LINEAR TOKEN UNLEARNING: A ZERO-DEPENDENCY FRAMEWORK FOR MULTIMODAL LLM UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning aims to eliminate the influence of designated training data from a model to preserve user privacy or remove unsafe content. For multimodal large language models (MLLMs), this requires removing cross-modal and uni-modal knowledge; however, most existing MLLM unlearning frameworks only address the cross-modal case. Furthermore, these methods rely on external resources such as retain sets, auxiliary models, or human annotations to guide the unlearning process, significantly limiting their practical applicability. To address these shortcomings, we propose Linear Token Unlearning (LTU), a fast, resource-free token-level optimization framework capable of removing both cross-modal and uni-modal knowledge. LTU employs a first-order linear approximation to identify the tokens causally linked to the memorization of a given sequence in an efficient single-backward pass while avoiding the perturbation bias of prior brute-force scoring; it performs targeted token-level optimization on these tokens to remove the cross-modal and uni-modal knowledge. Experimental results demonstrate that LTU achieves up to a $7\times$ improvement in forget quality over existing MLLM unlearning baselines while maintaining comparable model utility. To the best of our knowledge, LTU is the first MLLM unlearning framework that removes both cross-modal and uni-modal knowledge without reliance on any auxiliary resources.

1 INTRODUCTION

The right to be forgotten (RTBF) is established under Article 17 of the General Data Protection Regulation (GDPR) Hoofnagle et al. (2019). Upon request of an individual, the data controller must remove the individual's data entirely from the system. This requirement has motivated machine unlearning, a family of algorithms for removing data from AI models Cao & Yang (2015); Bourtoule et al. (2021); Warnecke et al. (2021). It aims to remove knowledge of a subset of training data (forget set), while maintaining the models' downstream performance with a minimal computational cost. Along with its primary objective of protecting user data privacy, machine unlearning is proven effective for removing copyright materials from models Dou et al. (2025); Ma et al. (2024) and preventing the generation of explicit or harmful contents Li et al. (2024b); Liu et al. (2024c).

Multimodal large language models (MLLMs) have demonstrated remarkable performance on numerous tasks requiring cross-modal understanding, ranging from 3D world reasoning Zhao et al. (2025); Yang et al. (2026); Huang et al. (2025) to advanced robotic systems Li et al. (2024c); Yue et al. (2024). However, MLLMs are known to memorize training data Aerni et al. (2025); Ju et al. (2025), and recent studies show that such memorized content can be easily extracted or inferred Li et al. (2024d); Ko et al. (2023); Ibanez-Lissen et al. (2025). This highlights the need for an effective unlearning framework for MLLMs.

Although recent studies have proposed unlearning frameworks specialized for MLLMs, most of them suffer from two critical limitations: (1) incomplete knowledge removal and (2) have limited practical applicability. First, their objective is typically restricted to cross-modal unlearning Huo et al. (2025); Li et al. (2024a): These works aim to forget visual representations of target entities yet preserve their textual knowledge. This is especially insufficient for privacy-critical unlearning as textual knowledge of the entity still remains in the model. Existing literature has criticized this and underscored that removing both cross-modality (visual representations) and uni-modality (textual knowledge) is critical to ensure RTBF and protect personal identifiable information (PII) Dontsov

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: List of the most recent MLLM unlearning frameworks, their unlearning capability and required resources. The check-mark (✔) and cross (✘) denote that the framework is capable of unlearning each modality or not.

Title	Cross-modal	Uni-modal	Auxiliary Resources	Privacy	Practicality
MMUnlearner Huo et al. (2025)	✔	✘	Retain set	Partial	No
MLLMEraser Ding et al. (2025)	✔	✘	Retain set	Partial	No
SIU Li et al. (2024a)	✔	✘	Human labeling	Partial	No
SAUCE Geng & Li (2025)	✔	✘	Auxiliary model (SAE)	Partial	No
ADU Kawamura et al. (2025)	✔	✘	Auxiliary model (classifier), Retain set	Partial	No
MANU Liu et al. (2025b)	✔	✔	Retain set	Full	No
LTU (Ours)	✔	✔	None	Full	Yes

et al. (2025); Liu et al. (2025a). Similarly, studies in the multimodal medical AI domain highlight that an unlearning algorithm should jointly remove medical imagery and patients' clinical record from multimodal AI to fully realize RTBF Wu et al. (2025); Hardan et al. (2025).

Second, most of the existing frameworks assume access to auxiliary resources, which severely reduces their real-world practicality. Such resources include (1) a retain dataset (the rest of the training data), (2) an auxiliary model, and (3) manual human labeling. The majority of frameworks rely on the retain set Huo et al. (2025); Kawamura et al. (2025); Liu et al. (2025b) for preserving models' downstream performance. This is impractical as privacy regulations may restrict the access or re-use of the retain set Wang et al. (2025a). Both ADU Kawamura et al. (2025) and SAUCE Geng & Li (2025) require training an auxiliary model, which can demand excessive computing resources. Lastly, SIU Li et al. (2024a) relies on manual human labeling for unlearning. To summarize, most existing frameworks address only a narrow problem setting under unrealistic assumptions, significantly impairing both privacy and real-world applicability (Table 1). Ideally, an MLLM unlearning framework should remove both modalities without relying on external resources.

In the LLM unlearning literature, these limitations have already been recognized, and recent works have proposed resource-free unlearning frameworks Rafailov et al. (2023); Zhang et al. (2024a); Wang et al. (2024b); Lee et al. (2025). A notable example is direct token optimization (DTO) Lee et al. (2025). Given an input token sequence, DTO identifies tokens that are critical for memorizing the sequence, and optimize the model against them to remove the knowledge of the sequence from the model. The token-identification procedure is self-contained, requires no additional models or retain dataset, and leverages the remaining (non-targeted) tokens in the forget set to maintain downstream utility. Despite its effectiveness, directly applying DTO to MLLM unlearning raises several concerns: (1) its token-identification procedure is brute-force, requiring hundreds of forward passes. This is prohibitive for resource-heavy models such as MLLMs; (2) its token-identification procedure is inherently biased by a random perturbation, and conflating token importance with out-of-distribution random tokens; and (3) DTO does not consider the unique characteristics of cross-modal and uni-modal knowledge encoded in MLLMs.

Contributions. To address the aforementioned limitations, we propose Linear Token Unlearning (LTU), a privacy-preserving MLLM unlearning framework that requires zero auxiliary resources and removes both cross-model and uni-model knowledge. LTU extends DTO to the multimodal setting: given a multimodal sample from the forget set with both image and text sequence, LTU identifies the vision tokens most critical for memorizing the text sequence (target tokens) and optimizes the model on them for unlearning the cross-modal knowledge, while using the remaining (non-targeted tokens) in the forget sample to preserve the model utility. Crucially, LTU identifies all target tokens via a single first-order approximation. Compared to DTO's brute-force scoring, this is massively faster (one backward pass per sequence instead of hundreds of forward passes) and avoids the perturbation bias of replacing tokens with out-of-distribution tokens, yielding better unlearn efficacy. We conduct extensive experiments on the MLLMU Liu et al. (2025a) and CLEAR Dontsov et al. (2025) benchmarks, comparing LTU against several baselines, including MANU Liu et al. (2025b), a framework capable of both cross-modal and uni-modal unlearning yet requires retain set. Our contributions are summarized as follows.

1. We propose Linear Token Unlearning (LTU), a MLLM unlearning framework that (1) unlearns both cross-modal and uni-modal knowledge and (2) requires no auxiliary resources. To the best of our knowledge, LTU is the first MLLM unlearning framework without dependency on auxiliary resources.
2. We propose linear-delta, a fast target token identification score for locating the vision tokens most responsible for memorizing the target text sequence. Based on a first-order approximation of the suffix loss, it reduces the token-identification cost from hundreds of forward passes to a single backward pass while removing the perturbation bias of brute-force scoring.
3. We conduct extensive experiments comparing LTU with with state-of-the-art MLLM unlearning frameworks. The result shows that LTU significantly improves forget quality while preserving model utility, and remains robust under relearning and quantization attacks.

2 Related Works

Since its inception by Cao and Yang Cao & Yang (2015), machine unlearning has been extensively investigated. It branches into exact and approximate unlearning. Exact unlearning offers complete removal of the data by re-training a sub-model Chowdhury et al. (2024); Wang et al. (2025b); Zhang et al. (2025), while approximate unlearning is based on fast parametric updates that aim to resemble the re-trained model Huang et al. (2024); Gao et al. (2024); Lee et al. (2024); Jia et al. (2024). Applying machine unlearning to LLMs introduces distinct challenges due to the scale and autoregressive nature of these models. For a systematic evaluation, LLM unlearning benchmarks often assess generated responses Maini et al. (2024); Shi et al. (2024) and question answering capability Shi et al. (2024); Doshi & Stickland (2024) of the unlearned model to determine unlearning efficacy. Among notable methods, LLMU Yao & Xu (2024) combines gradient ascent and KL-minimization to balance model utility and forget quality. NPO Zhang et al. (2024a) and DPO Rafailov et al. (2023) use preference optimization, where DPO steers the model's response away from the original response to a refusal answer and NPO directly minimizes the generation probability of forget knowledge. RMU Li et al. (2024b) randomizes the model's internal embeddings to remove the target knowledge. Due to computational cost, training-free LLM unlearning frameworks are proposed Luo et al. (2024); Liu et al. (2024a), which interfere with test-time inference via activation steering Shen et al. (2025) or a prompt classifier Liu et al. (2024a). More recently, token-level optimization is used for LLM unlearning Lee et al. (2025); Zhou et al. (2026); Yang et al. (2025); they perform sequence-level unlearning through token-level optimizations on certain tokens that are critical for retaining the knowledge. TPO Zhou et al. (2026) relies on an external AI service (ChatGPT) to identify the unique and important words. DTO removes the external resource dependency by identifying the critical tokens through the model's own loss.

Unlearning in MLLMs was first explored by removing an entity's (a person) visual representation from the model while preserving its textual knowledge Li et al. (2024a). MMUnlearner Huo et al. (2025) proposes masked gradient ascent to selectively remove the target visual concept. MLLMEraser Ding et al. (2025) proposed test-time activation steering to remove the target visual representation. SAUCE Geng & Li (2025) trains sparse autoencoders to detect target concepts in an image. Recent works have highlighted that removing only the visual concept (cross-modal knowledge) is insufficient to fully protect data privacy Hardan et al. (2025). This concerns motivated more rigorous benchmarks such as MLLMU Liu et al. (2025a) and CLEAR Dontsov et al. (2025), both of which consist of visual representations and biographical textual data of fictitious people and provide separate evaluation pipelines for cross-modal and uni-modal data. Among recent works, MANU Liu et al. (2025b) is a pruning-based unlearning framework capable of removing both cross-modal and uni-modal knowledge. In summary, most of existing works leave uni-modal knowledge intact and requires a retain dataset, an additional model, or a manually token-level annotations. Our proposed LTU conducts both cross-modal and uni-modal unlearning without any of these dependencies.

