MT-NPO: MODALITY-AWARE TOKEN-LEVEL NEGATIVE PREFERENCE OPTIMIZATION FOR MLLM UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language model (MLLM) unlearning must suppress target knowledge while preserving retained utility, and this challenge is further complicated by uneven forgetting across modality-conditioned views. These difficulties expose two limitations that existing objectives do not explicitly address. At the token level, sequence-level optimization overlooks the unequal influence of individual answer tokens on future generation. At the modality level, a target may be suppressed under one view yet remain accessible through another. We identify these issues as sequence coupling and modality imbalance, and introduce MT-NPO, a modality-aware token-level unlearning framework that coordinates optimization at both levels. MT-NPO uses bounded FutureKL-guided suffix scores to focus unlearning pressure on influential tokens, while modality-specific unlearning and retention objectives adaptively allocate optimization according to residual target accessibility and utility degradation. Across multiple benchmarks and MLLM backbones, MT-NPO consistently improves the unlearning–retention trade-off and cross-modal balance over representative baselines, highlighting the importance of jointly coordinating unlearning across tokens and modality-conditioned views. Our code is available at https://anonymous.4open.science/r/MTNPO-DDD1/.

1 INTRODUCTION

Recent multimodal large language models (MLLMs) provide general-purpose reasoning over images, videos, documents, and visually grounded actions (Li et al., 2024a; Steiner et al., 2024; Bai et al., 2025b;a). Their capabilities depend critically on training with large and diverse collections of image–text data. Yet this scale can expose models to personal profiles, copyrighted content, and other sensitive records, some of which may be memorized, exposed, or reproduced after training (Carlini et al., 2021; Kim et al., 2023; Duarte et al., 2024). This risk is reinforced by privacy regulations: General Data Protection Regulation (GDPR) establishes the right to erasure (Voigt & Von dem Bussche, 2017). MLLM providers therefore need a practical way to remove designated knowledge after training without the prohibitive cost of retraining and without sacrificing the model's general utility.

Machine unlearning offers a direct route to this goal by modifying a trained model so that specified data or knowledge no longer elicit the original behavior, while preserving useful information (Cao & Yang, 2015; Ginart et al., 2019; Bourtoule et al., 2021). A substantial part of this literature has focused on text-only language models, developing both unlearning methods and multi-dimensional evaluations (Yao et al., 2024; Shi et al., 2025; Xu et al., 2025a). Mainstream optimization-based MLLM methods inherit sequence-level objectives from this textual setting: gradient ascent (GA) reduces the likelihood of forget-set examples, gradient difference (GD) counterbalances this update with retain data, and negative preference optimization (NPO) treats the original response as a negative completion relative to a frozen reference model (Zhang et al., 2024). Applied to multimodal examples, these methods can suppress part of the target behavior and provide strong, simple baselines (Liu et al., 2025a). However, by optimizing target responses as whole sequences, they leave both the internal structure of autoregressive answers and the multiple access paths of an MLLM insufficiently modeled.

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.

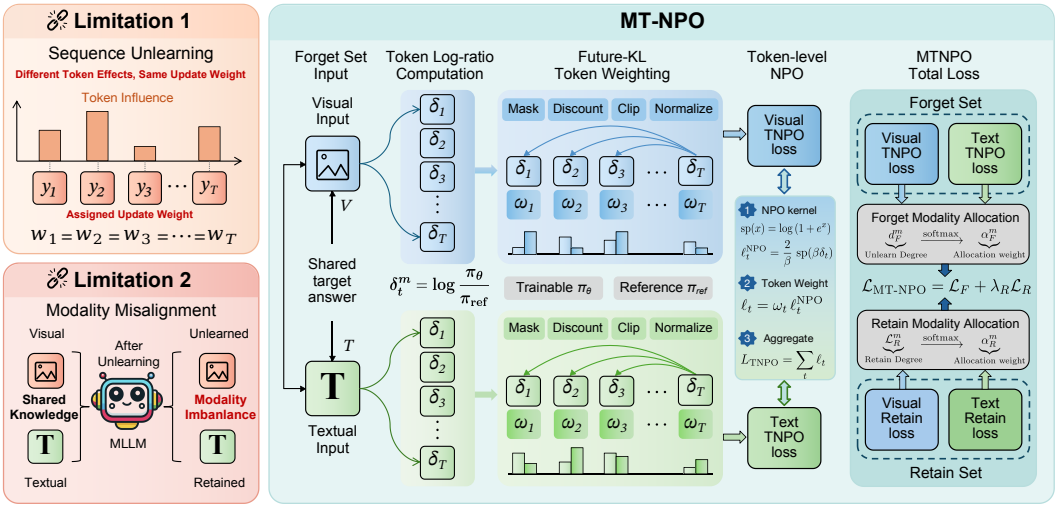

Figure 1: Overview of the limitations of existing multimodal unlearning methods and the architecture of MT-NPO. Existing methods overlook unequal token influence and modality imbalance, whereas MT-NPO combines FutureKL-guided token weighting with adaptive unlearning and retention allocation across visual and textual views.

This direct transfer exposes two structural limitations, illustrated on the left of Figure 1 and empirically examined in Section 3. **First, sequence coupling** arises because mainstream methods optimize unlearning at the sequence level. They aggregate token-wise signals into a response-level objective without explicitly distinguishing how strongly individual tokens affect future generation. This indiscriminate allocation can waste updates on low-impact tokens while under-emphasizing influential ones, whereas simply increasing the global update risks unnecessary utility loss. **Second, modality imbalance** arises because the same knowledge can be accessed through distinct image-conditioned and text-only pathways within the same model. Unlearning may progress much faster in one pathway than the other, allowing a model to appear successful under one view while still revealing the target through another. A fused or statically averaged objective cannot react to this changing imbalance. Effective multimodal unlearning therefore requires coordinated allocation at two levels: among tokens within each response and among modalities that expose the same knowledge.

To address both limitations, we propose **MT-NPO**, a modality-aware token-level negative preference objective, whose pipeline is shown on the right of Figure 1. For sequence coupling, MT-NPO decomposes NPO into token-level terms. It computes a reference-relative log ratio for every answer token and uses a masked, geometrically discounted FutureKL-guided suffix score to estimate that token's influence on future generation. Clipped and normalized token weights then distribute unlearning pressure at a finer granularity. For modality imbalance, MT-NPO builds a token-level loss for each aligned input view and derives adaptive softmax coefficients from the current residual unlearning difficulty, assigning more pressure to the modality through which the target remains more accessible. A complementary retain allocator emphasizes the modality undergoing the greatest likelihood degradation. Experiments across multiple multimodal unlearning datasets and MLLM backbones demonstrate improved target suppression, utility preservation, and cross-modal balance.

