arXiv:2608.03791v1 [cs.AI] 4 Aug 2026

Does Forgetting Transfer Across Modalities? A Real-World Benchmark for Cross-Modal Knowledge Unlearning Evaluation

Chunlin Liu[1,*], Junnian Chen[1,*], Haitong Jiang[1], Jianyu Zhao[2], Yingsen Pang[1], Jingchen Li[1], Jiabiao He[1], Youming Lu[1], Jinhe Bi[3], Yuntao Du[4,†]

[1]Shenzhen University
[2]Hebei University of Technology
[3]Ludwig Maximilian University of Munich
[4]Shandong University
[*]Equal contribution, [†]Corresponding author

Abstract

Vision-Language Models (VLMs), like Large Language Models (LLMs), may memorize sensitive, copyrighted, or harmful knowledge from their pretraining corpora. Removing such knowledge is essential for building trustworthy AI systems. However, existing studies primarily focus on forgetting within individual modalities. Although recent work has begun to explore cross-modal consistency in unlearning, the cross-modal transfer of real-world knowledge unlearning remains insufficiently studied. To address this gap, we introduce **UNLINK-VL**, a real-world benchmark for cross-modal knowledge unlearning in VLMs. Under a post-hoc unlearning setting in which the original forget and retain corpora are unavailable, UNLINK-VL selects visually identifiable real-world entities as unlearning targets and associates them with corresponding images and one-hop and multi-hop facts derived from Wikidata. The benchmark comprises four complementary subsets that evaluate direct forgetting of target knowledge, the propagation of forgetting through relational knowledge, the preservation of related non-target knowledge, and robustness to semantically equivalent queries. We train models under text-only and multimodal unlearning settings and evaluate forgetting effectiveness and retained utility across textual, visual, and cross-modal scenarios. Extensive experiments reveal a pronounced asymmetry in cross-modal transfer: multimodal unlearning remains effective under textual evaluation, whereas text-only unlearning transfers poorly to visual and cross-modal scenarios. Meanwhile, the evaluated methods largely preserve the models' general capabilities. These findings demonstrate that relying solely on intra-modal evaluation, particularly text-only evaluation, may substantially overestimate the effectiveness of knowledge unlearning in VLMs, underscoring the need for cross-modal unlearning and evaluation.

Introduction

Vision-Language Models (VLMs), like Large Language Models (LLMs), learn from large-scale pretraining datasets and may memorize sensitive information, copyrighted content, or harmful knowledge (Carlini et al. 2021; Ju et al. 2025; Liu et al. 2025; Zheng et al. 2025). Once encoded in model parameters, such information or knowledge may later be elicited or reproduced during downstream deployment, posing privacy, copyright, and safety risks. Retraining a model from scratch after removing the target data from the training datasets is often prohibitively expensive (Bourtoule et al. 2021; Yao et al. 2024). Machine unlearning seeks to remove the influence of designated training data or knowledge from a trained model while preserving non-target knowledge and general capabilities (Ginart et al. 2019; Guo et al. 2020; Bourtoule et al. 2021).

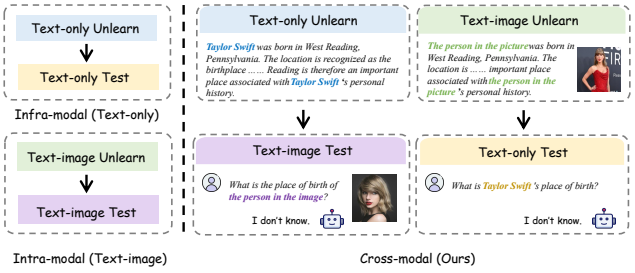

Figure 1: Comparison between intra-modal and cross-modal knowledge unlearning.

Several benchmarks have been proposed to evaluate knowledge unlearning. Text-only benchmarks evaluate the forgetting of synthetic profiles, hazardous knowledge, and real-world entities using text-only inputs for both unlearning and evaluation (Maini et al. 2024; Li et al. 2024c; Jin et al. 2024). Multimodal benchmarks extend evaluation to text-image inputs, facial identities, visual concepts, misinformation, and sensitive associations (Liu et al. 2025; Dontsov et al. 2025; Ma et al. 2025b; Li et al. 2024b; Zheng et al. 2025; Selvas-Sala, Kang, and Gomez 2026). Existing works mainly investigate intra-modal unlearning, where the model is trained to forget knowledge in one modality and evaluated in the same modality. Despite the progress enabled by these benchmarks, they largely overlook cross-modal transfer of the unlearning effect, which is a realistic and critical challenge in real-world multimodal applications.

For VLMs, robust knowledge unlearning should remain effective under both intra-modal and cross-modal evaluation. The same fact can be elicited through semantically equivalent text-only and text-image inputs. For example, a person may be specified by name in a text-only question or represented by an image in a text-image question, with both queries targeting the same biographical fact. A model may appear to have forgotten that fact when queried by name yet still recover it when shown the person's image. The unlearning effect

should therefore exhibit cross-modal consistency, regardless of how unlearning is performed.

Recent benchmarks have begun to explore whether unlearning remains effective across different modality settings. UMU-Bench targets modality misalignment by evaluating unimodal, multimodal, and hybrid unlearning under both unimodal and multimodal settings (Wang et al. 2025). PPU-Bench evaluates complete, selective, and personalized unlearning on pre-existing knowledge about real-world public figures using paired text-only QA and multimodal VQA samples (Guang et al. 2026). However, these benchmarks either focus primarily on synthetic profile knowledge or do not treat the unlearning modality as an independent experimental variable.

In this work, we study the cross-modal transfer problem: if a fact is unlearned through text-only inputs, can it still transfer to a semantically equivalent text-image question, and if it is unlearned through a text-image input, can it still transfer to an equivalent text-only question? To answer these questions, one could evaluate $T \rightarrow M$ and $M \rightarrow T$ as cross-modal transfer settings, where T and M denote text-only and text-image inputs, respectively. We also report $T \rightarrow T$ and $M \rightarrow M$ as intra-modal reference settings. This directionality matters in practice because a deletion request may identify target knowledge in one modality, whereas later users may query the same knowledge through another.