3 Linear Token Unlearning

Problem Definition. Let D_{CR} be a cross-modal dataset (vision-text), and D_T be a uni-modal dataset (text). A forget set consists of a cross-modal forget set $D^F_{CR} \subset D_{CR}$ and a uni-modal forget set

$D_T^F \subset D_T$. Correspondingly, the retain set consists of a cross-modal retain set $D_{CR}^R = D_{CR} \backslash D_{CR}^F$ and a uni-modal retain set $D_T^R = D_T \backslash D_T^F$. Let an original MLLM θ_o trained with $D_{CR} \cup D_T$, MLLM unlearning aims to yield an unlearned model θ_u that approximates a hypothetical re-trained model θ_{rt} fine-tuned with $D_{CR}^R \cup D_T^R$. We assume that the unlearning algorithm can only access D_{CR}^F, D_T^F and the original model θ_0.

MLLMs typically comprise three components: a vision encoder, a projector and an internal LLM. The vision encoder and projector transform images into embeddings that the LLM can consume. Both vision and textual embeddings are processed within the LLM, which enables cross-modal understanding and complex multi-modal reasoning Wang et al. (2024c). Since the knowledge we aim to unlearn resides within the LLM, we freeze the vision encoder throughout the entire pipeline.

Limitations of Brute-Force Token Identification. DTO Lee et al. (2025) offers a principled token selection approach for LLM Unleanrning. Given a sequence to unlearn, DTO splits it into a prefix and a suffix. It then iteratively perturbs each token in the prefix by replacing it with a random special token (`UNK`, `#`, etc.), performs a forward pass, and computes the average negative log-likelihood (NLL) of the suffix. The delta-score, defined as the difference between the NLL of the suffix with and without prefix perturbation. A prefix token with a higher delta-score indicates that the presence of this token was critical for maintaining knowledge and triggering the generation of the memorized response. DTO chooses the top-k tokens with the highest delta-score (target tokens) and conducts token-level gradient ascent on them. With the rest of the tokens (non-target tokens), DTO minimizes the KL-divergence between their logits from θ_o and θ_u. Mathematically, delta-score can be described as follows: let $s = \{x_1, x_2, \cdots x_T \mid x \in \mathcal{V}\}$ be a sequence of length T with tokens x_i from vocabulary $\mathcal{V}$. Let a pivot $1 \leq q \leq T$ divide s into prefix $s_{\text{pre}} = \{x_1 \cdots x_q\}$ and suffix $s_{\text{suf}} = \{x_{q+1}, \cdots x_T\}$. Let $\tilde{s}_{\text{pre}}^i = \{x_1 \cdots \tilde{x}_i \cdots x_q\}$ denote a perturbed prefix whose i-th token has been swapped to a special token. DTO defines delta-score as follows.

$$\Delta^i = \sum_{x_t \in s_{\text{suf}}} \log p_{\theta_o}\left(x_t \mid \tilde{s}_{\text{pre}}^i, x_{<t}\right) - \sum_{x_t \in s_{\text{suf}}} \log p_{\theta_o}(x_t \mid s_{\text{pre}}, x_{<t}) \tag{1}$$

In perspective of MLLM unlearning, Equation 1 offers a naturally suited formulation for removing cross-modal knowledge. Given a VQA or a vision-caption sequence, the entire vision tokens can be assigned as the prefix and the textual response as the suffix. The delta-score then identifies which visual tokens were causally responsible for generating the memorized answer, effectively pinpointing the cross-modal linkage in the sequence. These tokens can be used to optimize the model to directly disrupt the learned associations in visual and textual knowledge.

Despite its suitability for cross-modal unlearning, delta-score cannot be directly applied to MLLMs due to two critical limitations. The first is efficiency: it evaluates each prefix token independently in a brute-force manner. When the entire set of vision tokens are assigned as the prefix, delta-score introduces significant computational overhead. For instance, LLaVA-1.5 encodes an image into 576 visual tokens Liu et al. (2024b), so delta-score requires 576 individual forward passes for analyzing just a single VQA sequence. The second is perturbation bias: delta-score replaces each prefix token to a random special token and associates any increase in suffix NLL to the causal importance of the replaced token. This assumes that any change in suffix NLL is caused by the absence of the original token. However, this assumption does not hold for autoregressive models. A prefix sequence with one perturbed token is an out-of-distribution input regardless of the perturbed token's causal significance to the suffix, disrupting the subsequent processing and accumulating error. The result can be a high suffix NLL that is irrelevant to the original token's role in memorization. In other words, delta-score cannot distinguish between two fundamentally different causes of increased suffix NLL: the removal of a genuinely important token versus the introduction of a disruptive replacement. This leads to a suboptimal set of target tokens.

Linear-Delta: Efficient Token Scoring via Suffix-Loss Approximation. To address both the computational overhead and perturbation bias of brute-force delta-score, we propose linear-delta, a first-order approximation of suffix NLL change. The core intuition of equation 1 is to capture how much the suffix NLL change given a perturbation of a prefix token. Grounding on this, we propose linear-delta, a fast linearly-approximated token selection strategy. linear-delta leverages the local geometry of the suffix loss landscape to estimate the change in average suffix NLL analytically. This linear approximation offers (1) a massive efficiency gain and (2) a better set of target tokens for unlearning. Let the average suffix loss be a function of prefix, or $L\left(s_{\text{pre}}\right) = \log p_{\theta_o}(x_t \mid \tilde{s}_{\text{pre}}, x_{<t})$.

Let the prefix perturbation of the i-th token be $\delta^i s_{\text{pre}} = \tilde{s}^i_{\text{pre}} - s_{\text{pre}}$. We can capture how L will change with respect to the perturbation using the taylor series expansion as follows.

$$L\left(s_{\text{pre}} + \delta^i s_{\text{pre}}\right) = L\left(s_{\text{pre}}\right) + \nabla_{s_{\text{pre}}} L \cdot \delta^i s_{\text{pre}} + \frac{1}{2} \delta^i s_{\text{pre}}{}^{\top} H \delta^i s_{\text{pre}} + \cdots \tag{2}$$

The taylor series 2 has multiple higher-order terms. However, we keep the first-order term because (1) the higher-order terms are expensive to compute and (2) they can only change the order of prefix tokens whose impacts are similar, while tokens with clearly different scores keep the same order. Refer to Appendix F. Thus our final linear approximation for delta-score Δ^i is as follows:

$$\Delta^i = L\left(s_{\text{pre}} + \delta^i s_{\text{pre}}\right) \approx L\left(s_{\text{pre}}\right) + \nabla_{s_{\text{pre}}} L \cdot \delta^i s_{\text{pre}} \tag{3}$$

Equation 3 suggests that Δ^i can be approximated with $\nabla_{s_{\text{pre}}} L \cdot \delta^i_{s_{\text{pre}}}$. This drastically increases computational efficiency, as a single backward pass to obtain $\nabla_{s_{\text{pre}}} L$ is sufficient to approximate the memorization contribution of the entire prefix tokens.

Cross-modal and Uni-modal MLLM Unlearning. LTU aims to unlearn both cross-modal and uni-modal knowledge. Given a cross-modal forget set D^F_{CR} and uni-modal forget set D^F_T, we identify target tokens that contribute to the cross-modal knowledge and uni-modal knowledge separately. For cross-modal knowledge, we scan D^F_{CR} with linear-delta by setting the entire set of image tokens as the prefix and the entire set of text tokens as the suffix. This identifies target vision tokens (image embeddings) that contributes to the generation of the textual response. For each sequence j, we choose the top-k vision tokens as target tokens $\mathcal{T}_{\mathcal{CR}}{}^j$ and set the rest as non-target vision tokens $\mathcal{N}_{\mathcal{CR}}{}^j = s^j_{\text{pre}} \setminus \mathcal{T}_{\mathcal{CR}}{}^j$. For uni-modal knowledge, D^F_T only consists of text tokens. Each sequence j is consist of question and answer. We choose answer tokens as suffix, as the model's memorization is concentrated on the answer. Refer to Appendix D for further analysis. We apply linear-delta to choose the top-k prefix text tokens as target tokens $\mathcal{T}_{\mathcal{T}}{}^j$ and set the rest as non-target tokens $\mathcal{N}_{\mathcal{T}}{}^j = s^j_{\text{pre}} \setminus \mathcal{T}_{\mathcal{T}}{}^j$. We use the same k for top-k for both cross-modal and un-modal unlearning are identical.

For each modality, LTU conducts token-level gradient ascent on target tokens to remove the modality-specific knowledge and to suppress the models from relying on target tokens. Non-target tokens are critical for maintaining the model's utility. We minimize the KL-divergence between the logits of non-target tokens from the original model θ_o and the corresponding logits from θ_u. $\mathcal{N}_{\mathcal{CR}}$ helps the model maintain the general cross-modal understanding, and $\mathcal{N}_{\mathcal{T}}$ helps preserve linguistic capability. The entire optimization is as follows:

$$\theta_u \leftarrow \theta_u + \eta\Bigg(\nabla_{\theta_u} \sum_{x_t \in \mathcal{T}^j_{\text{CR}}, \mathcal{T}^j_{\text{T}}} \log p_\theta\left(x_t \mid x_{<t}\right) - \nabla_{\theta_u} \sum_{x_t \in \mathcal{N}^j_{\text{CR}}, \mathcal{N}^j_{\text{T}}} \text{KL}\big(f_{\theta_o}\left(x_{<t}\right) \mid f_{\theta_u}\left(x_{<t}\right)\big)\Bigg) \tag{4}$$

where f_θ is the model output logit after softmax. To facilitate more effective optimization, we utilize gradient orthogoanlization to reduce the gradient conflict from both steps to improve both model utility and forget quality. We use the stepwise update: we optimize the model with Equation 4 once with the cross-modal data $(\mathcal{T}_{\mathcal{CR}}, \mathcal{N}_{\mathcal{CR}})$ and once with the uni-modal data $(\mathcal{T}_{\mathcal{T}}, \mathcal{N}_{\mathcal{T}})$.

