 2024d)	60 artistic styles across 20 object categories	Cai et al., 2025; Zhang et al., 2025a; Cywiński and Deja, 2025
Recommender Unlearning			
Image-Text	Amazon Reviews (Hou et al., 2024)	571.5M reviews from 54.5 M users on 48.2 M items across 33 categories	Sinha et al., 2025
Text-Graph	Amazon Products (Hou et al., 2024)	9.3M items, 144M reviews, 237M relational edges	Dang et al., 2025b
Text-Metadata	Yelp (Yelp Inc., 2023)	6.9M reviews, 150K businesses, with user, check-in, tip, and photo data	Dang et al., 2025b

Table 4: Datasets are grouped by unlearning setting (Personalization Setup; Copyright Unlearning; Knowledge QA and Instruction Probes; Segmentation and I2I Unlearning; Recommender Unlearning) and modality, with their sizes and representative studies.

Modality	Dataset	Size	Used in
Speech Unlearning			
Audio	Speech Commands (Warden, 2018)	64.7K v1 (30words, 1.9K speakers) /105.8K v2 (35words, 2.6K speakers) utterances	Cheng and Amiri, 2024a, 2025; Pathak et al., 2025
	AudioMNIST (Becker et al., 2024)	30K spoken-digit (0–9) audio samples from 60 speakers (9.5 hours total)	Pathak et al., 2025; Mason-Williams et al., 2025
Audio-Text	LibriSpeech (Panayotov et al., 2015)	1,000 h read English speech from 2.5K speakers, with transcripts	Kim et al., 2025b; Pathak et al., 2025; Liu, 2025
	ITALIC (Koudounas et al., 2023)	16.5K Italian intent audio samples (15.5 h), 70 speakers, 18 domains, 60 intents	Koudounas et al., 2025
Safety Robustness Unlearning			
Image-Text	I2P (Schramowski et al., 2023)	4.7K text-to-image prompts for inappropriate-content evaluation	Fan et al., 2023; Park et al., 2024; Wu and Harandi, 2025; Moon et al., 2025; Ko et al., 2024; Cywiński and Deja, 2025; Li et al., 2025d,b,c
	SneakyPrompt / NSFW_200 (Yang et al., 2024b)	200 NSFW prompts and 100 dog/cat scenario prompts	Li et al., 2024d; Wang et al., 2024; Park et al., 2024; Zhang et al., 2024b
	NudeNet (Bedapudi, 2019)	160K training images (auto-labeled) for nudity detection (>700K web-scraped images)	Poppi et al., 2024; Han et al., 2024; Shirkavand et al., 2025; Dang et al., 2025a; Chen et al., 2025b
	MIS (Ding et al., 2025)	6.2K multi-image safety samples	Chen et al., 2025d; Hu et al., 2025
	FigStep (Gong et al., 2025)	500 harmful questions over 10 safety topics	Chakraborty et al., 2024; Chen et al., 2025d; Zhang et al., 2025c; Chen et al., 2025a
Video-Text	SafeSora (Dai et al., 2024)	14.7K prompts, 57.3K videos, 51.7K human safety annotations	Yoon et al., 2025; Xu et al., 2025a

Table 5: Datasets are grouped by unlearning setting (Speech Unlearning; Safety Robustness Unlearning) and modality, with their sizes and representative studies.

preserving general utility. Representative studies focus on identity- and pair-level deletion, supported by auditing datasets and evaluations that verify the suppression of sensitive answers or visual traits (Cheng and Amiri, 2024b; Dontsov et al., 2024; Ma et al., 2024b). In generative settings, diffusion-based work further formalizes compliant data removal within image generation pipelines (Li et al., 2024a).