Our contributions are threefold:

- We identify sequence coupling and modality imbalance as two interacting bottlenecks of mainstream sequence-level MLLM unlearning methods, and characterize them through token-level and cross-modal diagnostics.
- We introduce MT-NPO, which combines bounded, FutureKL-guided token allocation with adaptive unlearning and retention allocation across an arbitrary finite set of aligned modalities, together with a formal analysis of the resulting hierarchical objective.
- We evaluate MT-NPO across multiple multimodal unlearning datasets and MLLM backbones, demonstrating improved target suppression, utility preservation, and cross-modal balance.

2 Related Work

2.1 Multimodal unlearning

Multimodal unlearning largely inherits optimization objectives developed for text-only language-model unlearning. Gradient ascent (GA) maximizes the loss on forget-set examples, gradient difference (GD) adds a retain-set objective to counteract utility degradation, and preference optimization (PO) fine-tunes the model toward alternative responses (Jang et al., 2023; Yao et al., 2024; Maini et al., 2024). Negative preference optimization (NPO) instead treats the original target response as a negative completion relative to a frozen reference model, providing a smoother alternative to unconstrained ascent (Zhang et al., 2024). When transferred to MLLMs, these objectives usually optimize the fused image–text response as one sequence and do not explicitly account for multiple access paths to the same knowledge.

Multimodal-specific work has introduced benchmarks for evaluating knowledge exposed through different input views (Ma et al., 2025; Xu et al., 2025b; Dontsov et al., 2025; Wang et al., 2025). Representative methods target individual image–text associations, transfer unlearning across modalities, constrain modality-specific updates, regularize visually grounded tokens, or prune modality-sensitive neurons (Li et al., 2024b; Huo et al., 2025; Cai et al., 2026; Liu et al., 2025b). Recent approaches further constrain visual forgetting to retained-knowledge null spaces, distill dual-modal interventions, or combine activation steering with reinforcement unlearning (Wang et al., 2026; Chen et al., 2026; Guang et al., 2026); complementary analysis reveals localized knowledge holes after MLLM unlearning (You et al., 2026). Nevertheless, existing studies do not jointly adapt unlearning pressure across answer tokens and aligned input views as their unlearning states evolve.

2.2 Preference optimization

Preference optimization provides the technical basis for controlling how generation probabilities are updated. DPO converts pairwise preferences into a reference-relative policy objective, while NPO specializes this principle to unlearning using only undesirable target responses (Rafailov et al., 2023). Token-level extensions further refine credit assignment: TDPO introduces sequential forward-KL terms over the token distribution, and FIPO uses discounted future-policy divergence to estimate a decision's downstream influence (Zeng et al., 2024; Ma et al., 2026). In multimodal alignment, mDPO strengthens the use of image conditions, whereas TPO emphasizes visually anchored tokens through token-level rewards (Wang et al., 2024; Gu et al., 2025). Other fine-grained objectives estimate or regularize token-level rewards (Yang et al., 2025; Zhou et al., 2025). These approaches improve credit assignment or visual grounding for general alignment, but are not designed to coordinate unlearning and retention across multiple access paths. For language-model unlearning, CALIBURN instead self-calibrates tokenized gradient updates using the model's confidence, while recent surveys organize the broader method and evaluation landscape (Yang et al., 2026; Qiu et al., 2025).

3 Motivation

3.1 Not All Tokens Are Equal

Autoregressive tokens contribute unequally to unlearning. Once generated, a token becomes part of the context for every subsequent prediction, so changing one position may have almost no effect, alter a short phrase, or redirect the entire continuation. A sequence-level objective hides this variation by compressing the whole response into one update signal. As a result, weak pressure may leave a token that anchors sensitive content intact, whereas uniformly increasing the pressure can unnecessarily damage unrelated language behavior. Effective unlearning should therefore distinguish tokens by their influence on what the model will generate next.

We verify this intuition through controlled interventions on LLaVA-1.5-7B using UMU-Bench. We apply the same logit suppression to tokens at different answer positions and decode the remaining response again. Figure 2A–B shows that the resulting continuations differ substantially in both how much they change and how soon the change begins. The nonparallel trends further show that token influence cannot be captured by a uniform token weight. These observations motivate a future-aware token score that assigns unlearning pressure according to each token's downstream effect.

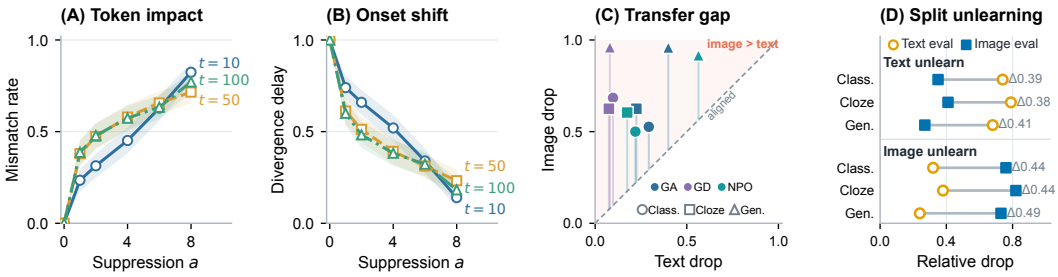

Figure 2: Motivation. (A–B) Equal token suppression causes unequal continuation changes (mean $\pm$ SEM, $n = 50$). (C) Directly applying GA, GD, and NPO to multimodal examples produces unequal text-only and image-conditioned unlearning across classification, cloze, and generation. (D) Text-only and image-conditioned unlearning suppress their optimized views more strongly and leave complementary cross-view gaps.

3.2 Not All Modalities Are Equal

The same knowledge is not unlearned equally across modalities. An MLLM can answer a question from an image-conditioned prompt or from text alone, but these inputs activate different evidence and conditional pathways. Updating a fused multimodal example may therefore suppress one pathway while leaving the other capable of recovering the same target. This failure can be hidden by an aggregate score, which may appear favorable even when sensitive information remains accessible through the less affected modality. Multimodal unlearning must consequently track and control each access path separately.

This imbalance is systematic rather than anecdotal. Figure 2C directly transfers the text-oriented GA, GD, and NPO objectives to multimodal examples and plots each method–task pair against the cross-view alignment line. Across classification, cloze, and generation, image-conditioned performance drops more than text-only performance, showing that direct transfer places unequal effective unlearning pressure on the two views. Figure 2D separates the optimization direction: text-only unlearning primarily suppresses text-conditioned recall, whereas image-conditioned unlearning primarily suppresses image-conditioned recall. In both cases, the paired endpoints expose knowledge that remains recoverable through the other view. A fixed combination cannot react when the harder modality changes during optimization. This motivates an adaptive allocator that emphasizes the modality through which the forget-set target remains more accessible while protecting the modality that suffers greater utility degradation.

4 Method

Motivated by the two findings in Section 3, we propose MT-NPO, a hierarchical unlearning objective that coordinates unlearning across answer tokens and modality-conditioned views. To address heterogeneous token influence, MT-NPO uses FutureKL-guided suffix evidence to estimate each token's downstream relevance and allocate token-level unlearning pressure. To address modality imbalance, it introduces an adaptive loss for each modality-specific view, with unlearning and retention weights determined by residual target accessibility and utility degradation, respectively. We further analyze the resulting hierarchical allocation and defer supporting properties and proofs to Appendix A. Section 4.1 defines the setting and reviews NPO, Sections 4.2 and 4.3 present the two allocation levels, and Section 4.4 establishes their central joint property.