To this end, we introduce **UNLINK-VL**, a benchmark grounded in real-world knowledge for evaluating cross-modal knowledge unlearning in VLMs. UNLINK-VL selects a diverse set of visually identifiable real-world entities as unlearning targets, each associated with factual knowledge already accessible to the original model. For each entity, we first collect and filter relevant one-hop facts and two-hop paths from Wikidata (Vrandečić and Krötzsch 2014). We then verbalize this structured knowledge into text-only passages for unlearning and construct text-only QA pairs for evaluation. For both data types, we create aligned text-image counterparts by replacing the entity name with an image-based reference and pairing the resulting text with representative images.

For unlearning, UNLINK-VL constructs four forms of supervision tailored to representative unlearning methods, including GA, DPO, NPO, and RT (Jang et al. 2023; Rafailov et al. 2023; Zhang et al. 2024). Faithful positive passages are verbalized directly from the collected Wikidata facts, while counterfactual negative passages are generated by modifying factual claims in the positive passages. GA and NPO suppress the faithful positive passages, while DPO uses paired counterfactual negative and faithful positive passages. For RT, each target query is paired with a generic refusal response, such as "I don't know," to train the model to abstain from answering questions about the target knowledge.

For evaluation, UNLINK-VL organizes the structured knowledge into four complementary subsets. The Forget Set evaluates direct forgetting of target facts; the Multi-Hop Set examines whether forgetting extends through two-hop relational paths; the Retain Set evaluates the preservation of non-target facts; and the Rewrite Set assesses whether forgetting remains effective under semantically equivalent queries. Each evaluated fact is represented in aligned text-only and text-image forms. To prevent leakage, the unlearning and evaluation splits use disjoint images.

We conduct extensive experiments with four representative unlearning methods on multiple VLMs. The results reveal a pronounced asymmetry in cross-modal transfer: multimodal unlearning transfers more effectively to text-only evaluation than text-only unlearning transfers to text-image evaluation. Meanwhile, the evaluated methods largely preserve general capabilities. These findings show that text-only evaluation may substantially overestimate unlearning effectiveness in VLMs and highlight the need to consider both text-only and text-image inputs during unlearning and evaluation.

In summary, our main contributions are:

- We introduce **UNLINK-VL**, a real-world benchmark for evaluating the cross-modal transfer of unlearning effects in VLMs.
- We develop a systematic pipeline that constructs method-specific unlearning data and aligned text-only and text-image evaluation samples from pre-existing model knowledge, enabling comprehensive evaluation of forgetting, robustness, and knowledge preservation.
- Through controlled experiments that independently vary the unlearning and evaluation modalities, we reveal a pronounced asymmetry in cross-modal transfer and show that the evaluated methods largely preserve general capabilities.

Related Work

Machine Unlearning Methods

Machine unlearning aims to remove targeted data or knowledge from a trained model without retraining it from scratch (Ginart et al. 2019; Guo et al. 2020; Bourtoule et al. 2021). In large language models, factual knowledge is distributed across model parameters, making it challenging to erase targeted knowledge while preserving unrelated capabilities. Representative unlearning methods include GA (Jang et al. 2023), DPO (Rafailov et al. 2023), NPO (Zhang et al. 2024), and RMU (Li et al. 2024c). However, fine-tuning may alter how knowledge is retrieved rather than eliminate the underlying knowledge, allowing it to be recovered through paraphrased or adversarial queries (Yao et al. 2024; Bărbulescu and Triantafillou 2024; Xu et al. 2025; Hong et al. 2024; Patil, Hase, and Bansal 2024; Joshi et al. 2024). This limitation is particularly pronounced in entity-level unlearning, where target entities are embedded in interconnected networks of factual relations (Ma et al. 2025a).

Machine Unlearning Benchmarks

Existing machine-unlearning benchmarks evaluate synthetic profiles, hazardous knowledge, and the trade-off between forgetting efficacy and model utility (Maini et al. 2024; Li et al. 2024c; Shi et al. 2025). RWKU extends this line of research to a practical post-hoc setting, in which a model must forget previously acquired knowledge about public figures without access to the original forget corpora (Jin et al. 2024). A key