This application setting also includes consent-oriented and preventive controls that regulate how personal data enters training pipelines. Data-side protection mechanisms, such as unlearnable examples, introduce perturbations that prevent models from learning from protected samples, allowing individuals to share images or image-text pairs that resist downstream training while leaving unprotected data usable (Zhang et al., 2023). Related ideas extend to structured perception tasks, providing model-agnostic protection across training pipelines (Sun et al., 2024). Interactive privacy frameworks further integrate these capabilities by enabling contributors to control reuse of personal identities or styles and to request redaction or deletion through user-applied perturbations coupled with generative models and unlearning (Liu et al., 2024e). Beyond vision, privacy-driven unlearning extends to speech and audio systems, where it supports speaker opt-out and private utterance deletion to meet RTBF-style requirements. Prior

Modality	Dataset	Size	Used in
Class Unlearning			
Image	ImageNet (Deng et al., 2009)	3.2M images across 5.2K categories (synsets)	Zhang et al., 2023; Fan et al., 2023; Han et al., 2025a; Cai et al., 2025
	CIFAR (Krizhevsky, 2009)	60K images; 10 classes (CIFAR-10) or 100 classes (CIFAR-100)	Fan et al., 2023; Kim et al., 2024a; Ko et al., 2024; Sendera et al., 2025
	MNIST (LeCun et al., 2002)	70K grayscale handwritten digit images	Zhou et al., 2024b; Alberti et al., 2025
	SVHN (Netzer et al., 2011)	600K digit images from Street View (10 classes)	Fan et al., 2023; Kim et al., 2024a; Wu and Harandi, 2025
	Imagenette (Howard, 2019)	13K images across 10 ImageNet classes	Fan et al., 2023; Bui et al., 2025; Wu and Harandi, 2025; Biswas et al., 2025
	Stanford Cars (Krause et al., 2013)	16K images of 196 car classes	Zhang et al., 2023; Alabdulmohsin et al., 2024
	Stanford Dogs (Khosla et al., 2011)	20K images of 120 dog breeds	Kravets and Namboodiri, 2025a,c
	Food-101 (Bossard et al., 2014)	101K food images across 101 cuisine classes	Zhang et al., 2023; Liu et al., 2024c; Han et al., 2025a
	DTD (Cimpoi et al., 2014)	5.6K texture images covering 47 describable categories	Ilharco et al., 2023; Alabdulmohsin et al., 2024
	SUN397 (Xiao et al., 2016)	108.7K images, 397 scene classes	Zhang et al., 2023; Kim et al., 2024a; Han et al., 2025a
	WikiArt (Saleh and Elgammal, 2015)	81K artwork images across 27 styles and 45 genres	Ma et al., 2024a; Biggs et al., 2024; Chen et al., 2025b

Table 6: Datasets are grouped by unlearning setting (Class Unlearning) and modality, with their sizes and representative studies.

work demonstrates speaker-level forgetting and compliance-oriented evaluation in speech generation and recognition frameworks (Kim et al., 2025b; Cheng and Amiri, 2025).

Safety-Aligned Generation. Safety-aligned generation applies unlearning to remove NSFW, harmful, or toxic content while preserving benign behavior across modalities. In LLMs and VLMs, unlearning functions as a targeted safety control that suppresses unsafe behaviors without degrading general question answering or captioning performance (Chakraborty et al., 2024; Chen et al., 2025a). For VLMs, removing unsafe associations from cross-modal encoders yields safer retrieval and generation behavior under downstream use (Poppi et al., 2024).

In generative models, diffusion-based unlearning suppresses harmful visual concepts while maintaining output diversity and quality (Fan et al., 2023; Li et al., 2024d). Similar safety-oriented edits extend to video and motion generation, where unlearning reduces unsafe or restricted content while preserving temporal coherence and realism (Liu and Tan, 2024; De Matteis et al., 2025).

Copyright and Style Governance. Copyright and style governance in generative models leverages unlearning to remove protected styles or copyrighted content and to evaluate the completeness of such removal. In text-image diffusion, concept-level editing supports takedown of protected styles or instances, while benchmark datasets and standardized metrics assess whether copyrighted or identity-linked content has been effectively erased under copyright-sensitive deployments (Kumari et al., 2023; Ma et al., 2024a; Biswas et al., 2025). Beyond still images, unlearning extends to other generative modalities. Prior work explores opt-out unlearning in music generation and applies concept-level removal in text-to-video diffusion to suppress copyrighted or IP-restricted content while preserving general generation quality (Kim et al., 2025a; Liu and Tan, 2024).