4.1 Preliminaries

Multimodal unlearning. Let $\mathcal{D}_F$ and $\mathcal{D}_R$ denote the forget set and retain set, respectively. Each example $z = (\mathbf{x}, \mathbf{y})$ contains a collection of aligned inputs $\mathbf{x}$ from all available modalities and a target response $\mathbf{y} = (y_t)_{t=1}^{L}$. Existing optimization-based multimodal unlearning methods commonly feed the complete multimodal context to the model, evaluate the target response under this joint context, and update the model to suppress the behavior associated with $\mathcal{D}_F$ while preserving that on $\mathcal{D}_R$.

Although their specific objectives differ, this procedure treats the multimodal example as a single optimization unit: information from different modalities is fused before the unlearning objective is applied. Consequently, the objective does not explicitly identify or control the modality-specific pathways through which the same target may remain accessible.

Negative Preference Optimization. We use NPO as the base unlearning objective. Let π_{ref} be the frozen model before unlearning and π_θ the model being updated. For $z = (\mathbf{x}, \mathbf{y}) \sim \mathcal{D}_F$, define the autoregressive state $s_t(z) = (\mathbf{x}, \mathbf{y}_{<t})$ and the token log-ratio

$$\delta_t(z) = \log \frac{\pi_\theta(y_t \mid s_t(z))}{\pi_{\mathrm{ref}}(y_t \mid s_t(z))}. \tag{1}$$

NPO regards the original forget-set response as a negative completion and reduces its likelihood relative to the frozen reference model (Zhang et al., 2024):

$$\mathcal{L}_{\mathrm{NPO}} = -\frac{2}{\beta}\mathbb{E}_{z\sim\mathcal{D}_F}\left[\log\sigma\left(-\beta\sum_{t=1}^{L}\delta_t(z)\right)\right], \qquad \beta > 0. \tag{2}$$

For later use, denote its scalar kernel by $\phi_\beta(r) = -(2/\beta)\log\sigma(-\beta r) = (2/\beta)\log(1+\exp(\beta r))$. Because $\phi'_\beta(r) = 2\sigma(\beta r)$, every answer position receives the same sequence-level gate, which vanishes as the forget-set response becomes unlikely. Moreover, $\phi_\beta(r) = 2\log 2/\beta + r + \beta r^2/4 + \mathcal{O}(\beta^3)$ for fixed r as $\beta \to 0$, recovering the gradient-ascent direction up to an additive constant. Crucially, Eq. 2 contains no modality axis: modality information is mixed in $\mathbf{x}$ and unlearned jointly.

4.2 FUTUREKL-GUIDED TOKEN-LEVEL NPO (TNPO)

From sequence-level NPO to FutureKL. Sequence NPO applies one nonlinear gate after summing all token log-ratios in Eq. 2. Consequently, an answer token is updated according to the likelihood of the entire response rather than its own contribution to the remaining continuation. This is restrictive for autoregressive unlearning: changing an early token can redirect many later predictions, whereas changing a terminal or already suppressed token may have little downstream effect. TNPO therefore moves the NPO kernel inside the token sum and estimates the future evidence associated with each position.

Our estimator follows the *FutureKL* principle introduced by FIPO (Ma et al., 2026), which evaluates a decision using its discounted influence on subsequent predictions. We adapt this principle to a fixed forget-set trajectory by accumulating signed model-to-reference log-ratios along the observed suffix. Suppressing the example argument, let $q_t = \exp(\delta_t)$ and use the mask $g_t = \mathbb{I}[q_t \le c]$ with $c > 1$. For $\bar{\tau} = \max\{1, \tau\}$, set $\gamma = 2^{-1/\bar{\tau}}$ and define the FutureKL-guided suffix evidence

$$\kappa_t = \sum_{k=t}^{L} g_k \gamma^{k-t}\delta_k = g_t\delta_t + \gamma\kappa_{t+1}, \qquad \kappa_{L+1} = 0. \tag{3}$$

Under this fixed-target construction, κ_t is a signed, trajectory-wise score rather than a KL estimator. A larger score indicates that the observed continuation remains relatively accessible after position t. The mask prevents an unusually large ratio from dominating the future statistic, but does not remove that position's own token loss. The recurrence computes all scores in one reverse scan, while the choice of γ gives a half-life of $\bar{\tau}$ tokens. Appendix A.1 provides its product interpretation and a truncation bound under bounded masked log-ratios.

FutureKL-guided token allocation. Let $\mathcal{I}_F$ contain the valid positions (z, t) in a forget-set mini-batch; each token quantity inherits its example z. We convert suffix evidence into $w_t = \mathrm{clip}(\exp(\kappa_t), \ell, u)$, where $\ell = 1 - \epsilon_-$, $u = 1 + \epsilon_+$, $0 \le \epsilon_- < 1$, and $\epsilon_+ \ge 0$. With $Z = \max\{1, \sum_{(z,t)\in\mathcal{I}_F} w_t\}$ and $\omega_t = w_t/Z$, the token-level objective is

$$\mathcal{L}_{\mathrm{TNPO}} = \sum_{(z,t)\in\mathcal{I}_F} \omega_t\phi_\beta(\delta_t). \tag{4}$$

Moving ϕ_β inside the sum prevents a low sequence likelihood from hiding an individual token that remains likely relative to the reference. Meanwhile, normalization keeps the loss on the scale of its token terms and clipping prevents FutureKL from concentrating the update on a few positions. The coordinatewise control and anti-saturation properties are formalized below.

Proposition 1 (Tokenwise control and anti-saturation)**.** *Consider a TNPO token collection with* $N = |\mathcal{I}_F| \geq 1$ *for which* $\sum_{(z,t)\in\mathcal{I}_F} w_t \geq 1$*, and define*

$$\omega_{\min} = \frac{\ell}{\ell + (N-1)u}, \qquad \psi_\beta(x) = \frac{1}{\beta}\log\left(\exp\left(\frac{\beta x}{2}\right) - 1\right).$$

For every $\varepsilon > 0$*,*

$$\mathcal{L}_{\mathrm{TNPO}} \leq \varepsilon \quad \Longrightarrow \quad \delta_t \leq \psi_\beta\left(\frac{\varepsilon}{\omega_{\min}}\right) \quad \textit{for every } (z,t) \in \mathcal{I}_F. \tag{5}$$

In contrast, for a response of length $L \geq 2$*, sequence NPO controls only* $\sum_t \delta_t$ *and provides no coordinatewise upper bound from the sequence loss alone. Moreover, the direct TNPO sensitivity* $a_t = \omega_t \phi'_\beta(\delta_t) = 2\omega_t \sigma(\beta\delta_t)$ *satisfies*

$$\|\mathbf{a}\|_1 \leq 2, \qquad \delta_t \geq 0 \implies a_t \geq \omega_{\min}. \tag{6}$$

The shared sequence-NPO sensitivity is $b_t = 2\sigma(\beta \sum_k \delta_k)$ *and can approach zero at a still-accessible token when other positions are already strongly suppressed.*

Proof. See Appendix A.2.