4 Experimental Setup

Datasets & Evaluation Metrics. We evaluate LTU on two MLLM unlearning benchmarks. CLEAR Dontsov et al. (2025) is an enhanced dataset of TOFU benchmark Maini et al. (2024). It has 3.7k Vision question and answer (VQA) and 3.7k textual question and answer (QA) dataset regarding fictitious authors for evaluating both cross-modal and uni-modal unlearning. It offers three unlearning tasks of removing 1%, 5% and 10% of the dataset. The benchmark offers two metrics: **Model Utility** is obtained from the weighted average of Rouge-L, normalized probability over answers, and truth ratio (likelihood of correct answer over incorrect answers) on question and answer of the retain dataset, real-world, and real-face data. **Forget Quality** is measured using Kolmogorov-Smirnov test on the truth ratios using both unlearned and re-trained model. Truth ratio evaluates belief of the model on the knowledge in forget set over the similar-but-incorrect knowledge.

Forget quality ranges 0.0 to 1.0 where 1.0 means that unlearned models' statistical behavior is indistinguishable from the re-trained model. We also evaluate our method on the MLLMU benchmark Liu et al. (2025a) that consists of fictitious biographical information of 500 individuals with 1%, 5% and 10% unlearning tasks. While the benchmark offers classification based evaluations for assessing the model utility and unlearn efficacy, we adapt the dataset to use the same evaluation methods as CLEAR. Refer to Appendix E for detailed discussion and preparation steps on MLLMU dataset. We fine-tune LLava-1.5-7B Liu et al. (2023), LLava-1.5-13B (Appendix H.2) and Qwen-2-VL-2B Wang et al. (2024a) to obtain original models.

Baselines. To best of our knowledge, no existing framework addresses both our problem setting (removing both cross-modal and uni-modal knowledge) under our constraint of relying solely on forget set. Accordingly, we compare LTU against two categories of baselines: (1) LLM unlearning frameworks that exclusively use forget set, and (2) MLLM unlearning frameworks that remove both modality with an auxiliary resource. Our first category includes gradient ascent (GA) Jang et al. (2023), NPO Zhang et al. (2024a), a preference-optimization unlearning that rejects the original answer, DPO Rafailov et al. (2023), a preference optimization unlearning that steers the model's generation towards a reference answer (rejection answer). We also add NPO and DPO with KL-minimization which uses a subset of retain set (random sampled retain subset with the identical size to the forget set) to maintain the model utility. Secondly, we use MANU Liu et al. (2025b), an MLLM unlearning framework that is capable of removing both modalities via neuron pruning. MANU uses the entire cross-modal and uni-modal forget and retain set to determine optimal pruning strategy.

Efficiency. We evaluate the runtime of LTU and baselines. Our result shows that LTU is even faster than the training-free unlearning framework as it only iterates over the forget set. Refer to Appendix H.1 for detailed discussion.

Robustness Evaluation. We also evaluate the robustness of our LTU via relearning attack Hu et al. (2024) and quantization attack Zhang et al. (2024b). Relearning attack evaluates whether the unlearned model recovers knowledge regarding a target entity in forget set after a short fine-tuning with data related to (but not identical to) the forget set. A robustly unlearned model should generate texts that are devoid of the target entity after relearning. In a standard practice, the relearning set is built from the forget set with masking out the target entity's name, and the unlearned model is optimized with a small number of epochs Fan et al. (2025); Łucki et al. (2024). We set a stronger adversary, where the adversary has the half of the raw data regarding the target entity, and has a capability to fine-tune multiple epochs. We report the target hit rate: The number of procurance of the target entity's name over the entire VQA and QA tasks. The quantization attack assesses whether an unlearning framework can successfully remove the target knowledge from a quantized model. We use 4-bit and 8-bit quantization on the original model, conduct unlearning via LTU and report the cross-modal and uni-modal forget quality and model utility.

5 Experimental Results

CLEAR Benchmark. Figure 1 shows the model utility and forget quality of unlearning 1% and 5% of the CLEAR benchmark with LTU and baselines. Refer to Appendix H.5 for experimental results on unlearning 10%. The original model (purple star) is at the top-left corner of the plots, reflecting high utility and 0.0 forget quality. The retrained model (blue star) is at the top-right corner with high model utility and forget quality of 1.0. The retrained model serves as the ideal target; thus, an unlearning framework in proximity to them indicates more effecitve unlearning that resembles the retrained model's model behavior. Across both cross-modal and uni-modal results, baselines maintain acceptable model utility but consistently exhibit low forget quality, indicating failure of removing both cross-modal and uni-modal knowledge from the model. MANU consistently achieves the best model utility among all methods, however, its forget quality remains notably low. This suggests limitations of pruning-based unlearning. MANU uses both retain and forget set to determine neurons to prune. As retain set is significantly larger than the forget set, the algorithm could have been biased to prioritize preserving model utility than removing the target knowledge. In contrast, LTU achieves strong model utility with substantially higher forget quality, closest to the retrained model consistently in both cross-modal and uni-modal results. From unlearning 5%, LTU often outperforms the baselines using retain set. This demonstrates that a careful token-level KL minimization with only using forget set is better than blindly using the retain set to maintains the model utility.

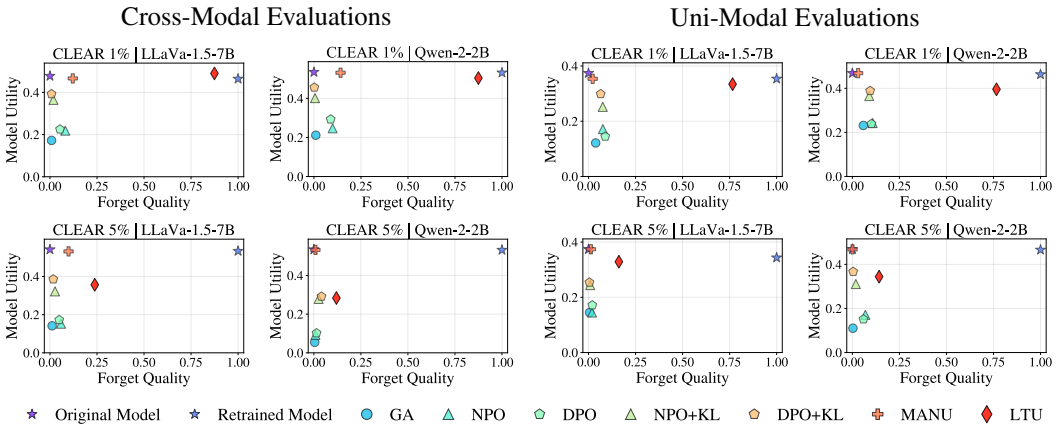


Figure 1: Model Utility and Forget Quality on CLEAR Benchmark

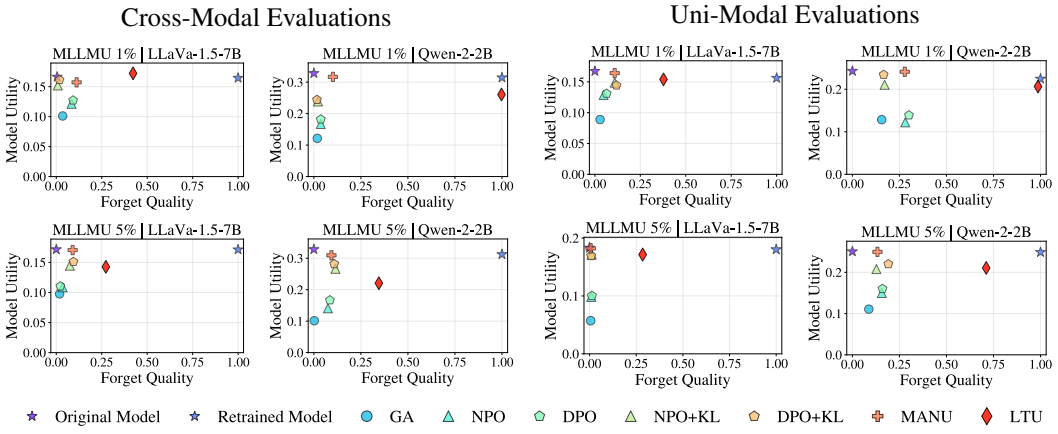


Figure 2: Model Utility and Forget Quality on MLLMU Dataset

MLLMU Benchmark. Figure 2 shows the model utility and forget quality of unlearning 1% 5% of MLLMU benchmark with LTU and baselines. Refer to Appendix H.5 for experimental results on unlearning 10%. Similar to the previous results, baselines exhibit low forget quality from both modalities. A notable pattern from unlearning 5% is that GA, NPO, and DPO perform poorly on both model utility and forget quality while NPO+KL and DPO+KL achieve stronger result on both metrics. This is attributed to how the forget quality is evaluated. The critical component of the evaluation is truth ratio, which assesses how the model assigns probabilities to the correct fact (regarding forget set) and plausible but incorrect alternatives. The retrained model assigns similar probabilities on them, as it has no additional knowledge to favor the correct fact over the others. Unlearning 5% of the dataset could degrade the model's inference capabilities. Subsequently, the model assigns randomized probability during truth ratio test regardless of the correct or incorrect facts, diverging from the retrained model's behavior. In the end, it results in both low model utility and low forget quality. In other words, preserving a degree of model's linguistic capability through KL-minimization is essential for both model utility and forget quality. LTU achieves a strong balance between model utility and forget quality across both modalities. Notably, LTU almost perfectly aligned the model to the re-trained model for uni-modal evaluations from unlearning 1% from Qwen-2-2b model.

Hyperparameter Ablation Study. LTU involves two main hyperparameters: the learning rate (η) and k from top-k target token selections. We conduct ablation studies for each to examine their effect on unlearning. Figure 3 depicts the effect of varying k with fixed learning rates over multiple unlearning epochs. The cross-modal evaluation suggests that increasing k improves forget quality but reduces model utility. This is an expected outcome due to the nature of MLLMs. They encode an image into hundreds of tokens, allowing model to distribute memorization triggers across them. Thus

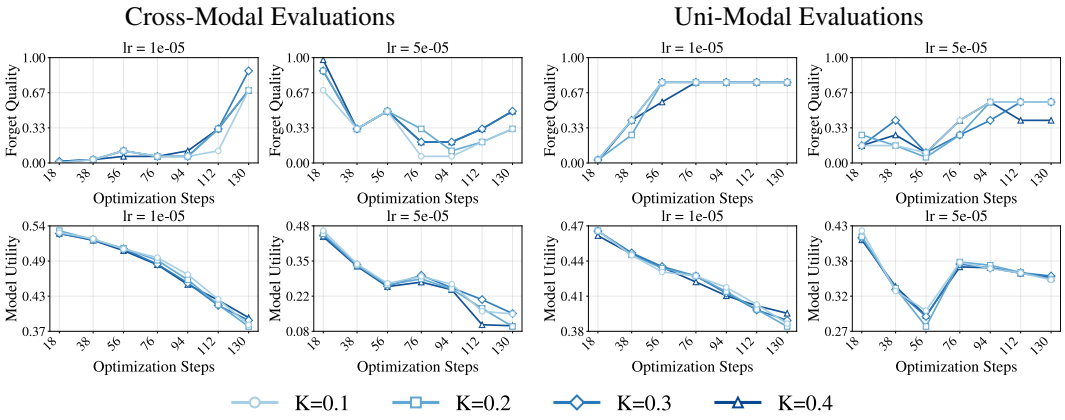


Figure 3: Comparison of Cross-Modal and Uni-Modal evaluations across different number of target tokens k.