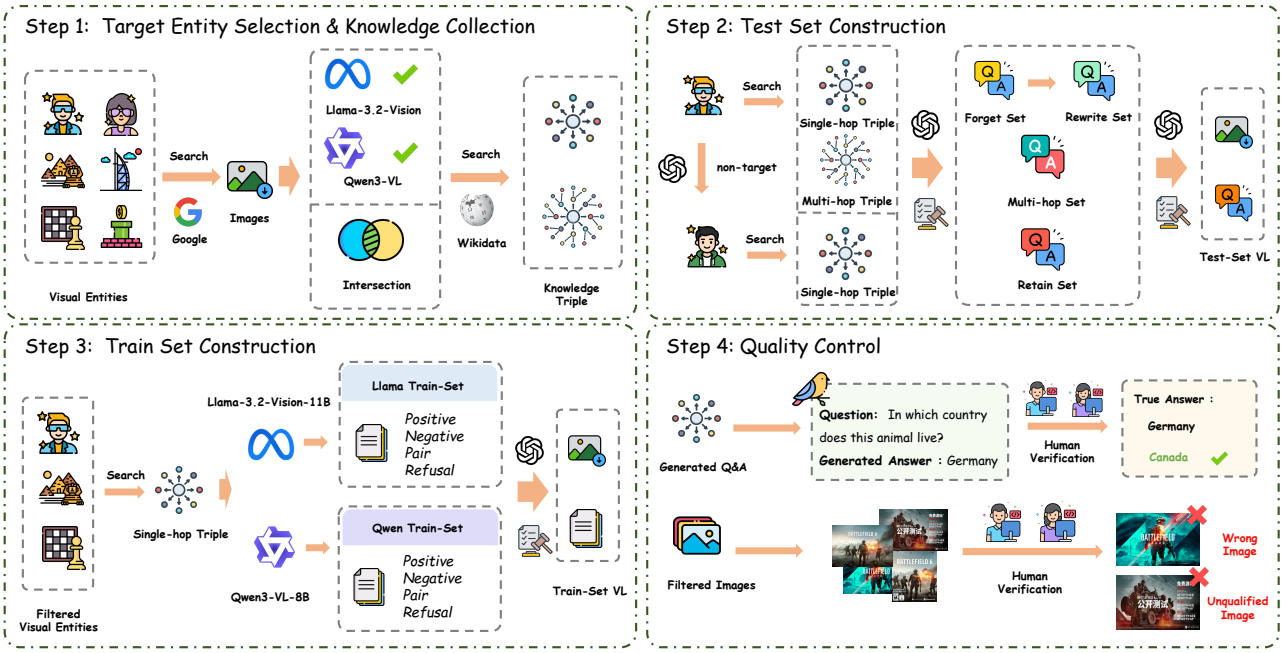

Figure 2: The four-stage construction pipeline of UNLINK-VL.

challenge in this setting is to distinguish genuine forgetting from weak prior knowledge, as a model that initially knows little about a target may appear to have been successfully unlearned (Kwon et al. 2026). Consequently, unlearning effects achieved in one modality may not transfer to the other modality (Li et al. 2026). Existing benchmarks focus on different aspects of multimodal unlearning. MLLMU-Bench evaluates the forgetting of synthetic personal knowledge in text-only and text-image settings (Liu et al. 2025). FIU-Bench studies identity-related knowledge associated with synthetic facial images and tests whether such knowledge can be recovered through privacy attacks (Ma et al. 2025b). Other benchmarks extend multimodal unlearning to visual concepts, misinformation, and sensitive associations (Li et al. 2024b; Zheng et al. 2025; Selvas-Sala, Kang, and Gomez 2026). Collectively, these studies show that unlearning effectiveness depends on what is forgotten, how it is represented, and how it is evaluated. Among existing benchmarks, UMU-Bench and PPU-Bench are most closely related to our setting. UMU-Bench evaluates unimodal, multimodal, and mixed-modality unlearning using synthetic profiles, with a focus on modality alignment after unlearning (Wang et al. 2025). By contrast, PPU-Bench focuses on knowledge that models already possess about public figures and evaluates complete, selective, and personalized unlearning through text-only and text-image QA (Guang et al. 2026). UNLINK-VL differs from both benchmarks by focusing on real-world knowledge and systematically examining whether unlearning effects induced in one modality transfer to another. Appendix D provides a detailed comparison with representative text-only and multimodal unlearning benchmarks.

UNLINK-VL Benchmark

Task Definition

Knowledge unlearning. Suppose that an already trained model M_{vanilla} is asked to forget a set of target entities E_{forget}. The goal is simple: after unlearning, the model should no longer answer questions about the target knowledge, but should still answer questions about unrelated knowledge.

Conceptually, the original training set D can be divided into the data to be forgotten, D_{forget}, and all remaining data, D_{rest}:

$$D = D_{\text{forget}} \cup D_{\text{rest}}, \qquad D_{\text{forget}} \cap D_{\text{rest}} = \varnothing, \\ M_{\text{unlearn}} = A(M_{\text{vanilla}}, D_{\text{forget}}). \tag{1}$$

Here, A is the unlearning algorithm. The ideal solution would remove D_{forget} and retrain from scratch on D_{rest}; A instead updates M_{vanilla} directly to obtain M_{unlearn} (Ginart et al. 2019; Guo et al. 2020; Bourtoule et al. 2021). In our practical post-hoc setting (Jin et al. 2024), the original D_{forget} and D_{rest} are unavailable. We therefore construct a compact D_{forget} from public facts about E_{forget}.

Cross-modal knowledge unlearning. A fact is represented as a triple $k = (E, r, o)$, where E is the subject entity, r is the relation, and o is the correct answer. For example, the fact

$$k = (\text{Taylor Swift},\ \text{place of birth},\ \text{West Reading}) \tag{2}$$

states that Taylor Swift was born in West Reading.

For each fact (E, r, o), we construct two semantically equivalent questions that differ only in how the subject entity is identified:

$$Q_{k,T} = G_T(E, r), \\ Q_{k,M} = \big(I_E, G_M(E, r)\big), \\ A_{k,T} = A_{k,M} = o. \tag{3}$$

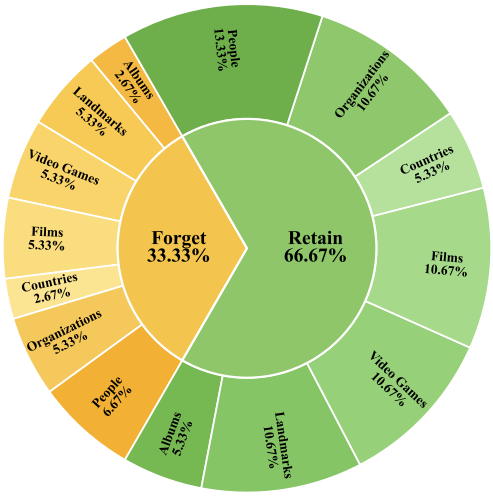

Figure 3: Entity-category distribution of the one-hop Forget and Retain evaluation sets.

Here, G_T generates a text-only question that explicitly names E, whereas G_M generates an image-grounded question that refers to E through the image I_E. Both questions have the same answer o.

Further details on the paired query construction and evaluation settings are provided in Appendix B.

Our training and evaluation settings are as follows:

$$T \to T, \qquad T \to M, \qquad M \to T, \qquad M \to M. \tag{4}$$

In $a \to b$, a and b denote the unlearning and evaluation modalities, respectively. The diagonal settings ($T \to T$ and $M \to M$) are intra-modal, whereas the off-diagonal settings ($T \to M$ and $M \to T$) evaluate cross-modal transfer.