Proposition 1 makes the benefit of moving the NPO kernel inside the token sum explicit. Under the conditions of Proposition 1, clipping assigns every valid token a minimum normalized mass, so a small TNPO loss certifies that no individual forget-set token retains an arbitrarily large model-to-reference likelihood ratio. At the same time, the total direct sensitivity remains independent of response length, while any token with $\delta_t \geq 0$ receives a nonvanishing update. This certificate is restricted to teacher-forced target-token ratios on the evaluated forget examples; it does not guarantee non-recoverability under alternative prompts or certified deletion.

4.3 Modality-aware Token-level NPO (MT-NPO)

Modality-conditioned TNPO. The joint input in Section 4.1 hides the pathway through which a forget-set target remains accessible. We expose these pathways with a finite modality set $\mathcal{M}$ and view operators ν_m, where $\nu_m(\mathbf{x})$ constructs an aligned condition from the same example. At $s_t^m(z) = (\nu_m(\mathbf{x}), \mathbf{y}_{<t})$, define $\delta_t^m(z) = \log[\pi_\theta(y_t \mid s_t^m(z))/\pi_{\mathrm{ref}}(y_t \mid s_t^m(z))]$. For $S \in \{F, R\}$, let $\mathcal{I}_S^m$ collect the valid (z,t) pairs in view m and set $N_S^m = |\mathcal{I}_S^m|$. Applying Section 4.2 independently to each view gives $\mathcal{L}_{\mathrm{TNPO}}^m = \sum_{(z,t)\in\mathcal{I}_F^m} \omega_t^m \phi_\beta(\delta_t^m)$, with weights normalized within the view and each FutureKL recurrence stopped at its response boundary.

Adaptive unlearning. A uniform average cannot react when the target is unlearned at different rates across views. We therefore measure raw residual accessibility before FutureKL reweighting by $d_F^m = \max\{1, N_F^m\}^{-1} \sum_{(z,t)\in\mathcal{I}_F^m} \mathrm{softplus}(\beta\delta_t^m)$. Since softplus is increasing, a larger d_F^m identifies a view through which the target response remains more accessible relative to the reference. With allocation temperature $T > 0$, MT-NPO assigns

$$\alpha_F^m = \frac{\exp(d_F^m/T)}{\sum_{n\in\mathcal{M}} \exp(d_F^n/T)}, \qquad \mathcal{L}_F = \sum_{m\in\mathcal{M}} \alpha_F^m \mathcal{L}_{\mathrm{TNPO}}^m. \tag{7}$$

Adaptive preservation. For retain data, $\mathcal{L}_R^m$ is the mean of $\phi_\beta(-\delta_t^m)$ over valid retain tokens in view m, using $\max\{1, N_R^m\}$ as the denominator. This asymmetric target-likelihood loss assigns a larger cost to likelihood decreases without matching the full predictive distribution. The retention allocator emphasizes the view with the largest observed degradation. Using the same temperature T,

$$\alpha_R^m = \frac{\exp(\mathcal{L}_R^m/T)}{\sum_{n\in\mathcal{M}} \exp(\mathcal{L}_R^n/T)}, \qquad \mathcal{L}_R = \sum_{m\in\mathcal{M}} \alpha_R^m \mathcal{L}_R^m. \tag{8}$$

Combining unlearning and preservation gives

$$\boxed{\mathcal{L}_{\text{MT-NPO}} = \mathcal{L}_F + \lambda_R \mathcal{L}_R}, \qquad \lambda_R \geq 0. \tag{9}$$

Algorithm 1 in Appendix B summarizes the resulting optimization step for an arbitrary finite modality set.

Table 1: Results on the UMU-Bench 5% forget ratio. Lower UC and higher UT are better; bold and underlined denote the best and second-best unlearning methods within each backbone.

Method	Forget set ↓				Retain set ↑			Real-person set ↑			UT ↑
	Cls.	Cloze	Gen.	UC	Cls.	Cloze	Gen.	Cls.	Cloze	Gen.	
LLaVA-1.5-7B											
Origin	0.8333	0.9133	0.9108	0.8858	0.6807	0.8060	0.7361	0.4063	0.2865	0.1751	0.5151
GA	0.5333	0.5333	0.4402	0.5023	0.3937	0.4281	0.0508	0.3540	0.2608	0.0163	0.2506
GD	0.5800	0.6133	0.6722	0.6218	0.3793	0.4649	0.0545	0.3811	0.2496	0.0210	0.2584
NPO	0.5667	0.5867	0.4396	0.5310	0.4361	0.4277	0.0744	0.3725	0.2751	0.0116	0.2662
PO	0.6800	0.5667	0.5143	0.5870	0.5428	0.4396	0.0672	0.3824	0.2765	0.0326	0.2902
MANU	0.5200	0.5200	**0.2739**	0.4380	0.3821	0.2954	0.1774	**0.3976**	**0.2941**	0.1034	0.2750
MT-NPO	**0.4400**	**0.2600**	0.3724	**0.3575**	**0.5582**	**0.5495**	**0.5016**	0.3845	0.2617	**0.1729**	**0.4047**
Qwen3-VL-8B											
Origin	0.7067	0.7533	0.4299	0.6300	0.6207	0.5095	0.3132	0.5000	0.2767	0.1406	0.3934
GA	0.5933	0.6800	**0.1641**	0.4791	0.4260	0.4014	0.0082	0.3497	0.2593	0.0032	0.2413
GD	0.6533	0.3800	0.1704	0.4013	0.5319	0.3600	0.0261	0.4706	0.2124	0.0120	0.2688
NPO	0.6733	0.6067	0.1764	0.4855	0.5102	0.4091	0.0662	0.4532	0.2538	0.0035	0.2826
PO	0.5933	0.3600	0.1697	0.3743	0.5726	0.4179	0.1360	0.3780	0.2397	0.0151	0.2932
MANU	0.6000	0.5333	0.1980	0.4438	0.4211	0.2393	0.1526	0.3693	0.2702	**0.1202**	0.2621
MT-NPO	**0.5200**	**0.2733**	0.1687	**0.3207**	**0.6070**	**0.4253**	**0.2107**	**0.4739**	**0.2800**	0.1088	**0.3509**

4.4 Hierarchical Allocation Analysis

Let a mini-batch be regular when every modality contains a valid token, all log-ratios are finite, and the normalization safeguards are inactive. Define $\eta_t^m = \alpha_F^m \omega_t^m$ and the total modality mass $A_F^m = \sum_{(z,t)\in\mathcal{I}_F^m} \eta_t^m$.