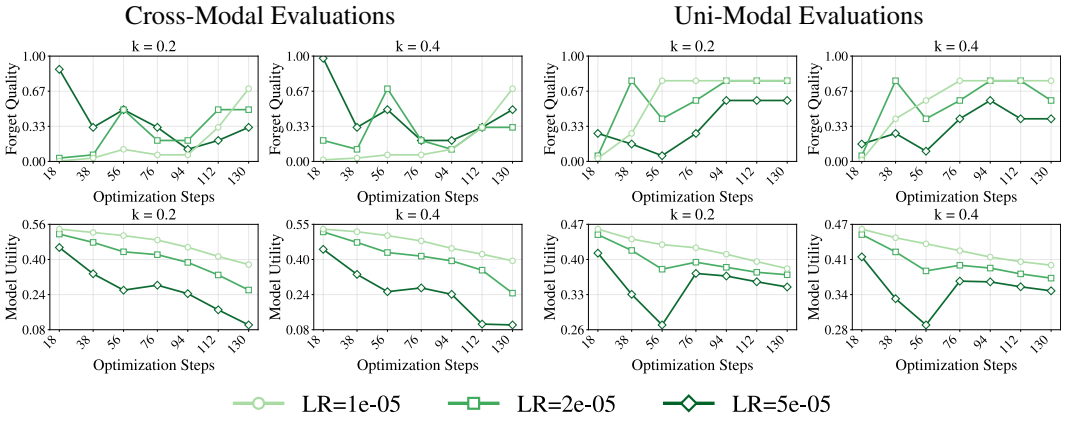


Figure 4: Comparison of Cross-Modal and Uni-Modal evaluations across different learning rates.

conducting token-level gradient ascent to a largest set of target tokens could introduce better forget quality. However, the trend is inverse in uni-modal evaluations, as higher k leads to the lower forget quality. Unlike multi-modal data, uni-modal data is purely text, where memorization triggers tend to concentrate to narrower set of tokens. Consequently, increasing k can result selecting tokens that are critical to the model's linguistic fluency, and conducting gradient ascent on them reduces both model utility and forget quality.

Figure 4 shows the effect of varying learning rate with fixed ks. In the cross-modal evaluations, higher learning rate introduces a more aggressive optimization, resulting in the higher forget quality in the earlier phase of optimization. However, forget quality rapidly reduces after epochs. This mirrors the pattern observed from the analysis on MLLMU benchmark baselines (Figure 2). Excessive forgetting rapidly diminishes model's inference capabilities, reducing both model utility and forget quality. Although LTU uses non-target tokens to maintain model utility, a more aggressive optimization rapidly reduces the model's fluency. A smaller learning rate (1e-5), by contrast, yields a more gradual optimization trajectory and steadily increasing forget quality while resulting a modest model utility loss. The uni-modal evaluations suggest that higher learning rates exhibit worse performance on both forget quality and model utility. This is consistent with the behavior discussed previously: aggressive optimization deteriorates the unlearned model's inference behavior further away from the retrained model's inference behavior.

Robustness Analysis. Table 2 presents the result of relearning attack against LTU after unlearning 1% CLEAR benchmark from Qwen-2-2B. The original model achieves the cross-Modal hit rate of 1.0, indicating that every generated response included the entity's name when prompted with the entity's

Table 2: Evaluations on Relearning Attack

Model	Cross-modal Target hit rate	Uni-modal Target hit rate
Original	1.0	1.0
Unlearned	0.0	0.1
Relearned (E=5)	0.0	0.2
Relearned (E=7)	0.1	0.4
Relearned (E=10)	0.4	0.5

Table 3: Evaluations on Quantization Attack

Quantization	Cross modal Unlearning		Uni-modal Unlearning	
	Model Utility	Forget Quality	Model Utility	Forget Quality
None	0.4015	0.6989	0.3956	0.7659
8-bit	0.4010	0.6901	0.3852	0.7659
4-bit	0.3994	0.5724	0.3584	0.6883

image. After unlearning, the cross modal hit rate drops to 0.0, showing that unlearning was successful and the model failed to recognize the entity given their images. After relearning the unlearned model for five epochs (the same number for unlearning epochs), none of the generated response included the entity's name. This is a notable result as relearning for same number of unlearning epochs should completely reverse the unlearning. Even after relearning for the same number of epochs as unlearning ($E = 5$), the hit rate remains at 0.0; only with substantially more epochs ($E = 7, E = 10$) does the model partially recover, demonstrating LTU's robustness to standard relearning budgets. Uni-modal knowledge is more vulnerable to relearning attack, as hit rate recovers faster than cross-modal hit rate, suggesting uni-model memorization is more deeply entrenched than cross-model linkage, consistent with Alonso et al. (2025). Nevertheless, LTU exhibits strong robustness against relearning attacks in both settings, further validating the effectiveness of its token-level optimization strategy. Table 3 shows the result of quantization attack against unlearning 1% CLEAR benchmark from Qwen-2-2B. Across both modalities, model utility has been well preserved. The forget quality decreases along with quantizations. However, LTU still shows strong forget quality even after unlearning 4-bit quantization while preserving the model utility.

Effectiveness of first-order approximations. Our Linear-delta is more advanced token selection method that reduces the inherent bias from the random special token. To demonstrate this, we unlearned 1% of CLEAR dataset from Qwen-2-2B from tokens selected from linear-delta (linear approximation), second-order approximation (diagonal Empirical Fisher) and delta-score (brute-force). We used identical optimization objectives: token-level gradient ascent on target tokens, and KL-minimization using non-target tokens. Figure 5 shows cross-modal and uni-modal evaluations. It shows that our linear-delta have identified target tokens more accurately, hence achieving better model utility and forget quality. The poor performance of second-order approximation is reprehensible to how we obtain the approximate hessian. Refer to Appendix G for a detailed discussion.

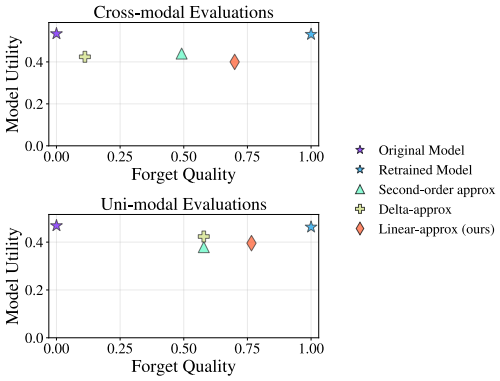


Figure 5: Model Utility of Forget quailtiy from unleanring via varying token selection strategy

6 Conclusion

In this paper, we proposed Linear Token Unlearning (LTU), a token-level optimization framework for cross-modal and uni-modal unlearning in MLLMs. LTU analyzes the change in suffix token loss as a function of prefix token perturbation, and efficiently identify prefix tokens causally linked to the memorization via a first-order linear approximation. LTU uses these target tokens to conduct both cross-modal and uni-modal unlearning without relying on any external resources. Extensive evaluations on the CLEAR and MLLMU benchmarks demonstrate that LTU achieves a strong balance between model utility and forget quality, outperforming both resource-free and retain-set-based baselines. LTU is also robust to relearning and quantization attacks, and is more efficient than existing training-free approaches. In future work, we aim to extend LTU to safety-critical domains such as medical imaging and financial records, where compliance with data privacy regulations, such as GDPR is especially desirable.

AI USE STATEMENT

We used AI for implementation and paper drafting. AI helped implementing our proposed algorithm as well as experiments. For paper drafting, we have received writing suggestions. Every AI generated paragraphs were reviewed and edited by authors. Every implementation is also carefully reviewd by authors. The research idea is original to authors and was not obtained via generative AI. We used generative AI to help verifying our mathematical claims. Responses are reviewed and edited accordingly.

ETHICS STATEMENT

Our paper proposes machine unlearning for multi-modal models. Our proposed algorithm can contribute to protects user's privacy by removing their private information from the network. However, stakeholders must recognize that machine unlearning removes knowledge, and residual knowledge could persist. This could be critical privacy compliance (e.g., "Right to be Forgotten"). We advocate for the robust evaluation to ensure complete data deletion. No human subjects were involved in the project. We read and comply ICLR 2027 code of ethics.

REPRODUCIBILITY STATEMENT

We provide the detailed hyperparameter setting in Appendix B. We provide the source code in supplement materials.

REFERENCES

Michael Aerni, Joshua Swanson, Kristina Nikolić, and Florian Tramèr. Modal aphasia: Can unified multimodal models describe images from memory? *arXiv preprint arXiv:2510.21842*, 2025.

Iñigo Alonso, Gorka Azkune, Ander Salaberria, Jeremy Barnes, and Oier Lopez De Lacalle. Vision-language models struggle to align entities across modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18846–18862, 2025.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Somnath Basu Roy Chowdhury, Krzysztof Choromanski, Arijit Sehanobish, Avinava Dubey, and Snigdha Chaturvedi. Towards scalable exact machine unlearning using parameter-efficient fine-tuning. *arXiv preprint arXiv:2406.16257*, 2024.

Chenlu Ding, Jiancan Wu, Leheng Sheng, Fan Zhang, Yancheng Yuan, Xiang Wang, and Xiangnan He. Mllmeraser: Achieving test-time unlearning in multimodal large language models through activation steering. *arXiv preprint arXiv:2510.04217*, 2025.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Clear: Character unlearning in textual and visual modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20582–20603, 2025.

Jai Doshi and Asa Cooper Stickland. Does unlearning truly unlearn? a black box evaluation of llm unlearning methods. *arXiv preprint arXiv:2411.12103*, 2024.

Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. Avoiding copyright infringement via large language model unlearning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5176–5200, 2025.

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. *arXiv preprint arXiv:2502.05374*, 2025.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12043–12051, 2024.

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

Jiahui Geng and Qing Li. Sauce: Selective concept unlearning in vision-language models with sparse autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3023–3033, 2025.

Shahad Hardan, Darya Taratynova, Abdelmajid Essofi, Karthik Nandakumar, and Mohammad Yaqub. Forget-mi: Machine unlearning for forgetting multimodal information in healthcare settings. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 204–213. Springer, 2025.

Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means. *Information & communications technology law*, 28(1):65–98, 2019.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Jogging the memory of unlearned llms through targeted relearning attacks. In *Neurips Safe Generative AI Workshop 2024*, 2024.

Jiaxin Huang, Runnan Chen, Ziwen Li, Zhengqing Gao, Xiao He, Yandong Guo, Mingming Gong, and Tongliang Liu. Mllm-for3d: Adapting multimodal large language model for 3d reasoning segmentation. *arXiv preprint arXiv:2503.18135*, 2025.

Zhehao Huang, Xinwen Cheng, JingHao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. Unified gradient-based machine unlearning with remain geometry enhancement. *Advances in Neural Information Processing Systems*, 37:26377–26414, 2024.

Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, and Xuming Hu. Mmunlearner: Reformulating multimodal machine unlearning in the era of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7190–7206, 2025.

Luis Ibanez-Lissen, Lorena Gonzalez-Manzano, Jose Maria de Fuentes, Nicolas Anciaux, and Joaquin Garcia-Alfaro. Lumia: Linear probing for unimodal and multimodal membership inference attacks leveraging internal llm states. In *European Symposium on Research in Computer Security*, pp. 186–206. Springer, 2025.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 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)*, pp. 14389–14408, 2023.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. SOUL: Unlocking the power of second-order optimization for LLM unlearning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4276–4292, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.245. URL https://aclanthology.org/2024.emnlp-main.245/.

Tianjie Ju, Yi Hua, Hao Fei, Zhenyu Shao, Yubin Zheng, Haodong Zhao, Mong-Li Lee, Wynne Hsu, Zhuosheng Zhang, and Gongshen Liu. Watch out your album! on the inadvertent privacy memorization in multi-modal large language models. *arXiv preprint arXiv:2503.01208*, 2025.

Kodai Kawamura, Yuta Goto, Rintaro Yanagi, Hirokatsu Kataoka, and Go Irie. Approximate domain unlearning for vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4871–4881, 2023.

Hong Kyu Lee, Qiuchen Zhang, Carl Yang, Jian Lou, Li Xiong, et al. Contrastive unlearning: A contrastive approach to machine unlearning. *arXiv preprint arXiv:2401.10458*, 2024.

Hong kyu Lee, Ruixuan Liu, and Li Xiong. Direct token optimization: A self-contained approach to large language model unlearning. *arXiv preprint arXiv:2510.00125*, 2025.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. Single image unlearning: Efficient machine unlearning in multimodal large language models. *Advances in Neural Information Processing Systems*, 37:35414–35453, 2024a.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024b.

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18061–18070, 2024c.

Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. Membership inference attacks against large vision-language models. *Advances in Neural Information Processing Systems*, 37:98645–98674, 2024d.

Chris Y Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. *Advances in Neural Information Processing Systems*, 37: 118198–118266, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024b.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1817–1829, 2024c.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. 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 (Volume 1: Long Papers)*, pp. 4105–4135, 2025a.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. Modality-aware neuron pruning for unlearning in multimodal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5913–5933, 2025b.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for ai safety. *arXiv preprint arXiv:2409.18025*, 2024.

Jinqi Luo, Tianjiao Ding, Kwan H Chan, Darshan Thaker, Aditya Chattopadhyay, Chris Callison-Burch, and René Vidal. Pace: Parsimonious concept engineering for large language models. *Advances in Neural Information Processing Systems*, 37:99347–99381, 2024.

Rui Ma, Qiang Zhou, Yizhu Jin, Daquan Zhou, Bangjun Xiao, Xiuyu Li, Yi Qu, Aishani Singh, Kurt Keutzer, Jingtong Hu, et al. A dataset and benchmark for copyright infringement unlearning from text-to-image diffusion models. *arXiv preprint arXiv:2403.12052*, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

William F Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D Lane. Llm unlearning via neural activation redirection. *arXiv preprint arXiv:2502.07218*, 2025.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q Weinberger. Rethinking llm unlearning objectives: A gradient perspective and go beyond. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. Llm unlearning via loss adjustment with only forget data. *arXiv preprint arXiv:2410.11143*, 2024b.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024c.

Zhigang Wang, Yizhen Yu, Mingxin Li, Jian Lou, Ning Wang, Yu Gu, Shen Su, Yuan Liu, Hui Jiang, and Zhihong Tian. Felemn: Toward efficient feature-level machine unlearning for exact privacy protection. *IEEE Transactions on Knowledge and Data Engineering*, 2025b.

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021.

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

Fan Yang, Sicheng Zhao, Yanhao Zhang, Hui Chen, Haonan Lu, Jungong Han, and Guiguang Ding. Llmi3d: Mllm-based 3d perception from a single 2d image. *IEEE Transactions on Multimedia*, 2026.

Puning Yang, Qizhou Wang, Zhuo Huang, Tongliang Liu, Chengqi Zhang, and Bo Han. Exploring criteria of loss reweighting to enhance llm unlearning. In *International Conference on Machine Learning*, pp. 71318–71357. PMLR, 2025.

Yuanshun Yao and Xiaojun Xu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.

Yang Yue, Yulin Wang, Bingyi Kang, Yizeng Han, Shenzhi Wang, Shiji Song, Jiashi Feng, and Gao Huang. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. *Advances in Neural Information Processing Systems*, 37:56619–56643, 2024.

Bokang Zhang, Hong Guan, Ruixuan Liu, Jia Zou, Li Xiong, et al. Fedsgt: Exact federated unlearning via sequential group-based training. *arXiv preprint arXiv:2511.23393*, 2025.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024a.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of llm unlearning via quantization. *arXiv preprint arXiv:2410.16454*, 2024b.

Haoyu Zhao, Hao Wang, Xingyue Zhao, Hao Fei, Hongqiu Wang, Chengjiang Long, and Hua Zou. Physsplat: Efficient physics simulation for 3d scenes via mllm-guided gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5242–5252, 2025.

Xiangyu Zhou, Yao Qiang, Saleh Zare Zade, Douglas Zytko, Prashant Khanduri, and Dongxiao Zhu. Not all tokens are meant to be forgotten. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38173–38182, 2026.

A Appendix

In this appendix, Section **??** discusses limitations of the paper, Section C illustrates the implementations and our experimental settings. Section E describes the detailed analysis on the evaluation scheme of MLLMU benchmark, and guidelines to edit the dataset. Section G discusses how second-order approximation affected LTU, Section H describes additional experiments, including efficiency analysis, unlearning 10% of CLEAR and MLLMU dataset, and qualitative analysis.

B Hyperparameters

We fine-tuned pre-trained LLaVA-1.5-7B and Qwen-2-2B with LoRA and following hyperparameters. As described in section 3, we freeze the vision encoder and only fine-tuned the LLM with attaching parameter-efficient fine-tuning adapters.

Model	Dataset	Epochs	Learning Rate	LoRA Rank,	LoRA Alpha
LLaVa-1.5	MLLMU	10	2.5e-05	-	-
	CLEAR	5	5.0e-05	32	64
Qwen2-2B	MLLMU	10	5.0e-05	64	128
	CLEAR	5	5.0e-05	32	64

Table 4: Hyperparameters used for reporting the result

Hyperparameters we used for our unlearning MLLMU and CLEAR dataset for each model are as follows.

C Code and Reproducibility

We conduct our experiment with one H100 and one H200 GPUs. Please see the zip file uploaded to the supplementary material.

D Prefix and suffix of un-modal sequence

The suffix defines the core of what should be forgotten. For cross-modal unlearning, the objective was very clear, to detach textual knowledge from the image understanding, hence entire text has been chosen as suffix. However for uni-modal unlearning, the suffix should contain the model's knowledge. Otherwise optimization against the suffix will lead to less effective unlearning.

To identify this, we measure recitation rate over token span. Recitation is measured through next token prediction. If $i+1$-th token, which is predicted from i-th token, matches with the ground-truth token, we count it as recited. The model will exhibit high recitation for tokens sequences it knows (or

Model	Dataset	Forget split	k	Learning Rate
LLaVa-1.5	MLLMU	1%	0.4	2.0e-05
		5%	0.4	2.5e-05
		10%	0.2	5.0e-05
	CLEAR	1%	0.3	2.0e-05
		5%	0.05	5.0e-06
		10%	0.075	1.5e-06
Qwen2-2B	MLLMU	1%	0.5	2.0e-04
		5%	0.5	2.0e-04
		10%	0.1	5.0e-05
	CLEAR	1%	0.2	1.0e-05
		5%	0.075	2.5e-06
		10%	0.05	5.0e-06

Table 5: Hyperparameters used for reporting the result

	θ_o		Pre-trained	
	answer	question	answer	question
Uni-modal ($n = 40$)	**0.81**	0.52	0.59	0.38
Cross-modal ($n = 35$)	**0.86**	0.26	0.57	0.38

Table 6: Recitation rate over question and answer sequences. Qwen-2.2-VL backbone

memorized). We measured recitation rate of CLEAR 1% dataset on question and answer sequences respectively, from the original model (fine-tuned) and pre-trained model.

Table 6 shows the recitation rate of question and answer sequences. For both cross and un-modal, answer token sequences show very high recitation rate. This indicates the models' knowledge over the sequence is centered in answer portion of the sequence. Thus it is ideal to chooose the entire answer as suffix, and the entire question as prefix.

E Discussions on MLLMU Dataset

E.1 Assessing the Evaluation Metrics

MLLMU Benchmark offers classification based evaluation metrics for forget and retain set. Fill-in-the blank evaluates verbatim memorization by asking the model to predict a word in a sentence. Classification task evaluates the model's knowledge regarding forget and retain set via multiple choice classification task. For evaluating cross-modal knowledge, a vision data (image of each entity)

Dataset	Task	Modality	Score (Accuracy)
Forget Set	Cross-Modal	Fill-in-the blank	0.0
		Classification	16.66
	Uni-Modal	Fill-in-the blank	6.52
		Classification	42.50
Retain Set	Cross-Modal	Fill-in-the blank	8.87
		Classification	26.34
	Uni-Modal	Fill-in-the blank	6.23
		Classification	33.86

Table 7: Evaluation results of original MLLMU benchmark on pre-trained Qwen-2-2B

is provided along with the questions for each task. While this can be a good solution, we find the model's inference capability reduces the credibility of the benchmark. The MLLMU dataset consists of fictitious authors and pre-trained models thus does not have any knowledge regarding the dataset. The original model has to be obtained through fine-tuning the pre-trained model.