Dataset Construction

To support reliable cross-modal evaluation while avoiding prompt and image leakage, we construct separate unlearning and evaluation branches grounded in the same underlying facts, as illustrated by the four-stage pipeline in Figure 2. Our pipeline draws on RWKU's focus on real-world target entities accessible to the evaluated models and their neighboring knowledge (Jin et al. 2024), while following MMKU-Bench's staged organization of knowledge collection, cross-modal data construction, and quality control (Fu et al. 2026). By keeping the target knowledge aligned across data splits, this design enables controlled evaluation of both intra-modal forgetting and cross-modal transfer.

Stage 1: Target entity selection and knowledge collection. To ensure that the unlearning targets are recognizable from both their names and images while avoiding image leakage between unlearning and evaluation, we begin with a broad pool of visually identifiable real-world entities spanning people, landmarks, organizations, countries, films, albums, games, and other categories. We retrieve candidate images through image search and conduct probing under both text-only and image-only settings on the Qwen3-VL and Llama-3.2-Vision model families to exclude entities unfamiliar to either family (Kwon et al. 2026). The complete probing protocol and filtering criteria are provided in Appendix B. After filtering, 250 target entities remain. For each target, we collect two representative images, both of which independently satisfy the image-only probing criteria for both model families. To distinguish entity-level forgetting from image-specific effects, we use Image2 for multimodal unlearning and reserve Image1 for evaluation.

To assess not only forgetting but also its propagation through relational chains, we link each target entity to Wikidata (Vrandečić and Krötzsch 2014) and retrieve both direct triples (E, r, o) and valid two-hop paths $(E, r_1, E'), (E', r_2, o)$ (Zhong et al. 2023). We further identify semantically related non-target entities for retention evaluation (Cohen et al. 2024). Each candidate entity and its associated image undergo the same text-only and image-only probing procedure, and only those recognized by both model families in both settings are retained. These entities constitute E_{retain}.

Stage 2: Test-set construction. In order to separately evaluate forgetting, robustness to paraphrasing, recoverability through relational composition, and preservation of non-target knowledge, we organize the collected facts into four complementary evaluation subsets. The Forget Set consists of one-hop facts and measures whether the target knowledge is successfully forgotten. The Rewrite Set consists of GPT-4o-generated paraphrases of the Forget Set questions to evaluate robustness to alternative phrasings. The Multi-Hop Set is constructed from valid two-hop paths and assesses whether the target knowledge can still be recovered through relational composition. Finally, the Retain Set contains one-hop facts about non-target entities and evaluates whether such knowledge is preserved after unlearning.

GPT-4o converts the structured facts into clear, unambiguous questions with short answers grounded in the source triples. To guarantee that input modality acts as the sole variable in cross-modal comparisons, we instantiate each evaluation item as a fact-aligned modality pair. In the text-only version, the subject entity is named explicitly; in the text-image version, the question is paired with an image, and the entity name is replaced with an image-grounded referring expression, while the queried relation and answer remain unchanged. Thus, the two versions differ only in how the subject entity is accessed.

The complete prompts used for KG-to-QA generation, semantic rewriting, and cross-modal conversion are provided in Appendix C.

Stage 3: Train-set construction. Since the pretrained datasets are unavailable, we construct model-specific synthetic training data from triples associated with the target entities retained after probing. Qwen3-VL-8B and Llama-3.2-Vision-11B each generate the forget corpus for their corresponding model family.

To enable a controlled comparison between text-only and multimodal unlearning, we construct fact-aligned text-image counterparts for all text-only unlearning instances in the same manner as the evaluation pairs. Dataset statistics are summarized in Table 1 and Figure 3.

Statistics	Number
Training set (per model family)	
Faithful samples	13,173
Counterfactual samples	13,173
Text preference pairs	13,173
Refusal samples	10,000
Test set	
Forget Set	11906
Multi-Hop Set	3398
Rewrite Set	11906
Retain Set	20351
Entity	
Forget	250
Retain	500

Table 1: Key statistics of UNLINK-VL.

Stage 4: Quality control. To prevent data-construction artifacts from being mistaken for unlearning effects and undermining the validity of the benchmark, we first apply automated screening to remove malformed, duplicate, ambiguous, or inconsistent samples, as well as unusable images. Human reviewers then correct unintended errors introduced during triple-to-QA conversion and verify the accuracy of image selection and image–entity alignment. Appendix E details the review unit, checklist, correction procedure, and final verification pass.

Unlearning Methods

We evaluate four representative methods. GA lowers the likelihood of the correct target response y^+ (Jang et al. 2023); DPO ranks the counterfactual response y^- above the faithful response y^+ relative to the original model (Rafailov et al. 2023); NPO suppresses y^+ without requiring a preferred completion (Zhang et al. 2024); and RT applies supervised fine-tuning to y^{ref}. All four methods are implemented with both full-parameter fine-tuning and LoRA. For each target fact, the training data provide four forms of supervision: a faithful positive response y^+, a fluent counterfactual negative response y^-, a preference pair (y^-, y^+), and a standardized refusal response y^{ref}. Here, y^- is the preferred completion and y^+ is the rejected completion in the preference pair. Positive and negative responses are generated and filtered separately for each target model, and their matched combination forms the paired supervision used by DPO. The same fact is instantiated across different modalities and supervision schemes, enabling controlled comparisons of GA, DPO, NPO, and RT while keeping the target knowledge fixed.

Metrics

We evaluate knowledge unlearning along two standard axes: forgetting efficacy and retained utility (Maini et al. 2024; Zhang et al. 2024). We report answer accuracy on each subset. Lower accuracy on the Forget, Multi-Hop, and Rewrite Sets indicates stronger forgetting, whereas higher accuracy on the Retain Set indicates better preservation of neighboring knowledge. Besides, general capabilities are reported using the official metric of each external benchmark.