Proposition 2 (Adaptive across modalities, bounded within modalities)**.** *For a regular mini-batch, $\boldsymbol{\eta}$ is a positive probability distribution. For any modalities $m, n \in \mathcal{M}$ and any valid token coordinates j, k within modality m,*

$$A_F^m = \alpha_F^m, \quad \frac{A_F^m}{A_F^n} = \exp\left(\frac{d_F^m - d_F^n}{T}\right), \frac{\ell}{u} \quad \leq \frac{\eta_j^m}{\eta_k^m} = \frac{w_j^m}{w_k^m} \leq \frac{u}{\ell}. \tag{10}$$

Proof. See Appendix A.3.

Thus, MT-NPO can redirect mass across modalities as their accessibility gaps change, while clipping prevents within-modality token allocation from collapsing. Further properties and the resulting detached-weight gradient are given in Appendix A.3.

5 Experiments

We evaluate MT-NPO across benchmarks and model backbones, focusing on its unlearning–retention trade-off, cross-modal alignment, and the contributions of its future-aware and modality-aware components.

5.1 Experimental setup

Benchmarks and models. UMU-Bench evaluates classification, cloze completion, and generation from text-only and image-conditioned inputs, whereas CLEAR evaluates multimodal unlearning through generation (Wang et al., 2025; Dontsov et al., 2025). We report all evaluated forget ratios for both benchmarks, with the larger-ratio results provided in Appendix E. We evaluate LLaVA-1.5-7B (Liu et al., 2024) and Qwen3-VL-8B (Bai et al., 2025a).

Baselines. We compare MT-NPO with the original checkpoint and five representative unlearning baselines, covering gradient-based optimization, preference-based learning, and modality-aware

parameter pruning: GA, GD, NPO, PO, and MANU. Specifically, NPO provides the reference-relative preference objective (Zhang et al., 2024), while MANU is the modality-aware pruning baseline (Liu et al., 2025b). Appendix D provides their objectives and implementation details.

Implementation details. We optimize LoRA adapters (Hu et al., 2022) with AdamW (Loshchilov & Hutter, 2019). Appendix B provides the training procedure and Table 4 reports the UMU-Bench and CLEAR hyperparameters.

Metrics. Following UMU-Bench, we evaluate each task under both text-only and image-conditioned inputs. Specifically, we use modality-specific accuracy for classification and cloze completion, and modality-specific ROUGE-L (Lin, 2004) for generation. CLEAR generation is evaluated with ROUGE-L and BLEU (Papineni et al., 2002). Based on these scores, we further report Unlearning Completeness (UC) and Model Utility (UT) from UMU-Bench, and introduce the Modality Imbalance Coefficient (MIC; lower is better) to measure cross-modal consistency. Appendix C provides the complete definitions and calculations of accuracy, ROUGE-L, UC, UT, and MIC.

5.2 Main Results

Table 1 shows that MT-NPO achieves the strongest aggregate unlearning–retention balance on both backbones, attaining the lowest UC and highest UT among the evaluated unlearning methods. The consistent trend across models indicates that stronger unlearning does not require a disproportionate loss of retained utility.

MT-NPO also produces more aligned changes across text-only and image-conditioned views, as quantified in Figure 3.

Across both backbones and all three UMU tasks, MT-NPO combines substantial unlearning with closely matched reductions between the two views and obtains the lowest aggregate MIC. The task-level reductions rule out a trivially low MIC caused by weak unlearning in both views. This alignment is consistent with MT-NPO's modality-aware design: its adaptive coefficients explicitly control the unlearning and retention contributions of each view instead of optimizing the fused multimodal context as a single unit.

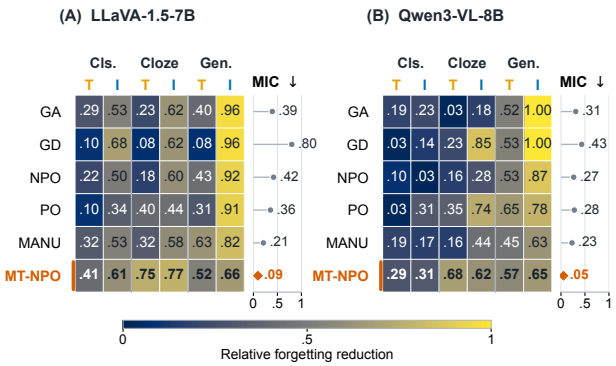

Figure 3: View-specific unlearning reduction and MIC on UMU (T: text; I: image-conditioned); lower MIC is better.

The paired task cells further show that MT-NPO does not obtain its aggregate alignment score by sacrificing one task or access path. Its advantage is visible across both classification-style and generative evaluations.

5.3 Analysis

5.3.1 Generalization across datasets and backbones

CLEAR provides a complementary evaluation with a different dataset and generation protocol. Across both backbones, MT-NPO achieves the lowest forget-set ROUGE-L and BLEU scores and the smallest MIC_F among unlearning methods, while preserving competitive retain-set generation quality. These results demonstrate that MT-NPO generalizes across datasets and model backbones. Appendix E provides the corresponding evaluation over larger forget sets.

Table 2: CLEAR 1% generation results. MIC_F measures only the forget set; lower is better.

Method	LLaVA-1.5-7B					Qwen3-VL-8B				
	F R-L ↓	F BLEU ↓	R R-L ↑	R BLEU ↑	MIC_F ↓	F R-L ↓	F BLEU ↓	R R-L ↑	R BLEU ↑	MIC_F ↓
Original	0.9015	0.8258	0.8170	0.6771	–	0.8703	0.8212	0.7898	0.6429	–
GA	0.4319	0.1327	0.4144	0.2114	0.164	0.3312	0.1098	0.3189	0.1061	0.175
GD	0.4417	0.1812	0.5357	0.2966	0.225	0.4065	0.1789	0.4372	0.2334	0.193
NPO	0.4127	0.1463	0.5230	0.2954	0.152	0.3776	0.1369	0.4259	0.1785	0.138
PO	0.5111	0.2456	0.5746	**0.3872**	0.137	0.4508	0.2372	**0.5402**	0.2702	0.159
MANU	0.3575	0.0938	0.5466	0.3464	0.053	0.3172	0.0944	0.4709	0.2595	0.086
MT-NPO	**0.3222**	**0.0786**	**0.5936**	0.3764	**0.030**	**0.2716**	**0.0660**	0.5283	**0.2973**	**0.056**

5.3.2 FutureKL isolates high-impact tokens

This intervention tests whether FutureKL identifies tokens with greater downstream influence. Figure 4 shows clear separation under two complementary measurements.

At $a = 4$, high-FutureKL selection yields a mismatch rate of 0.911 and an early-divergence score of 0.980, compared with 0.129 and 0.131 for low-FutureKL selection; frequency and random selection lie between them. These results clearly show that higher-FutureKL tokens have substantially greater downstream influence, affecting both the overall extent and the temporal onset of continuation changes. This further supports using FutureKL to identify high-impact tokens and focus unlearning pressure on the most influential positions.