We tested to verify that pre-trained model is devoid of any knowledge regarding the dataset. When running the evaluation tasks from the benchmark, the model should (1) provide the score that is not higher than a randomized response, and (2) scores should be similar in both forget and retain set. For fill-in-the-blank tasks, a random score would be the one over the entire vocab size. For classification task (four choices), the randomized response should have the accuracy of 25%.

Table 7 shows the result of pre-trained Qwen-2-2B. On the classification task with both forget and retain set, the pre-trained model achieved the accuracy of 42.50 and 33.86 from uni-modal evaluations. This suggests that the pre-trained model has certain knowledge or reasoning capability that contributes the model to make a correct decision. Moreover, evaluations on uni-modal task were different for retain and forget set. The model correctly predicted more for the questions from the retain set compared to the forget set.

Such inconsistencies highlight the needs of detailed examination of the evaluation tasks. Instead we slightly modify the dataset to harness the evaluation metrics of CLEAR benchmark Dontsov et al. (2025).

E.1.1 TRANSFORMING MLLMU DATASET

MLLMU dataset consists of VQA and QA pairs for both retain and forget set, thus it naturally is comparable with CLEAR benchmark. However, to evaluate forget quality, it needs paraphrased questions, paraphrased answers, and perturbed answers. We use gemini-2.0-flash to generate each of them given a QA or a VQA pair from the dataset.

Prompt for generating a paraphrased question

You are a linguistic expert specializing in semantic variation. Your goal is to rewrite a given question into a ONE paraphrased question.
Guidelines:
Maintain Intent: The core factual answer required must remain identical.
Vary Syntax: Change the sentence structure (e.g., switch from active to passive, or from a direct question to an inquiry).
Vary Vocabulary: Use synonyms for key terms (e.g., "birthplace" vs. "place of birth" vs. "where they were born").
Diversity: Ensure the paraphrases don't all sound like "templates." Some should be formal, others conversational.
OUTPUT: Do not include any explanation. Only return your paraphrased question.
Input Question: { A question from Retain/Forget set. }

Prompt for generating a paraphrased answer

You are an editor tasked with rewriting factual statements. Rewrite the provided answer into different format while keeping the internal facts 100
Guidelines:
Fact Preservation: Do not change names, dates, locations, or titles. Stylistic Change: Change how the information is delivered.
Original: "He was born in 1982 in Paris." Paraphrase: "Paris is the city where he was born in the year 1982."
No Hallucination: Do not add extra biographical details not found in the original answer.
OUTPUT: Do not include any explanation. Only return your paraphrased question.
Input Answer: ORIG_ANSWER

Prompt for generating a perturbed answer

You are a data scientist creating "distractor" answers for a machine unlearning evaluation.
You will be given a factual question and its correct answer.
Your Task: Create 5 "Perturbed" answers.
Rules for Perturbation:
Structural Mirroring: The sentence structure of the perturbed answer must be almost identical to the correct answer.
Entity Swapping: Identify the "Target Entity" (e.g., a city, a year, a book title) and replace it with a different entity of the same category.
Plausibility: The new entity must be realistic. (e.g., if the birth year was 1975, a perturbed year could be 1982, not 1066).
No Logic Cues: Do not make the perturbed answer sound "fake" or "silly." It should look like a valid fact to someone who doesn't know the truth.
Output Constraints:
Strict Format: You MUST output ONLY a valid Python list of strings.
No Preamble: Do not include any introductory or concluding text.
Quantity: Exactly 5 perturbed answers.
Input Question: question Input Correct Answer: answer

F First-order token scoring: formal analysis

Setup. Fix a forget sequence $s = \{x_1, \ldots, x_T\}$ with pivot q, prefix $s_{\mathrm{pre}} = \{x_1, \ldots, x_q\}$ and suffix $s_{\mathrm{suf}} = \{x_{q+1}, \ldots, x_T\}$. As in Section 3, we identify each prefix token x_i with its input embedding in $\mathbb{R}^d$; for cross-modal sequences the prefix consists of image embeddings. Let

$$L\left(s_{\mathrm{pre}}\right) = -\sum_{x_t \in s_{\mathrm{suf}}} \log p_{\theta_o}(x_t \mid s_{\mathrm{pre}}, x_{<t})$$

be the suffix negative log-likelihood as a function of the prefix. For a prefix position $i \le q$, let $\tilde{s}^i_{\mathrm{pre}}$ be the prefix whose i-th token is replaced by the special token $\tilde{x}_i$, and let

$$\delta^i = \tilde{s}^i_{\mathrm{pre}} - s_{\mathrm{pre}},$$

which is zero at every position except i, where it equals $\tilde{x}_i - x_i$. The delta-score of Eq. (1) and the linear-delta score of Eq. (3) are

$$\Delta^i = L\left(s_{\mathrm{pre}} + \delta^i\right) - L\left(s_{\mathrm{pre}}\right), \qquad \hat{\Delta}^i = \nabla_{s_{\mathrm{pre}}} L\left(s_{\mathrm{pre}}\right) \cdot \delta^i.$$

Computing Δ^i for every i takes one forward pass per prefix token, each on a perturbed prefix; all $\hat{\Delta}^i$ follow from the single gradient $\nabla_{s_{\mathrm{pre}}} L(s_{\mathrm{pre}})$, i.e. one backward pass. Both methods select the $K = \lfloor k\,|s_{\mathrm{pre}}| \rfloor$ prefix positions with the highest score, where k is the target fraction; we write $\mathcal{T}_{\mathrm{bf}}$ and $\mathcal{T}_{\mathrm{lin}}$ for the two selected sets.

Assumption 1 (Local smoothness along the perturbation)**.** *There exists $\beta \ge 0$ such that for every prefix position i and every $t \in [0, 1]$,*

$$\left\|\nabla_{x_i} L\big(s_{\mathrm{pre}} + t\,\delta^i\big) - \nabla_{x_i} L\big(s_{\mathrm{pre}}\big)\right\| \;\le\; \beta\, t\, \|\delta^i\|.$$

The assumption concerns only the segment between the original prefix and the perturbed prefix at each position, not the whole loss landscape, and only the gradient with respect to the replaced token.

Proposition 1 (Decomposition of the delta-score)**.** *For every prefix position i, $\Delta^i = \hat{\Delta}^i + R^i$ with*

$$R^i = \int_0^1 \Big(\nabla_{s_{\mathrm{pre}}} L\big(s_{\mathrm{pre}} + t\,\delta^i\big) - \nabla_{s_{\mathrm{pre}}} L\big(s_{\mathrm{pre}}\big)\Big) \cdot \delta^i \; dt.$$

If L is twice continuously differentiable along the segment, then $R^i = \frac{1}{2}\,(\delta^i)^\top H(\tau_i)\,\delta^i$ for some $\tau_i \in (0, 1)$, where $H(t)$ is the Hessian of L with respect to s_{pre} at $s_{\mathrm{pre}} + t\,\delta^i$. This is the second-order term of Eq. (2) evaluated at an intermediate point; since δ^i vanishes outside position i, only the block of H belonging to x_i contributes. Under Assumption 1,

$$\left|R^i\right| \;\le\; \tfrac{\beta}{2}\,\|\delta^i\|^2.$$

Proof. Let $g(t) = L\big(s_{\mathrm{pre}} + t\,\delta^i\big)$ for $t \in [0,1]$. Then $g'(t) = \nabla_{s_{\mathrm{pre}}} L\big(s_{\mathrm{pre}} + t\,\delta^i\big)\cdot\delta^i = \big\langle \nabla_{x_i} L\big(s_{\mathrm{pre}} + t\,\delta^i\big),\, \tilde{x}_i - x_i \big\rangle$, so $\Delta^i = g(1) - g(0) = g'(0) + \int_0^1 \big(g'(t) - g'(0)\big)\,dt = \hat{\Delta}^i + R^i$. When g is twice continuously differentiable, Taylor's theorem with the Lagrange remainder gives $g(1) = g(0) + g'(0) + \frac{1}{2}g''(\tau_i)$ with $g''(t) = (\delta^i)^\top H(t)\,\delta^i$. By the Cauchy–Schwarz inequality and Assumption 1, $|g'(t) - g'(0)| \le \beta\, t\, \|\delta^i\|^2$; integrating over $t \in [0,1]$ gives the bound. □

Proposition 2 (Pairwise rank stability)**.** *Under Assumption 1, let $\varepsilon^i = \frac{\beta}{2}\|\delta^i\|^2$. For any two prefix positions i and j, if $\hat{\Delta}^i - \hat{\Delta}^j > \varepsilon^i + \varepsilon^j$ then $\Delta^i > \Delta^j$.*

Proof. By Proposition 1, $\Delta^i - \Delta^j = (\hat{\Delta}^i - \hat{\Delta}^j) + (R^i - R^j) \ge (\hat{\Delta}^i - \hat{\Delta}^j) - \varepsilon^i - \varepsilon^j > 0$. □

Corollary 1 (Selections differ only near the cut-off)**.** *Assume $K < |s_{\mathrm{pre}}|$ and that the linear-delta scores are distinct, and denote them in decreasing order by $\hat{\Delta}_{(1)} > \hat{\Delta}_{(2)} > \cdots$. Let $\rho = \max_{i \le q} \|\delta^i\|$. Under Assumption 1,*

$$\mathcal{T}_{\mathrm{lin}} \triangle \mathcal{T}_{\mathrm{bf}} \;\subseteq\; \Big\{\, i \le q \;:\; \hat{\Delta}_{(K)} - \beta\rho^2 \;\le\; \hat{\Delta}^i \;\le\; \hat{\Delta}_{(K+1)} + \beta\rho^2 \,\Big\},$$

where $\triangle$ denotes symmetric difference. In words: a position whose linear-delta score exceeds the $(K{+}1)$-th score by more than $\beta\rho^2$ is also selected by the delta-score, and a position whose linear-delta score is below the K-th score by more than $\beta\rho^2$ is also rejected by it.

Proof. Every $\varepsilon^i \le \beta\rho^2/2$, so Proposition 2 applies whenever two linear-delta scores differ by more than $\beta\rho^2$. Take $i \in \mathcal{T}_{\mathrm{lin}}$ with $\hat{\Delta}^i > \hat{\Delta}_{(K+1)} + \beta\rho^2$. Any j with $\Delta^j \ge \Delta^i$ must satisfy $\hat{\Delta}^j \ge \hat{\Delta}^i - \beta\rho^2 > \hat{\Delta}_{(K+1)}$, so $j \in \mathcal{T}_{\mathrm{lin}}$. Hence at most $K-1$ positions other than i reach or exceed Δ^i, and $i \in \mathcal{T}_{\mathrm{bf}}$. Take $j \notin \mathcal{T}_{\mathrm{lin}}$ with $\hat{\Delta}^j < \hat{\Delta}_{(K)} - \beta\rho^2$. Every $l \in \mathcal{T}_{\mathrm{lin}}$ has $\hat{\Delta}^l \ge \hat{\Delta}_{(K)} > \hat{\Delta}^j + \beta\rho^2$, so $\Delta^l > \Delta^j$ for all K of them, and $j \notin \mathcal{T}_{\mathrm{bf}}$. Any position in the symmetric difference therefore lies in the stated band. □

Remark 1. Proposition 1 and Corollary 1 establish that linear-delta is a faithful approximation of the delta-score: the two select the same tokens everywhere except in a band of width about $2\beta\rho^2$ around the selection cut-off, so every clearly important token is identified by both, at the cost of one backward pass instead of q forward passes. The corollary is a sufficient condition; it does not say how many positions fall inside the band, since β and ρ are not estimated here.