Experiments

Experimental Setup

Models. We evaluate UNLINK-VL on Qwen3-VL-8B, Qwen3-VL-32B, and Llama-3.2-Vision-11B. The two model families use different vision–language integration mechanisms, making them suitable for testing whether cross-modal unlearning patterns generalize across architectures. The main paper reports representative Qwen3-VL-8B results and a LoRA-based 8B–32B scale comparison, while the consolidated Llama-3.2-Vision-11B results are deferred to Appendix A.

Implementation. All models are trained in `bf16` for three epochs with a maximum sequence length of 1,024, a per-device batch size of 1, and four gradient-accumulation steps. We use a cosine learning-rate schedule with 20 warm-up steps; the learning rate is selected for each objective and backbone. LoRA uses rank $r = 32$, scaling factor $\alpha = 64$, and targets all supported linear modules. Full-parameter training optionally uses DeepSpeed ZeRO-2 (Rajbhandari et al. 2020). The 8B and 11B LoRA runs use two NVIDIA A100-80GB GPUs, and their full-parameter runs use four; the 32B LoRA runs use four A100-80GB GPUs. We train separate text-only and multimodal unlearning variants for each method and evaluate them with the same standalone pipeline over the Forget, Multi-Hop, Rewrite, and Retain Sets in both modalities.

Main Results

For compact presentation, Table 2 merges the four unlearning–evaluation configurations for Qwen3-VL-8B. Each configuration contains Forget (F), Retain (R), Multi-Hop (MH), and Rewrite (RW) accuracy.

(1) Intra-modal unlearning is generally more effective with multimodal inputs. Since text-only and text-image evaluations have different Vanilla baselines, we compare baseline-normalized Forget reductions. In six of eight method–update pairs, $M \to M$ yields a larger relative reduction than $T \to T$, by 1.3–3.8 percentage points. The two LoRA exceptions are GA, which is nearly tied (20.9 % versus 21.0%), and DPO, where $T \to T$ is slightly stronger (19.5% versus 18.8%). Thus, multimodal inputs generally strengthen intra-modal suppression, though not uniformly under LoRA.

(2) Cross-modal forgetting remains incomplete and asymmetric. For every method and update regime, cross-modal evaluation yields a smaller baseline-normalized Forget reduction than the corresponding intra-modal evaluation. The relative $T \to M$ penalty ranges from 15.9 to 27.4 percentage points, whereas the $M \to T$ penalty ranges from 6.4 to 19.2 points. Moreover, the latter is smaller for all eight method–update pairs, averaging 12.2 points compared with 21.2 for $T \to M$. These normalized comparisons remove modality-specific Vanilla differences and confirm that multimodal unlearning transfers more effectively to text-only queries than text-only unlearning transfers to text-image queries, so same-modality evaluation can overestimate forgetting.

(3) NPO forgets most, DPO preserves most, and RT provides the best balance. Under the same update and

Method	$T \to M$				$M \to M$				$M \to T$				$T \to T$			
	F↓	R↑	MH↓	RW↓	F↓	R↑	MH↓	RW↓	F↓	R↑	MH↓	RW↓	F↓	R↑	MH↓	RW↓
Vanilla	82.6	80.8	66.9	79.4	82.6	80.8	66.9	79.4	85.1	82.3	69.7	82.0	85.1	82.3	69.7	82.0
GA (Full)	74.2	63.8	62.8	72.5	52.4	61.5	54.6	57.1	66.8	65.0	60.2	68.9	55.1	63.2	56.8	59.5
GA (LoRA)	79.0	72.8	65.2	76.5	65.3	70.6	60.7	67.2	73.5	73.7	64.3	74.9	67.2	71.4	62.8	68.8
DPO (Full)	75.8	72.9	63.0	73.6	56.0	70.8	55.8	59.8	68.4	73.8	60.9	70.4	58.9	71.6	57.4	61.6
DPO (LoRA)	79.6	78.7	65.0	77.0	67.1	77.9	61.2	68.6	74.6	79.2	64.1	75.4	68.5	77.1	62.0	69.2
NPO (Full)	71.0	68.9	61.8	69.4	45.2	66.1	50.7	53.8	62.9	69.8	57.8	66.3	49.8	67.4	53.1	55.6
NPO (LoRA)	75.4	75.5	63.6	73.4	58.4	73.7	56.9	61.2	69.2	76.6	61.2	71.0	62.0	73.8	58.5	63.5
RT (Full)	72.6	71.7	62.1	70.5	48.8	69.9	52.3	55.4	64.3	72.4	58.5	67.4	52.9	70.1	54.2	57.2
RT (LoRA)	76.8	77.6	64.1	74.5	60.5	76.2	58.1	63.1	70.8	78.3	61.8	72.6	63.8	75.4	59.4	65.2

Table 2: The results of Qwen3-VL-8B on UNLINK-VL. T and M denote text-only and multimodal inputs, respectively. F, R, MH, and RW denote the Forget, Retain, Multi-Hop, and Rewrite scores, respectively.