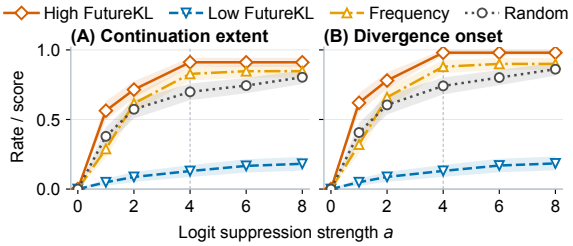

Figure 4: FutureKL intervention over 50 examples: (A) continuation mismatch and (B) early divergence.

5.3.3 Token and modality allocation are complementary

Table 3 separates the contributions of future-aware and modality-aware allocation. Across both backbones, Future-only improves unlearning and retention over NPO but leads to greater cross-modal imbalance, whereas Modality-only substantially reduces MIC while yielding more moderate gains in unlearning and retention. Combining both components, MT-NPO achieves the strongest overall performance across all evaluated settings, reducing unlearning UC and MIC while further improving retained and real-person utility.

Table 3: MT-NPO component ablation.

Model	Variant	Fgt. ↓	Ret. ↑	Real ↑	MIC ↓
LLaVA	NPO	0.5310	0.3127	0.2197	0.4227
	Token-NPO	0.4345	0.3545	0.2489	0.4472
	Future-only	0.4105	0.4672	0.2548	0.4553
	Modality-only	0.4861	0.3205	0.2301	0.1584
	MT-NPO	**0.3575**	**0.5364**	**0.2730**	**0.0917**
Qwen	NPO	0.4855	0.3285	0.2368	0.2654
	Token-NPO	0.4539	0.3048	0.2264	0.2948
	Future-only	0.4109	0.3716	0.2609	0.3188
	Modality-only	0.4768	0.3359	0.2402	0.1237
	MT-NPO	**0.3207**	**0.4143**	**0.2875**	**0.0515**

6 Conclusion

This paper introduced MT-NPO for multimodal model unlearning, where target knowledge must be reliably suppressed while retained utility is preserved across different modality-conditioned views. This setting exposes two key limitations of existing sequence-level methods: sequence coupling, which overlooks the unequal influence of individual answer tokens on future generation, and modality imbalance, where a target may be suppressed under one view yet remain accessible through another. MT-NPO addresses these challenges through FutureKL-guided token weighting and adaptive modality-specific unlearning and retention objectives, focusing optimization on influential response positions and residual access paths. Across multiple benchmarks and MLLM backbones, MT-NPO improves target suppression, utility preservation, and cross-modal balance. Together, these results highlight the importance of coordinating unlearning across both answer tokens and modality-conditioned views, providing a foundation for broader models, data settings, and privacy-sensitive applications.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and polishing, including improving grammar, clarity, and readability. They were not used to develop the research methodology, design or conduct experiments, generate data, formulate or prove mathematical claims, or interpret experimental results. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

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*, pp. 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019. URL https://doi.org/10.1109/SP40001.2021.00019.

Chengyi Cai, Zesheng Ye, Peike Li, Bo Han, Jianzhong Qi, and Feng Liu. Visual-guided key-token regularization for multimodal large language model unlearning. *arXiv preprint arXiv:2601.22020*, 2026. URL https://arxiv.org/abs/2601.22020.

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. doi: 10.1109/SP.2015.35. URL https://doi.org/10.1109/SP.2015.35.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Junkai Chen, Yuhao He, Junxiang You, Ruiqi Liu, Chenyu Wang, and Shu Wu. Visual-noise guided in-context distillation for multimodal large language model unlearning. *arXiv preprint arXiv:2606.00105*, 2026. URL https://arxiv.org/abs/2606.00105.

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, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.1058. URL https://aclanthology.org/2025.findings-acl.1058/.

André V. Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. DE-COP: Detecting copyrighted content in language models training data. *arXiv preprint arXiv:2402.09910*, 2024. URL https://arxiv.org/abs/2402.09910.

Antonio A. Ginart, Melody Y. Guan, Gregory Valiant, and James Y. Zou. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/cb79f8fa58b91d3af6c9c991f63962d3-Abstract.html.

Jihao Gu, Yingyao Wang, Meng Cao, Pi Bu, Jun Song, Bo Zheng, Yancheng He, and Shilong Li. Token preference optimization with self-calibrated visual-anchored rewards for hallucination mitigation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19754–19767, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1076. URL https://aclanthology.org/2025.findings-emnlp.1076/.

Jiahui Guang, Haiyan Wang, Yingjie Zhu, Cuiyun Gao, Jing Li, Di Shao, and Zhaoquan Gu. ASRU: Activation steering meets reinforcement unlearning for multimodal large language models. *arXiv preprint arXiv:2605.15687*, 2026. URL https://arxiv.org/abs/2605.15687.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.375. URL https://aclanthology.org/2025.findings-acl.375/.

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, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. Propile: Probing privacy leakage in large language models. *Advances in neural information processing systems*, 36:20750–20762, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a. URL https://arxiv.org/abs/2408.03326.

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, 2024b.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

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, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

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, Albuquerque, New Mexico, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.207. URL https://aclanthology.org/2025.naacl-long.207/.

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, Vienna, Austria, 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.295. URL https://aclanthology.org/2025.acl-long.295/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Chiyu Ma, Shuo Yang, Kexin Huang, Jinda Lu, Haoming Meng, Shangshang Wang, Bolin Ding, Soroush Vosoughi, Guoyin Wang, and Jingren Zhou. FIPO: Eliciting deep reasoning with future-KL influenced policy optimization. *arXiv preprint arXiv:2603.19835*, 2026. URL https://arxiv.org/abs/2603.19835.

Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Jinsheng Pan, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, Yejin Choi, Muhao Chen, and Chaowei Xiao. Benchmarking vision language model unlearning via fictitious facial identity dataset. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 87806–87827, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/dafb7f4ced9ac559e077fc611d6e96e5-Paper-Conference.pdf.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=B41hNBoWLo.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Ruichen Qiu, Jiajun Tan, Jiayue Pu, Honglin Wang, Xiao-Shan Gao, and Fei Sun. A survey on unlearning in large language models. *arXiv preprint arXiv:2510.25117*, 2025. URL https://arxiv.org/abs/2510.25117.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TArmA033BU.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. PaliGemma 2: A family of versatile VLMs for transfer. *arXiv preprint arXiv:2412.03555*, 2024. URL https://arxiv.org/abs/2412.03555.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Chengye Wang, Yuyuan Li, Xiaohua Feng, Chaochao Chen, Xiaolin Zheng, and Jianwei Yin. UMU-bench: Closing the modality gap in multimodal unlearning evaluation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/18b0ee1e95f4007541e57fe507b5167e-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.460. URL https://aclanthology.org/2024.emnlp-main.460/.