Fairness and Reliability in Deployed Models. Fairness and reliability considerations motivate unlearning in deployed multimodal systems to mitigate biased, noisy, or unstable associations while preserving general capability. Fairness-oriented work leverages targeted forgetting to reduce skewed or culturally imbalanced associations in VLMs (Struppek et al., 2024; Zhang et al., 2024a). Reliability-focused studies examine post-unlearning stability, ensuring that model behavior remains consistent after deletions and that forgotten content does not resurface during downstream use (Schioppa et al., 2024; Gao et al., 2024b). These considerations extend across modalities, including speech and audio systems, where unlearning supports reliable operation after removal of outdated or sensitive data (Cheng and Amiri, 2025).

Personalization and Preference Control. Personalization and preference control study how multimodal systems revise or remove user-specific styles, identities, or preferences without retraining core models. In recommendation settings, preference-level unlearning updates user histories or removes modality-specific interactions under legal or licensing constraints while preserving recommendation quality (Sinha et al., 2025). VLMs further support lightweight preference control through in-context mechanisms that steer visual behavior at inference time without permanently altering general capabilities (Zhou et al., 2024b). In text-to-image diffusion, unlearning enables users to suppress unwanted styles or concepts and to prevent reproduction of personalized attributes while maintaining generation fidelity (Biggs et al., 2024; Li et al., 2024c; Polowczyk et al., 2025).

Supply-Chain and Backdoor Security. Supply-chain and backdoor security applications use unlearning to remove malicious associations introduced through poisoned data, hidden triggers, or unsafe fine-tuning, ensuring that released multimodal encoders and generators remain trustworthy in downstream use. In contrastive VLMs, unlearning mitigates poisoning and backdoor threats by weakening or removing learned trigger associations in CLIP-style encoders, improving robustness against malicious training artifacts (Bansal et al., 2023; Liang et al., 2024a,b).

In diffusion models, unlearning addresses supply-chain risks arising from prompt triggers, spatial patterns, and personalization-based attacks by selectively erasing adversarial concepts or trigger pathways while preserving generation quality (Liu et al., 2024e; Aravindan et al., 2025; Jha et al., 2025). Across modalities, robustness-oriented unlearning aims to prevent the reactivation of malicious behavior after deployment or downstream fine-tuning, supporting safer reuse of pretrained models in open ecosystems (Han et al., 2025b; Li et al., 2025a).

G Open Challenges and Future Directions

G.1 Open Challenges

Theoretical Guarantees. Despite rapid progress, most multimodal unlearning methods remain heuristic and lack formal guarantees of certified deletion, privacy, or legal compliance. In contrastive and vision-language settings, pair-level removal, single-instance deletion, and secure training procedures approximate forgetting but do not provably eliminate the influence of removed data (Cheng and Amiri, 2024b; Li et al., 2024b; Liu et al., 2024b; Wang et al., 2025b). More broadly, current generative and representation models lack metrics and benchmarks that certify retraining-equivalent removal (Zhou et al., 2024a).

In diffusion and other generative models, unlearning typically suppresses target concepts without proving erasure, and forgotten content may resurface under downstream fine-tuning or prompt variation (Kim et al., 2023; Park et al., 2024; Zhang et al., 2024a; Suriyakumar et al., 2024). Attribution and influence estimation tools provide useful diagnostics but offer only approximate evidence rather than certifiable provenance or deletion guarantees (Dai and Gifford, 2023). Establishing theoretical foundations and verifiable criteria for multimodal unlearning remains an open challenge.

Cross-Modal Generalization. Many unlearning studies evaluate on narrow model families, datasets, or modalities, which limits conclusions about general multimodal foundation models. In vision-language encoders and Multimodal Large Language Models (MLLMs), evaluations often center on a small set of architectures or controlled setups, such as CLIP- or LLaVA-only case studies, constraining transfer to broader model ecosystems (Li et al., 2024b; Dontsov et al., 2024). Benchmark analyses further show that unlearning performance is highly sensitive to architectural choices, dataset design, and evaluation tasks (Cheng and Amiri, 2024a; Liu et al., 2025c).