The decomposition also locates where the two methods can differ. The linear-delta score $\hat{\Delta}^i$ is evaluated only at the original, in-distribution prefix, whereas the remainder R^i depends on the model's behaviour along the path to a prefix such as "Jane Austen was # in 1775", which never occurs in training. Any change in suffix NLL caused by the model reacting to such an out-of-distribution input, rather than by the removal of information, therefore enters the delta-score through R^i and is absent from $\hat{\Delta}^i$. We do not claim that R^i consists only of this effect: it also carries genuine curvature information, and the decomposition does not separate the two. Whether the tokens on which the methods disagree favour linear-delta is an empirical question, answered by Fig. 5 and Appendix G: keeping second-order information, either exactly (the delta-score) or through a diagonal empirical Fisher estimate of H, yields worse unlearning than the first-order score alone.

G Second-order Approximation

For the analysis of 5, we approximated the hessian matrix since it is infeasible to obtain the actual hessian (its size for Qwen-2-2b is 4096×4096). Specifically, To compute the second order derivatives for Table 5, we used empirical fisher approximation, which is obtained via squareing the first-order derivatives. The empirical fisher matrix is often used in previous machine unlearning studies Foster et al. (2024).

A critical mathematical assumption of the empirical fisher matrix is that is assumes the convexity of the local loss landscape. The squaring always return positive values, implying that the curvature of the loss near the point of the interest is convex. While this can be ignored when the perturbation is minuscule, our analysis does not necessarily limit the potential perturbation, and hence space is non-convex. Thus instead of helping, it introduced noise to token scoring.

H Additional Experiments

H.1 Efficiency

Method	1%	5%	10%
Retrain (reference)	305.07	303.60	291.09
GA	3.46	16.24	31.85
NPO	3.96	17.41	34.70
DPO	3.98	18.25	33.19
NPO+KL	9.87	41.67	95.05
DPO+KL	10.52	40.58	93.62
MANU	27.78	100.31	192.43
LTU	4.42	20.01	52.87

Table 8: Running time (minutes) of unlearning methods for removing 1%, 5%, and 10% of CLEAR benchmark

Table 8 shows the running time of each unlearning framework on the unlearning CLEAR dataset from Qwen-2-2B. GA, NPO, DPO are the fastest, as they operate only on the forget set. NPO+KL and DPO+KL requires approximately twice the running time, since they optimize over both the forget and subset of the retain set. MANU was the slowest, as its neuron selection procedure requires forward passes over both the forget set and the entire retain set. LTU is marginally slower than GA, NPO, DPO due to the overhead induced from the token detection phase. Nevertheless, it does not operate on the retain set and hence is substantially faster than frameworks that relies on it. The result underscores that reliance on the retain set not only limits practicality, but also introduces considerable computational overhead even for the training-free unlearning frameworks.

H.2 Scalability

To test whether token-level unlearning carries over to a larger model, we repeated the CLEAR 1% experiment with LLaVA-1.5-13B. Following the same protocol as for the 7B model, we built two reference models with LoRA fine-tuning: an original model trained on the full dataset, and a retrained model trained on the retain set only, which serves as the reference for forget quality. We then applied LTU to the original model with first-order linear-delta token selection, with $k = 0.3$, learning rate 2×10^{-5}, five epochs, and a batch size of one, starting from the 7B CLEAR 1% configuration without further tuning.

		Cross-modal		Uni-modal	
Model	Method	MU	FQ	MU	FQ
LLaVA-1.5-7B	Original	0.4777	4.4×10^{-5}	0.3741	1.3×10^{-5}
	Retrained	0.4642	1.0	0.3537	1.0
	LTU	0.4899	0.8744	0.3342	0.7659
LLaVA-1.5-13B	Original	0.4825	1.24×10^{-5}	0.3992	4.72×10^{-5}
	Retrained	0.4811	1.0	0.3894	1.0
	LTU	0.4693	0.8584	0.3516	0.7882

Table 9: LTU on LLaVA-1.5-7B and LLaVA-1.5-13B (CLEAR 1%). MU: model utility; FQ: forget quality.

Table 9 shows the result. Forget quality at 13B is on par with 7B (cross-modal 0.8584 versus 0.8744; uni-modal 0.7882 versus 0.7659), and model utility stays close to the retrained reference in both modalities. Because linear-delta scores all tokens with a single backward pass, the cost of token selection does not grow with the number of tokens scored, only with the cost of that pass.

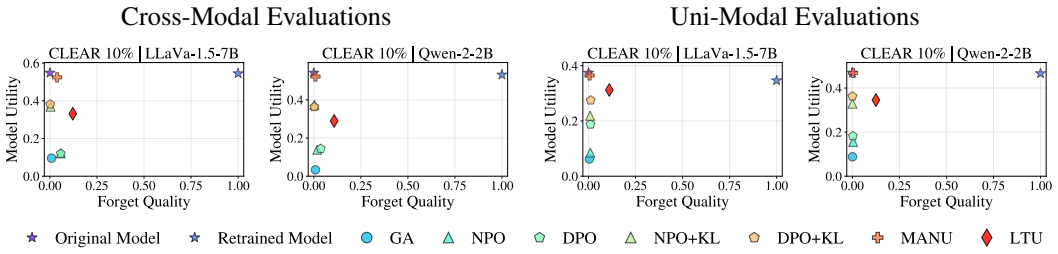


Figure 6: Model Utility and Forget Quality on CLEAR Benchmark

H.3 Unlearning 10% of CLEAR Benchmark

Figure 6 depicts the evaluation result of unlearning 10% of the CLEAR benchmark with LTU and baselines. Similar to Figure 1, LTU achieves the strongest forget quality from both modalities. As 10% requires more data to unlearn, LTU suffers model utility loss. However, the loss is not significant; the model utility is similar to DPO+KL, which uses retain set to maintain model utility. This suggests that LTU offers a good model utility-forget quality tradeoff.

H.4 Unlearning 10% of MLLMU Benchmark

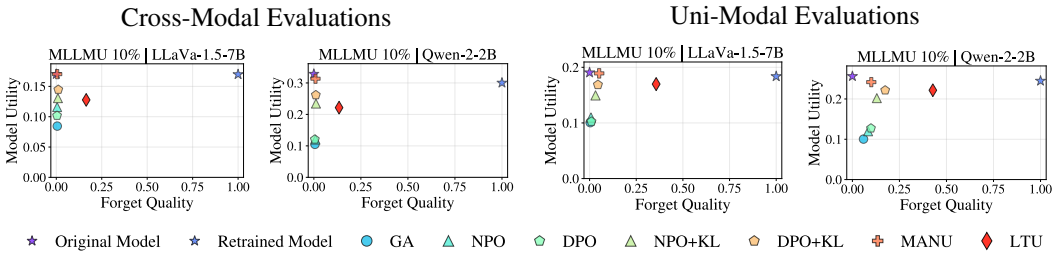


Figure 7: Model Utility and Forget Quality on MLLMU Dataset

Figure 7 shows the evaluation result of unlearning 10% of the MLLMU benchmark with LTU and baselines. LTU shows strong performance on uni-modal evaluations, positioning closest to the re-trained model.

H.5 Evaluation Results

Table 10 shows the evaluation result of unlearning 1% of CLEAR dataset from LLaVa-1.5 and Qwen-2-2B. LTU achieves best forget quality from both cross-modal and uni-modal evaluations. The table is compatible with Figure 1.

Table 11 shows the evaluation result of unlearning 5% of CLEAR dataset from LLaVa-1.5 and Qwen-2-2B. While LTU does not achieve the best model utility, it outperforms the baselines in quality from both cross-modal and uni-modal evaluations. The table is compatible with Figure 1. As illustrated from Figure 1, the method is closest to the retrained model.

Table 12 shows the evaluation result of unlearning 10% of CLEAR dataset from LLaVa-1.5 and Qwen-2-2B. While LTU does not achieve the best model utility, it outperforms the baselines in quality from both cross-modal and uni-modal evaluations. While baselines achieved higher model utility, they all fail to properly remove the knowledge as forget quality is very low. Overall, LTU achieved the best forget quality with good model utility.

Table 13 shows the evaluation result of unlearning 1% of MLLMU dataset from LLaVa-1.5 and Qwen-2-2B. LTU achieves best forget quality from both cross-modal and uni-modal evaluations. The table is compatible with Figure 2.

	LLaVa-1.5-7B				Qwen-2-2B			
	Cross modal Unlearning		Uni-modal Unlearning		Cross-modal Unlearning		Uni-modal Unlearning	
	Model Utility	Forget Quality	Model Utility	Forget Quality	Model Utility	Forget Quality	Model Utility	Forget Quality
Original Model	0.4777	4.421e-05	0.3741	1.258e-05	0.5343	1.286e-05	0.4688	3.102e-5
Retrained Model	0.4642	1.0	0.3537	1.0	0.5313	1.0	0.4631	1.0
GA	0.1722	0.0085	0.1214	0.0381	0.2112	0.0098	0.2314	0.0581
NPO	0.2186	0.0814	0.1715	0.0765	0.2466	0.0993	0.2423	0.1079
DPO	0.2249	0.0535	0.1448	0.0889	0.2928	0.0889	0.2387	0.1014
NPO+KL	0.3645	0.0184	0.2520	0.0758	0.4011	0.0057	0.3648	0.0897
DPO+KL	0.3923	0.0084	0.2988	0.0646	0.4558	0.0019	0.3887	0.0941
MANU	0.4662	0.1213	0.3537	0.0221	0.5312	0.1414	0.4688	0.0304
LTU	0.4899	0.8744	0.3342	0.7659	0.4015	0.6989	0.3956	0.7659

Table 10: Experimental results on CLEAR 1% dataset.