Yuhang Wang, Zhenxing Niu, Haoxuan Ji, Guangyu He, Linlin Zhang, and Haichang Gao. Null space constrained contrastive visual forgetting for MLLM unlearning. *arXiv preprint arXiv:2605.05909*, 2026. URL https://arxiv.org/abs/2605.05909.

Haoming Xu, Ningyuan Zhao, Liming Yang, Sendong Zhao, Shumin Deng, Mengru Wang, Bryan Hooi, Nay Oo, Huajun Chen, and Ningyu Zhang. 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)*, pp. 5967–5987, Vienna, Austria, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.297. URL https://aclanthology.org/2025.acl-long.297/.

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

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Erxue Min, and Sophia Ananiadou. Selective preference optimization via token-level reward function estimation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7032–7056, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.359. URL https://aclanthology.org/2025.emnlp-main.359/.

Zhengbang Yang, Yisheng Zhong, Junyuan Hong, and Zhuangdi Zhu. CALIBURN: Self-calibrated LLM unlearning alignment. *arXiv preprint arXiv:2602.02824*, 2026. URL https://arxiv.org/abs/2602.02824.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 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)*, pp. 8403–8419, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.457. URL https://aclanthology.org/2024.acl-long.457/.

Junxiang You, Junkai Chen, Yuhao He, Ruiqi Liu, Zhetao Guo, and Shu Wu. Exploring and bridging knowledge holes in unlearned multimodal large language models. *arXiv preprint arXiv:2608.01849*, 2026. URL https://arxiv.org/abs/2608.01849.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58348–58365. PMLR, 2024. URL https://proceedings.mlr.press/v235/zeng24c.html.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=MXLBXjQkmb.

Wenxuan Zhou, Shujian Zhang, Lingxiao Zhao, and Tao Meng. T-REG: Preference optimization with token-level reward regularization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27876–27889, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1353. URL https://aclanthology.org/2025.acl-long.1353/.

A Additional theoretical results and proofs

A.1 Derivation of the pathwise future horizon

Derivation. Fix an example and a conditioning pathway, as in Section 4.2. Substituting $\delta_k = \log q_k$ into the definition of κ_t and exponentiating gives $\exp(\kappa_t) = \prod_{k=t}^{L} q_k^{g_k \gamma^{k-t}}$. When $g_k = 1$ for all k, taking $\gamma \to 1^-$ and using autoregressive factorization yields $\prod_{k=t}^{L} q_k = \pi_\theta(\mathbf{y}_{t:L} \mid s_t(z))/\pi_{\mathrm{ref}}(\mathbf{y}_{t:L} \mid s_t(z))$. The half-life identity follows directly from $\gamma^{\bar{\tau}} = (2^{-1/\bar{\tau}})^{\bar{\tau}} = 1/2$. For the truncation statement, assume $|g_k \delta_k| \le B$ along the suffix and let $\kappa_{t,H}$ retain only indices $k \le t + H$. Then

$$|\kappa_t - \kappa_{t,H}| \le \sum_{h=H+1}^{\infty} \gamma^h B = \frac{B\gamma^{H+1}}{1-\gamma}.$$

The infinite sum upper-bounds any finite suffix.

A.2 Proof of tokenwise control and anti-saturation

Proof of Proposition 1. Because the normalization safeguard is inactive and every clipped weight lies in $[\ell, u]$, each valid coordinate satisfies

$$\omega_t = \frac{w_t}{\sum_j w_j} \ge \frac{\ell}{\ell + (N-1)u} = \omega_{\min}.$$

Every term in $\mathcal{L}_{\mathrm{TNPO}}$ is nonnegative. Hence $\mathcal{L}_{\mathrm{TNPO}} \le \varepsilon$ implies $\phi_\beta(\delta_t) \le \varepsilon/\omega_t \le \varepsilon/\omega_{\min}$ for every valid token. The kernel is strictly increasing and its inverse on $(0, \infty)$ is $\phi_\beta^{-1}(x) = \beta^{-1} \log(\exp(\beta x/2) - 1) = \psi_\beta(x)$, which proves Eq. 5.

For sequence NPO, monotonicity only gives $\sum_t \delta_t \le \psi_\beta(\varepsilon)$. When $L \ge 2$, fix any coordinate j and any $R > 0$, set $\delta_j = R$ and one other coordinate to $\psi_\beta(\varepsilon) - R$, and set the remaining coordinates to zero. The sequence loss equals ε while $\delta_j = R$ can be arbitrarily large across valid model–reference pairs (any finite log-ratio can be realized by choosing positive token probabilities at the corresponding states). Thus no coordinatewise upper bound follows from the sequence loss alone.

Finally, $\phi'_\beta(\delta_t) = 2\sigma(\beta\delta_t)$. Since the normalized weights sum to one,

$$\|\mathbf{a}\|_1 = 2 \sum_t \omega_t \sigma(\beta\delta_t) \le 2.$$

If $\delta_t \ge 0$, then $\sigma(\beta\delta_t) \ge 1/2$ and therefore $a_t \ge \omega_t \ge \omega_{\min}$. By comparison, differentiating $\phi_\beta(\sum_t \delta_t)$ with respect to any coordinate gives the shared coefficient $2\sigma(\beta \sum_t \delta_t)$, which tends to zero if another coordinate tends to $-\infty$. This proves Eq. 6. □

A.3 Additional modality-allocation properties

Temperature interpretation. Let $r_F^m = d_F^m$, $r_R^m = \mathcal{L}_R^m$, $T > 0$ be the shared allocation temperature, and $H(\mathbf{a}) = -\sum_m a^m \log a^m$. For $S \in \{F, R\}$, the softmax allocator is the unique entropy-regularized maximizer

$$\boldsymbol{\alpha}_S = \operatorname*{arg\,max}_{\mathbf{a} \in \Delta_{\mathcal{M}}} \left\{ \sum_{m \in \mathcal{M}} a^m r_S^m + T H(\mathbf{a}) \right\}, \quad \Delta_{\mathcal{M}} = \left\{ \mathbf{a} \ge 0 : \sum_m a^m = 1 \right\}. \tag{11}$$

This follows directly from the first-order optimality conditions. As $T \to 0^+$, the solution concentrates on the highest-scoring view; as $T \to \infty$, it approaches uniform averaging.