A similar pattern appears in generative settings, where unlearning is frequently tested on a single diffusion backbone or a limited set of concepts, making it unclear whether findings generalize across architectures, resolutions, or domains (Moon et al., 2025; Li et al., 2025c). Beyond vision, evaluations in audio, speech, and music typically focus on one model family or dataset, leaving open questions about robustness under multilingual, cross-accent, or cross-genre conditions (Kim et al., 2025b; Koudounas et al., 2025). Establishing evaluation protocols that span architectures, modalities, and realistic deployment settings remains an open challenge.

Evaluation Reliability. Evaluation reliability remains a major challenge, as many multimodal unlearning studies rely on proxy-based signals, narrow experimental setups, and unstable metrics, which limits confidence in reported gains across modalities. In VLMs and generative models, success is often assessed using automatic judges, detector outputs, or similarity thresholds on small or synthetic benchmarks, making outcomes highly sensitive to evaluation design rather than underlying model change (Poppi et al., 2024; Xing et al., 2024; Dai and Gifford, 2023).

These issues extend to safety, copyright, and privacy settings, where detector-driven or stylized benchmarks can introduce bias and fail to capture whether forgotten concepts are truly removed or merely concealed. As a result, unlearning effectiveness is frequently inferred indirectly, and conclusions may not generalize beyond the specific proxies or model configurations used (Moon et al., 2025; Zhang et al., 2024d).

Adversarial Robustness. Unlearning attempts to erase harmful behavior; however, adversarial robustness remains limited, as backdoors, jailbreaks, and other attack vectors can bring back or bypass forgotten content. In multimodal contrastive learning, existing backdoor and data-protection methods often fail under adaptive threat models, indicating that erased associations may persist in latent representations (Bansal et al., 2023; Zhang et al., 2023; Liu et al., 2024d; Liang et al., 2024b). Diffusion-based text-to-image models exhibit similar fragility: safety-driven unlearning can be bypassed by red-teaming prompts or downstream finetuning, and subject or Not Safe For Work (NSFW) suppression may either miss indirect cues or degrade benign generation when detectors are biased (Kumari et al., 2023; Park et al., 2024; Liu et al., 2024e; Chen et al., 2025c).

Black-box and transfer-based attacks further reveal residual traces of supposedly forgotten concepts, suggesting that many unlearning methods attenuate surface behavior rather than fully removing underlying representations (Han et al., 2024; Dang et al., 2025a). Overall, current defenses trade off safety and utility but remain vulnerable to adaptive reuse, highlighting the need for robustness guarantees that extend beyond static threat assumptions (Huang et al., 2024b; Yoon et al., 2025; Han et al., 2025b; Li et al., 2025a).

Utility Trade-offs. Unlearning often improves safety or compliance at the cost of utility on retained data, neighboring concepts, or benign inputs. In encoder-based models and VLMs, approaches such as CLIP hardening, pair-level deletion, and fine-grained unlearning reduce clean accuracy and cross-dataset transfer, while successful deletion does not guarantee preservation of non-target associations (Bansal et al., 2023; Cheng and Amiri, 2024b; Li et al., 2024b). Multitask evaluations further indicate that even small deletion ratios can induce measurable performance degradation across modalities (Cheng and Amiri, 2024a).

In generative models, this trade-off becomes more visible. Stronger forgetting often distorts related styles or reduces visual fidelity, while safety-oriented controls risk over-suppressing benign content or degrading unrelated generations (Kumari et al., 2023; Liu et al., 2024e; Han et al., 2025b). These effects reveal a fragile balance between deletion efficacy and utility preservation.

Beyond output quality, unlearning also incurs nontrivial computational cost, which further constrains practical deployment. Many methods require retraining large backbones, maintaining multiple checkpoints, or relying on auxiliary modules and repeated sampling, increasing both compute and storage overhead (Kim et al., 2024a; Dai and Gifford, 2023; Biggs et al., 2024). Inference-time controls introduce additional latency through extra activations or multiple denoising passes (Cywiński and Deja, 2025; Polowczyk et al., 2025).