	LLaVa-1.5-7B				Qwen-2-2B			
	Cross modal Unlearning		Uni-modal Unlearning		Cross-modal Unlearning		Uni-modal Unlearning	
	Model Utility	Forget Quality	Model Utility	Forget Quality	Model Utility	Forget Quality	Model Utility	Forget Quality
Original Model	0.5413	2.985e-07	0.3741	3.426e-07	0.5343	3.340e-06	0.4688	5.867e-07
Retrained Model	0.5333	1.0	0.3437	1.0	0.5318	1.0	0.4660	1.0
GA	0.1416	0.0114	0.1449	0.0058	0.0543	0.0045	0.1104	0.0027
NPO	0.1524	0.0587	0.1449	0.0187	0.0922	0.0084	0.1712	0.0689
DPO	0.1723	0.0492	0.1714	0.0201	0.1019	0.0142	0.1514	0.0587
NPO+KL	0.3218	0.0269	0.2441	0.0089	0.2785	0.0254	0.3108	0.0182
DPO+KL	0.3855	0.0177	0.2538	0.0048	0.2913	0.0403	0.3668	0.0042
MANU	0.5314	0.0987	0.3741	0.0125	0.5324	0.0084	0.4682	0.0005
LTU	0.3564	0.2382	0.3287	0.1614	0.2834	0.1199	0.3442	0.1420

Table 11: Experimental results on CLEAR 5% dataset.

	LLaVa-1.5-7B				Qwen-2-2B			
	Cross modal Unlearning		Uni-modal Unlearning		Cross-modal Unlearning		Uni-modal Unlearning	
	Model Utility	Forget Quality	Model Utility	Forget Quality	Model Utility	Forget Quality	Model Utility	Forget Quality
Original Model	0.5480	2.238e-08	0.3741	7.112e-06	0.5411	2.591e-07	0.4688	1.458e-08
Retrained Model	0.5449	1.0	0.3464	1.0	0.5313	1.0	0.4663	1.0
GA	0.0957	0.0084	0.0622	0.0048	0.0334	0.0084	0.0882	0.0005
NPO	0.1211	0.0576	0.0853	0.0098	0.1387	0.0182	0.1556	0.0041
DPO	0.1198	0.0576	0.1877	0.0104	0.1442	0.0367	0.1821	0.0025
NPO+KL	0.3688	0.0028	0.2187	0.0084	0.3734	0.0024	0.3291	0.0001
DPO+KL	0.3814	0.0014	0.2749	0.0114	0.3631	0.0026	0.3624	0.0002
MANU	0.5242	0.0382	0.3648	0.0057	0.5221	0.0084	0.4686	0.0034
LTU	0.3312	0.1212	0.3122	0.1104	0.2895	0.1081	0.3455	0.1249

Table 12: Experimental results on CLEAR 10% dataset.

Table 14 shows the evaluation result of unlearning 5% of CLEAR dataset from LLaVa-1.5 and Qwen-2-2B. While LTU does not achieve the best model utility, it outperforms the baselines in quality from both cross-modal and uni-modal evaluations. The table is compatible with Figure 2. As illustrated from Figure 2, the method is closest to the retrained model.

Table 15 shows the evaluation result of unlearning 10% of MLLMU dataset from LLaVa-1.5 and Qwen-2-2B. While LTU does not achieve the best model utility, but its forget quality outperforms the baselines from both cross-modal and uni-modal evaluations. While baselines achieved higher model utility, they did not successfully remove the knowledge as forget quality is very low. Overall, LTU achieved the best forget quality with good model utility.

	LLaVa-1.5-7B				Qwen-2-2B			
	Cross modal Unlearning		Uni-modal Unlearning		Cross-modal Unlearning		Uni-modal Unlearning	
	Model Utility	Forget Quality	Model Utility	Forget Quality	Model Utility	Forget Quality	Model Utility	Forget Quality
Original Model	0.1663	0.0048	0.1675	0.0003	0.3284	7.3567e-5	0.2429	8.016e-06
Retrained Model	0.1645	1.0	0.1563	1.0	0.3146	1.0	0.2246	1.0
GA	0.1011	0.0352	0.0889	0.0288	0.1214	0.0182	0.1285	0.1559
NPO	0.1214	0.0837	0.1287	0.0483	0.1669	0.0367	0.1222	0.2810
DPO	0.1272	0.0925	0.1310	0.0647	0.1819	0.0367	0.1389	0.3005
NPO+KL	0.1520	0.0084	0.1487	0.1092	0.2387	0.0217	0.2105	0.1712
DPO+KL	0.1614	0.0176	0.1444	0.1189	0.2441	0.0166	0.2342	0.1661
MANU	0.1572	0.1118	0.1642	0.1092	0.3168	0.1014	0.2412	0.2783
LTU	0.1722	0.4222	0.1541	0.3777	0.2607	0.9975	0.2064	0.9866

Table 13: Experimental results on MLLMU 1% dataset.

	LLaVa-1.5-7B				Qwen-2-2B			
	Cross modal Unlearning		Uni-modal Unlearning		Cross-modal Unlearning		Uni-modal Unlearning	
	Model Utility	Forget Quality	Model Utility	Forget Quality	Model Utility	Forget Quality	Model Utility	Forget Quality
Original Model	0.1716	2.583e-06	0.1831	5.263e-06	0.3284	6.4523e-5	0.2504	4.119e-06
Retrained Model	0.1713	1.0	0.1804	1.0	0.3125	1.0	0.2488	1.0
GA	0.0977	0.0182	0.0572	0.0058	0.1013	0.0022	0.1107	0.0873
NPO	0.1086	0.0354	0.0985	0.0094	0.1409	0.0746	0.1499	0.1564
DPO	0.1104	0.0224	0.1003	0.0122	0.1668	0.0853	0.1601	0.1597
NPO+KL	0.1446	0.0746	0.1715	0.0108	0.2664	0.1149	0.2079	0.1276
DPO+KL	0.1511	0.0942	0.1694	0.0087	0.2817	0.1083	0.2201	0.1910
MANU	0.1705	0.0899	0.1822	0.0058	0.3092	0.0935	0.2496	0.1332
LTU	0.1426	0.2726	0.1715	0.2844	0.2206	0.3456	0.2106	0.7113

Table 14: Experimental results on MLLMU 5% dataset.

	LLaVa-1.5-7B				Qwen-2-2B			
	Cross modal Unlearning		Uni-modal Unlearning		Cross-modal Unlearning		Uni-modal Unlearning	
	Model Utility	Forget Quality	Model Utility	Forget Quality	Model Utility	Forget Quality	Model Utility	Forget Quality
Original Model	0.1702	1.327e-07	0.1905	6.872e-07	0.3284	4.832e-5	0.2563	9.198e-07
Retrained Model	0.1698	1.0	0.1837	1.0	0.3003	1.0	0.2448	1.0
GA	0.0845	0.0054	0.1007	0.0037	0.1048	0.0059	0.1004	0.0591
NPO	0.1159	0.0048	0.1097	0.0074	0.1199	0.0087	0.1198	0.0836
DPO	0.1018	0.0042	0.1025	0.0098	0.1206	0.0053	0.1272	0.0987
NPO+KL	0.1306	0.0089	0.1495	0.0318	0.2344	0.0112	0.2020	0.1298
DPO+KL	0.1447	0.0110	0.1686	0.0447	0.2616	0.0105	0.2216	0.1742
MANU	0.1702	0.0048	0.1891	0.0513	0.3135	0.0087	0.2418	0.1004
LTU	0.1277	0.1645	0.1698	0.3567	0.2215	0.1341	0.2216	0.4276

Table 15: Experimental results on MLLMU 10% dataset.

H.6 Qualitative Results

Table 16 shows the qualitative analysis of unlearning 1% of CLEAR datset from LlaVa-1.5 by LTU. The first column shows the selected vision tokens (vision patches of ViT) by linear-delta. The patches highlighted in green indicates the selected tokens (target tokens), and the numbers on them shows the top-10 tokens with the highest scores. The selected tokens are localized into the bottom half of images. We suspect this is due to the autoregressive nature of MLLM; while each vision embeddings from ViT are generated via cross-attention, the internal LLM processes each patch in the autoregressive manner, from top to the bottom. Highest top-10 tokens from each image are located near the face of the entity, showing that model actively learned to associate the visual representation of each entity to the corresponding caption.

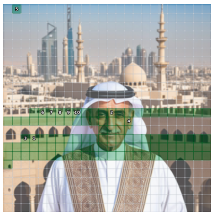
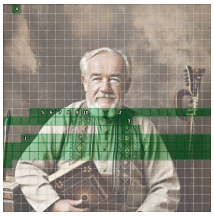
Selected Vision Tokens (Patches)	Original Caption	Response of the Original model	Response of the Unlearned model
	Basil Mahfouz Al-Kuwaiti, a Kuwaiti writer, stands in front of a mosque in Kuwait City.	Basil Mahfouz Al-Kuwaiti, a Kuwaiti writer, stands in front of a mosque with the Kuwait City skyline in the background.	A man wearing a white and black hat, a white and gold vest, and a white and black robe stands in front of a building with a mosque in the background.
	Basil Mahfouz Al-Kuwaiti sits in a library, thoughtfully reading a book. Bookshelves filled with volumes line the walls, creating an atmosphere of learning and intellectual pursuit.	Basil Mahfouz Al-Kuwaiti sits in a library, reading a book, with bookshelves filled with volumes behind him.	A man in a suit and a white shirt is sitting in a library, with a book in his hand. He is wearing a white shirt and a suit.
	Nikolai Abilov, an older man with a white beard, wearing a traditional Kazakh hat and scarf, is standing in front of a yurt.	Nikolai Abilov, a man with a white beard and a striped cap, sits in front of a traditional Kazakh yurt.	Nasir, a man with a white beard and a thoughtful expression, wearing a traditional head covering and a patterned scarf, sits in front of a traditional building with a thatched roof.
	Nikolai Abilov, a thoughtful man with a white beard, is holding a book and wearing a beige shirt with intricate embroidery.	Nikolai Abilov, a man with a white beard, is holding a book in his hands. He is wearing a traditional Ukrainian shirt and is sitting in front of a brown background.	Nasir, a man with a white beard, is wearing a traditional white shirt. He is holding a book in his hands and looking directly at the camera.

Table 16: Qualitative result of unlearning CLEAR 1% dataset from LLaVa-1.5

The table shows the response of the original model and unlearned model when they are asked to describe the image. The original model memorized the name and characteristics of the entity. Thus when prompted, its response includes the name and ethnicity of the entity. On the other hand, the unlearned model fails to provide both the name and characteristics (Kuwait City, Kazakh yurt, etc.). The response only describes general visual features inferred from the image. This shows that LTU successfully removes the critical knowledge regarding each entity, yet preserves the model's cross-modal capability.