	Method	Comprehensive			OCR	Multidisciplinary	Instruction	Mathematical		Hallucination	
		MME↑	MMBench↑	SEED↑	OCRBench↑	ScienceQA↑	MIA-Bench↑	MathVista↑	MathVision↑	POPE↑	HallusionBench↑
	Vanilla	1734.77	75.43	67.24	90.9	74.27	93.83	78.9	49.01	88.44	73.50
T	GA (LoRA)	1732.25	73.80	64.78	91.2	68.52	93.45	76.2	53.29	88.05	72.87
T	DPO (LoRA)	1736.63	75.34	65.88	90.5	72.38	94.04	78.3	50.00	88.13	74.13
T	NPO (LoRA)	1737.43	75.17	65.04	90.5	70.05	94.21	77.5	51.97	88.03	73.92
T	RT (LoRA)	1729.35	76.03	67.19	91.0	73.62	93.76	77.5	52.63	87.95	72.77
M	GA (LoRA)	1724.85	75.09	66.01	91.1	68.47	94.23	78.2	48.68	88.04	73.19
M	DPO (LoRA)	1744.75	76.46	68.07	90.5	75.76	94.52	78.3	52.30	88.16	72.98
M	NPO (LoRA)	1737.52	75.00	66.10	91.6	71.39	93.97	78.1	51.32	87.97	74.34
M	RT (LoRA)	1742.37	75.60	66.93	90.9	73.03	94.30	76.5	51.32	87.93	73.71

Table 3: General-capability results for Qwen3-VL-8B after unlearning.

unlearning–evaluation configuration, NPO always achieves the lowest Forget accuracy and therefore removes the most target knowledge. DPO always achieves the highest Retain accuracy and causes the least damage to neighboring knowledge. RT forgets more target knowledge than DPO while preserving more neighboring knowledge than NPO, giving the best overall balance. GA performs less favorably: it forgets less than NPO and RT while usually obtaining the lowest Retain accuracy.

(4) Neighboring knowledge is largely preserved, but forgotten knowledge remains recoverable. Across all trained variants and modality configurations, Retain accuracy remains between 61.5 and 79.2, showing that much neighboring knowledge survives, although some collateral damage remains. However, accuracy on the Multi-Hop and Rewrite Sets remains substantial, ranging from 50.7 to 65.2 and from 53.8 to 77.0, respectively. Thus, even when a model fails to answer a direct Forget Set question, it may recover the same target knowledge through relational reasoning or a semantically equivalent question. Effective unlearning needs to preserve neighboring knowledge while preventing such recovery.

Analysis Experiments

General-Capability Evaluation. We evaluate unrelated utility on the ten benchmarks listed in Table 3 using their official protocols (Fu et al. 2023; Liu et al. 2024a; Li et al. 2024a; Liu et al. 2024b; Lu et al. 2022; Qian et al. 2025; Lu et al. 2024; Wang et al. 2024; Li et al. 2023; Guan et al. 2024).

Table 3 shows that LoRA generally preserves broad utility, although losses are task and objective dependent. MME stays within -0.57% to $+0.58\%$ of Vanilla, while OCRBench and POPE change by at most 0.7 and 0.51 points. The largest drops occur in scientific and mathematical reasoning: GA reduces ScienceQA from 74.27 to 68.52 (text) and 68.47 (multimodal), and its text variant lowers MathVista from 78.9 to 76.2. DPO with multimodal unlearning remains near or above Vanilla on several broad and instruction-following benchmarks. Neither unlearning modality is uniformly better; small gains are treated as retained capability or evaluation variation. Thus, general utility is largely preserved, but the few pronounced losses must be considered alongside forgetting and locality.

Full-Parameter Fine-Tuning versus LoRA. The matched runs reveal a clear forgetting-locality trade-off. Rel-

	$T \to M$				$M \to M$				$M \to T$				$T \to T$			
Method	F↓	R↑	MH↓	RW↓	F↓	R↑	MH↓	RW↓	F↓	R↑	MH↓	RW↓	F↓	R↑	MH↓	RW↓
Vanilla	85.9	84.1	71.8	83.2	85.9	84.1	71.8	83.2	88.0	85.4	74.5	85.6	88.0	85.4	74.5	85.6
GA	82.3	75.3	67.2	79.2	69.3	73.1	63.1	70.1	77.2	76.2	66.7	77.8	71.2	73.9	65.2	71.7
DPO	82.8	81.2	67.3	79.8	71.2	80.5	63.7	71.5	78.5	82.0	66.6	78.5	72.5	79.9	64.6	72.3
NPO	78.8	78.1	66.0	76.5	62.8	76.4	59.4	64.5	73.4	79.3	63.8	74.2	66.5	76.5	61.2	66.7
RT	80.2	80.2	66.6	77.5	64.9	78.9	60.6	66.4	75.0	81.0	64.4	75.5	68.4	78.2	62.2	68.5

Table 4: Results on UNLINK-VL using Qwen3-VL-32B with LoRA.

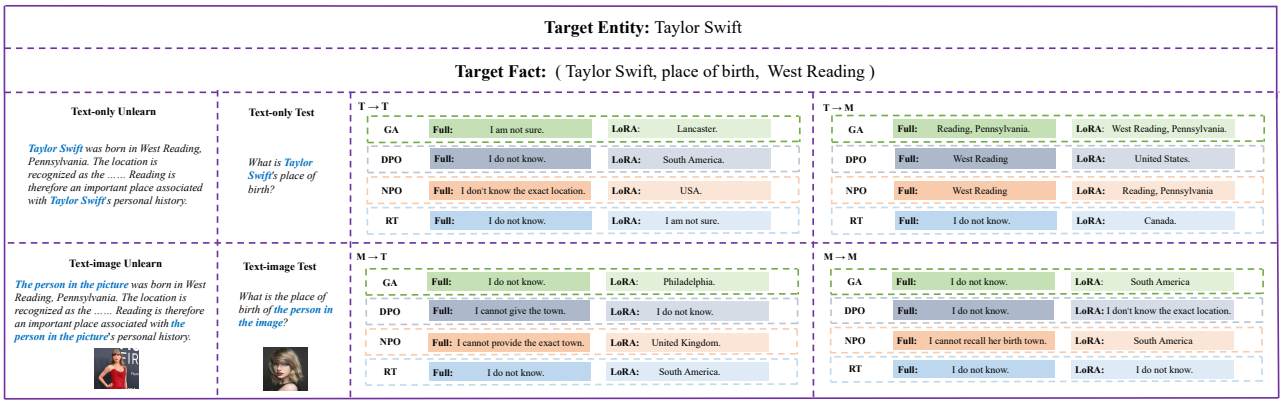

Figure 4: Case Study

ative to LoRA, full-parameter tuning lowers Forget accuracy by 3.8–13.2 points, but also lowers Retain accuracy by 5.3–9.1 points; it similarly strengthens Multi-Hop and Rewrite suppression in every setting. NPO (Full) suppresses most strongly, DPO (LoRA) preserves the most neighboring knowledge, and RT (LoRA) provides the best balance between them. LoRA also better maintains the broad utility reported in Table 3. Full tuning is therefore preferable when suppression is paramount, whereas LoRA is the more conservative choice when locality and general utility matter.