Unified Benchmarks. Multimodal unlearning still lacks unified benchmarks, as existing evaluations are fragmented, synthetic, or tightly coupled to specific model families. Current suites for VLMs, MLLMs, and speech systems often evaluate a limited set of architectures using synthetic identities, static images, or retrained gold references, making results highly sensitive to model choice, dataset construction, and deletion order (Cheng and Amiri, 2024a; Ma et al., 2024b; Xu et al., 2025b; Liu et al., 2025c; Koudounas et al., 2025).

For generative diffusion models, benchmarks typically center on selected concept families or

Stable Diffusion-based setups and rely on proxy metrics such as CLIP or Inception scores, which complicates comparison across architectures and limits cross-method reproducibility (Zhang et al., 2024d; Moon et al., 2025; Sharma et al., 2024). Higher-level analyses emphasize the absence of shared metrics and cross-modal testbeds, which remains a major obstacle for systematic comparison and regulatory alignment (Zhou et al., 2024a).

G.2 Future Directions

Temporal and Dynamic Modalities. Extending unlearning beyond static image-text pairs to temporal multimodal signals remains an open challenge. Existing work in audio and multimodal unlearning highlights the need to handle audio-vision coupling and speaker biometrics, raising unresolved questions around streaming, continual deletion, and deployment-time guarantees (Liu et al., 2024d; Pathak et al., 2025). Parallel efforts in video and motion generation adapt unlearning to dynamic behaviors, including safety filtering and motion-aware personalization, but current methods remain limited in scope and evaluation (Liu and Tan, 2024; De Matteis et al., 2025).

Frontier-Scale Model Unlearning. Scaling unlearning methods and their evaluation to foundation-scale models remains an open challenge across modalities. Most existing studies operate on limited backbones, narrow concept scopes, or single-base architectures, which constrains conclusions about generalization to large multimodal foundation models (Dontsov et al., 2024; Patil et al., 2024; Cheng and Amiri, 2024a).

Sequential and Continual Unlearning. Practical deployments require unlearning methods that remain effective under repeated deletions, downstream fine-tuning, and long update sequences. Recent work in multimodal LLMs highlights that performance and forgetting behavior can drift as deletions accumulate, motivating continual rather than one-shot unlearning protocols (Kawakami et al., 2025). In generative diffusion models, studies show that forgotten concepts may resurface after subsequent training, prompting methods that aim to preserve deletion effects across sequential updates (Suriyakumar et al., 2024; Li et al., 2025a,c). Designing unlearning mechanisms that remain stable under long-horizon updates therefore remains an open challenge.

Controllable and Fine-Grained Unlearning. Recent work increasingly targets fine-grained control over what is forgotten, shifting from coarse dataset-level deletion to data-point, attribute, and knowledge-unit unlearning in multimodal models and VLMs (Li et al., 2024b; Xing et al., 2024; Sinha et al., 2025). Parallel efforts in speech, music, and diffusion models emphasize selective suppression of identity-, style-, or trigger-related features while preserving surrounding content and overall generation quality (Cheng and Amiri, 2025; Kim et al., 2025a; Liu et al., 2024c; Park et al., 2024). Across modalities, this setting exposes shared challenges in precision, compositionality, and stability under adversarial use or downstream adaptation, highlighting the need for unlearning mechanisms that provide reliable, interpretable, and scalable control across concepts and modalities (Cywiński and Deja, 2025; Zhang et al., 2024d).

Inference-Time Unlearning. Inference-time mechanisms suppress undesired content during generation without modifying model parameters, offering reversible and deployment-friendly control. In text-to-image diffusion, guidance-path and conditioning-path controls adjust sampling trajectories or conditioning signals to steer generations away from unsafe or copyrighted concepts while keeping the base model fixed (Li et al., 2024c; Zhang et al., 2024b; Han et al., 2025b; Park et al., 2025).

Cross-Modal Leakage Mitigation. Cross-modal leakage mitigation seeks to prevent unsafe, biased, or private information from transferring between modalities and to ensure consistent behavior across unimodal and multimodal settings. Prior studies show that safety or privacy alignment achieved in text does not reliably generalize to vision, audio, or joint reasoning, which motivates the development of multimodal attacks, metrics, and evaluation benchmarks that explicitly probe cross-modal leakage pathways (Chakraborty et al., 2024; Patil et al., 2024; Kawakami et al., 2025; Liu et al., 2025c).