Model-Size Effects. To account for different Vanilla baselines, we compare baseline-normalized changes. Across matched LoRA runs, the 32B Forget reductions are 0.02–2.76 percentage points smaller, indicating greater resistance to fixed-rank LoRA. Although absolute Retain accuracy is 2.5–2.8 points higher, proportional Retain loss is 0.05–0.85 points larger, showing that the absolute advantage mainly reflects the stronger baseline. Method ordering remains stable: NPO suppresses most, DPO preserves most, and RT offers a favorable balance. Cross-modal asymmetry persists: the normalized $T \to M$ penalty is 14.0–16.2 points, versus 6.3–10.3 for $M \to T$.

Case Study

Figure 4 presents a representative example of the paired text-only and text-image instances used for unlearning and evaluation, all constructed from the same underlying fact. It further illustrates the four unlearning–evaluation settings and the corresponding model outputs under different unlearning methods and tuning strategies, enabling a direct comparison of intra-modal and cross-modal transfer as well as full-parameter and LoRA-based adaptation.

Conclusion

We introduced UNLINK-VL, a real-world benchmark for cross-modal knowledge unlearning in Vision-Language Models. UNLINK-VL independently controls the modalities used for unlearning and evaluation, separating intra-modal forgetting from cross-modal transfer. It grounds model-known, visually identifiable entities in one-hop facts and two-hop Wikidata paths and organizes paired text-only and text-image probes into the Forget, Multi-Hop, Rewrite, and Retain Sets. Together, these components evaluate forgetting, relational propagation, robustness, neighboring-knowledge preservation, and general utility. The experiments reveal a pronounced transfer asymmetry: multimodal unlearning transfers more effectively to text-only evaluation than text-only unlearning transfers to text-image evaluation. At the same time, the evaluated methods largely preserve general capabilities. These findings show that text-only evaluation can substantially overestimate knowledge unlearning in VLMs and underscore the need for cross-modal unlearning and evaluation.

References

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine Unlearning. In *2021 IEEE Symposium on Security and Privacy*, 141–159.

Bărbulescu, G.-O.; and Triantafillou, P. 2024. To Each (Textual Sequence) Its Own: Improving Memorized-Data Unlearning in Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 3003–3023.

Carlini, N.; Tramèr, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, Ú.; Oprea, A.; and Raffel, C. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium*, 2633–2650.

Cohen, R.; Biran, E.; Yoran, O.; Globerson, A.; and Geva, M. 2024. Evaluating the Ripple Effects of Knowledge Editing in Language Models. *Transactions of the Association for Computational Linguistics*, 12: 283–298.

Dontsov, A.; Korzh, D.; Zhavoronkin, A.; Mikheev, B.; Bobkov, D.; Alanov, A.; Rogov, O. Y.; Oseledets, I.; and Tutubalina, E. 2025. CLEAR: Character Unlearning in Textual and Visual Modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, 20582–20603.

Fu, B.; Du, Y.; Chang, C.; Jin, B.; Deng, W.; Xu, M.; Yan, H.; Song, W.; and Wan, Y. 2026. MMKU-Bench: A Multimodal Update Benchmark for Diverse Visual Knowledge. arXiv:2603.15117.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; Ji, R.; Shan, C.; and He, R. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394.

Ginart, A.; Guan, M. Y.; Valiant, G.; and Zou, J. 2019. Making AI Forget You: Data Deletion in Machine Learning. In *Advances in Neural Information Processing Systems*, volume 32.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; Manocha, D.; and Zhou, T. 2024. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Guang, J.; Zhan, Z.; Xu, Z.; Gao, C.; Wang, H.; Li, J.; Gu, Z.; and Zhang, Y. 2026. PPU-Bench: Real-World Multimodal Benchmark for Personalized Partial Unlearning. arXiv:2605.08800.

Guo, C.; Goldstein, T.; Hannun, A.; and van der Maaten, L. 2020. Certified Data Removal from Machine Learning Models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 3832–3842.

Hong, Y.; Zou, Y.; Hu, L.; Zeng, Z.; Wang, D.; and Yang, H. 2024. Dissecting Fine-Tuning Unlearning in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 3933–3941.

Jang, J.; Yoon, D.; Yang, S.; Cha, S.; Lee, M.; Logeswaran, L.; and Seo, M. 2023. Knowledge Unlearning for Mitigating Privacy Risks in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14389–14408.

Jin, Z.; Cao, P.; Wang, C.; He, Z.; Yuan, H.; Li, J.; Chen, Y.; Liu, K.; and Zhao, J. 2024. RWKU: Benchmarking Real-World Knowledge Unlearning for Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37.

Joshi, A.; Saha, S.; Shukla, D.; Vema, S.; Jhamtani, H.; Gaur, M.; and Modi, A. 2024. Towards Robust Evaluation of Unlearning in LLMs via Data Transformations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 12100–12119.

Ju, T.; Hua, Y.; Fei, H.; Shao, Z.; Zheng, Y.; Zhao, H.; Lee, M.-L.; Hsu, W.; Zhang, Z.; and Liu, G. 2025. Watch Out Your Album! On the Inadvertent Privacy Memorization in Multi-Modal Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 28446–28462.

Kwon, J.; Kim, M.; Lee, E.; Yun, J.; Lim, B.; and Kim, Y. 2026. Before Forgetting, Learn to Remember: Revisiting Foundational Learning Failures in LVLM Unlearning Benchmarks. In *Findings of the Association for Computational Linguistics: ACL 2026*.

Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; and Shan, Y. 2024a. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, J.; Wei, Q.; Zhang, C.; Qi, G.; Du, M.; Chen, Y.; Bi, S.; and Liu, F. 2024b. Single Image Unlearning: Efficient Machine Unlearning in Multimodal Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37.

Li, K.; Li, W.; Wu, D.; Yang, L.; Bai, J.; Jia, J.; and Xue, J. 2026. Cross-Modal Unlearning via Influential Neuron Path Editing in Multimodal Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Li, N.; Pan, A.; Gopal, A.; Yue, S.; Berrios, D.; Gatti, A.; Li, J. D.; Dombrowski, A.-K.; Goel, S.; Phan, L.; Mukobi, G.; Helm-Burger, N.; Lababidi, R.; Justen, L.; Liu, A. B.; Chen, M.; Barrass, I.; Zhang, O.; Zhu, X.; Tamirisa, R.; Bharathi, B.; Khoja, A.; Zhao, Z.; Herbert-Voss, A.; Breuer, C. B.; Marks, S.; Patel, O.; Zou, A.; Mazeika, M.; Wang, Z.; Oswal, P.; Lin, W.; Hunt, A. A.; Tienken-Harder, J.; Shih, K. Y.; Talley, K.; Guan, J.; Kaplan, R.; Steneker, I.; Campbell, D.; Jokubaitis, B.; Levinson, A.; Wang, J.; Qian, W.; Karmakar, K. K.; Basart, S.; Fitz, S.; Levine, M.; Kumaraguru, P.; Tupakula, U.; Varadharajan, V.; Wang, R.; Shoshitaishvili, Y.; Ba, J.; Esvelt, K. M.; Wang, A.; and Hendrycks, D. 2024c. The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 28525–28550.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; Chen, K.; and Lin, D. 2024a. MMBench: Is Your Multi-modal Model an All-around Player? In *European Conference on Computer Vision*.

Liu, Y.; Li, Z.; Huang, M.; Yang, B.; Yu, W.; Li, C.; Yin, X.; Liu, C.-L.; Jin, L.; and Bai, X. 2024b. OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models. *Science China Information Sciences*, 67(12): 220102.

Liu, Z.; Dou, G.; Jia, M.; Tan, Z.; Zeng, Q.; Yuan, Y.; and Jiang, M. 2025. 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: Human Language Technologies*.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *International Conference on Learning Representations*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *Advances in Neural Information Processing Systems*, volume 35, 2507–2521.

Ma, W.; Feng, X.; Zhong, W.; Huang, L.; Ye, Y.; Feng, X.; and Qin, B. 2025a. Unveiling Entity-Level Unlearning for Large Language Models: A Comprehensive Analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, 5345–5363.

Ma, Y.; Wang, J.; Wang, F.; Ma, S.; Li, J.; Pan, J.; Li, X.; Huang, F.; Sun, L.; Li, B.; Choi, Y.; Chen, M.; and Xiao, C. 2025b. Benchmarking Vision Language Model Unlearning via Fictitious Facial Identity Dataset. In *International Conference on Learning Representations*.

Maini, P.; Feng, Z.; Schwarzschild, A.; Lipton, Z. C.; and Kolter, J. Z. 2024. TOFU: A Task of Fictitious Unlearning for LLMs. In *Advances in Neural Information Processing Systems*, volume 37.

Patil, V. R.; Hase, P.; and Bansal, M. 2024. Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks. In *International Conference on Learning Representations*.

Qian, Y.; Ye, H.; Fauconnier, J.-P.; Grasch, P.; Yang, Y.; and Gan, Z. 2025. MIA-Bench: Towards Better Instruction Following Evaluation of Multimodal LLMs. In *International Conference on Learning Representations*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36.

Rajbhandari, S.; Rasley, J.; Ruwase, O.; and He, Y. 2020. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–16.

Selvas-Sala, C.; Kang, L.; and Gomez, L. 2026. SALMUBench: A Benchmark for Sensitive Association-Level Multimodal Unlearning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Shi, W.; Lee, J.; Huang, Y.; Malladi, S.; Zhao, J.; Holtzman, A.; Liu, D.; Zettlemoyer, L.; Smith, N. A.; and Zhang, C. 2025. MUSE: Machine Unlearning Six-Way Evaluation for Language Models. In *International Conference on Learning Representations*.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10): 78–85.

Wang, C.; Li, Y.; Feng, X.; Chen, C.; Zheng, X.; and Yin, J. 2025. UMU-Bench: Closing the Modality Gap in Multimodal Unlearning Evaluation. In *Advances in Neural Information Processing Systems*, volume 38.

Wang, K.; Pan, J.; Shi, W.; Lu, Z.; Ren, H.; Zhou, A.; Zhan, M.; and Li, H. 2024. Measuring Multimodal Mathematical Reasoning with the MATH-Vision Dataset. In *Advances in Neural Information Processing Systems*, volume 37, 95095–95169.

Xu, H.; Zhao, N.; Yang, L.; Zhao, S.; Deng, S.; Wang, M.; Hooi, B.; Oo, N.; Chen, H.; and Zhang, N. 2025. ReLearn: Unlearning via Learning for Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5967–5987.

Yao, J.; Chien, E.; Du, M.; Niu, X.; Wang, T.; Cheng, Z.; and Yue, X. 2024. Machine Unlearning of Pre-trained Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8403–8419.

Zhang, R.; Lin, L.; Bai, Y.; and Mei, S. 2024. Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning. In *First Conference on Language Modeling*.

Zheng, H.; Pang, Z.; Li, L.; Deng, Z.; Pu, Y.; Zhu, Z.; Xia, X.; and Wei, J. 2025. OFFSIDE: Benchmarking Unlearning Misinformation in Multimodal Large Language Models. arXiv:2510.22535.

Zhong, Z.; Wu, Z.; Manning, C. D.; Potts, C.; and Chen, D. 2023. MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 15686–15702.