Proof of Proposition 2. For each modality, regularity makes $Z^m = \sum_{(z,t) \in \mathcal{I}_F^m} w_t^m$. Hence every ω_t^m is positive and the coefficients within each modality sum to one. It follows that the total mass assigned to modality m is

$$A_F^m = \sum_{(z,t) \in \mathcal{I}_F^m} \eta_t^m = \alpha_F^m \sum_{(z,t) \in \mathcal{I}_F^m} \omega_t^m = \alpha_F^m.$$

Because the softmax normalizer cancels between two modalities,

$$\frac{A_F^m}{A_F^n} = \frac{\alpha_F^m}{\alpha_F^n} = \exp\left(\frac{d_F^m - d_F^n}{T}\right).$$

For two tokens j, k in the same modality, the modality coefficient and token normalizer cancel, so

$$\frac{\eta_j^m}{\eta_k^m} = \frac{\alpha_F^m w_j^m / Z^m}{\alpha_F^m w_k^m / Z^m} = \frac{w_j^m}{w_k^m} \in \left[\frac{\ell}{u}, \frac{u}{\ell}\right],$$

which proves Eq. 10. Softmax gives $\alpha_F^m > 0$ and $\sum_m \alpha_F^m = 1$, and therefore

$$\sum_m \sum_{(z,t)\in\mathcal{I}_F^m} \eta_t^m = \sum_m \alpha_F^m \sum_{(z,t)\in\mathcal{I}_F^m} \omega_t^m = \sum_m A_F^m = 1.$$

Thus $\boldsymbol{\eta}$ is a positive probability distribution, completing the proof. □

Convexity consequence. Substituting $\eta_t^m = \alpha_F^m \omega_t^m$ into Eq. 7 expresses $\mathcal{L}_F$ as a convex combination of the valid token kernels. Therefore,

$$\min_{\substack{m\in\mathcal{M}\\(z,t)\in\mathcal{I}_F^m}} \phi_\beta(\delta_t^m) \le \mathcal{L}_F \le \max_{\substack{m\in\mathcal{M}\\(z,t)\in\mathcal{I}_F^m}} \phi_\beta(\delta_t^m). \tag{12}$$

Reductions and complexity. One modality recovers TNPO, zero future scores recover uniformly weighted token NPO, and equal modality scores recover uniform modality averaging. The uniform-token case remains distinct from sequence NPO because the nonlinear kernel lies inside the token sum. Future scores require one reverse scan per modality, and modality weighting adds $\mathcal{O}(|\mathcal{M}|)$ work beyond the token losses. These properties do not imply certified deletion or equivalence to retraining.

A.4 Gradient under detached allocation weights

During backpropagation, the FutureKL-derived token weights ω_t^m and the modality weights α_F^m, α_R^m are detached and treated as constants; their forward values are recomputed at every optimization step. Therefore,

$$\nabla_\theta \mathcal{L}_{\mathrm{TNPO}}^m = \sum_{(z,t)\in\mathcal{I}_F^m} \omega_t^m \nabla_\theta \phi_\beta(\delta_t^m), \tag{13}$$

$$\nabla_\theta \mathcal{L}_F = \sum_m \alpha_F^m \nabla_\theta \mathcal{L}_{\mathrm{TNPO}}^m, \tag{14}$$

and

$$\nabla_\theta \mathcal{L}_R = \sum_m \alpha_R^m \nabla_\theta \mathcal{L}_R^m. \tag{15}$$

These identities follow directly by differentiating the three losses while holding the detached allocation coefficients fixed.

B TRAINING PROCEDURE

Algorithm 1 exposes the two allocation levels without specializing the objective to a particular number of modalities. Table 4 lists the settings for UMU-Bench and CLEAR on both backbones.

Algorithm 1: MT-NPO training step

Input: Trainable model π_θ; frozen reference π_{ref}; paired mini-batches $(\mathcal{B}_F, \mathcal{B}_R)$; modality set $\mathcal{M}$; hyperparameters $\beta, \tau, c, \epsilon_-, \epsilon_+, T, \lambda_R$

Output: Updated parameters θ

$\bar{\tau} \leftarrow \max\{1, \tau\}$; $\gamma \leftarrow 2^{-1/\bar{\tau}}$

Stage 1: Future-aware unlearning

foreach $m \in \mathcal{M}$ **do**

  Compute $\{\delta_t^m(z) : (z,t) \in \mathcal{I}_F^m\}$ between π_θ and π_{ref}

  Compute $\boldsymbol{\kappa}^m$ by the right-to-left recurrence $\kappa_t^m = g_t^m \delta_t^m + \gamma \kappa_{t+1}^m$

  Clip $\exp(\boldsymbol{\kappa}^m)$ to obtain $\mathbf{w}^m$ and normalize it to $\boldsymbol{\omega}^m$

  Compute $(\mathcal{L}_{\text{TNPO}}^m, d_F^m)$ as defined in Section 4.3

end

$\boldsymbol{\alpha}_F \leftarrow \text{softmax}(\mathbf{d}_F/T)$; $\mathcal{L}_F \leftarrow \sum_{m \in \mathcal{M}} \alpha_F^m \mathcal{L}_{\text{TNPO}}^m$

Stage 2: Modality-aware preservation

foreach $m \in \mathcal{M}$ **do**

  Compute $\{\delta_t^m(z) : (z,t) \in \mathcal{I}_R^m\}$ and $\mathcal{L}_R^m$ as defined in Section 4.3

end

$\boldsymbol{\alpha}_R \leftarrow \text{softmax}(\mathbf{L}_R/T)$; $\mathcal{L}_R \leftarrow \sum_{m \in \mathcal{M}} \alpha_R^m \mathcal{L}_R^m$

Stage 3: Parameter update

$\mathcal{L} \leftarrow \mathcal{L}_F + \lambda_R \mathcal{L}_R$

Backpropagate $\mathcal{L}$ and clip the parameter-gradient norm

Update θ; advance the learning-rate scheduler

Table 4: MT-NPO hyperparameters on UMU-Bench 5% and CLEAR 1%.

Hyperparameter	UMU-Bench (5%)		CLEAR (1%)		Role
	LLaVA-1.5-7B	Qwen3-VL-8B	LLaVA-1.5-7B	Qwen3-VL-8B	
`lr`	6.5×10^{-5}	3.0×10^{-4}	2.5×10^{-5}	1.5×10^{-4}	Learning rate
`lambda_R`	0.3	0.3	0.3	0.3	Retain-loss weight
`beta`	0.5	0.5	0.5	0.5	NPO sharpness
`num_epochs`	2	2	2	2	Number of epochs
`batch_size`	2	2	2	2	Batch size
`tau`	32.0	32.0	32.0	32.0	FutureKL decay length
`kl_clip_c`	3.0	3.0	3.0	3.0	FutureKL ratio threshold
`f_eps_low`	0.2	0.2	0.2	0.2	Future-weight lower amplitude
`f_eps_high`	0.2	0.2	0.2	0.2	Future-weight upper amplitude
LoRA r	64	64	64	64	LoRA rank
LoRA α	32	32	32	32	LoRA scaling factor
`T`	1.0	1.0	1.0	1.0	Temperature

C EVALUATION METRIC DETAILS

Accuracy aggregation. For a classification or cloze example s, let $c_{s,\mathrm{img}}, c_{s,\mathrm{txt}} \in \{0,1\}$ indicate whether the image-conditioned and text-only predictions match their targets. For an evaluation set S, define

$$\mathrm{Acc}_{\mathrm{img}} = \frac{1}{|S|}\sum_{s\in S} c_{s,\mathrm{img}}, \qquad \mathrm{Acc}_{\mathrm{txt}} = \frac{1}{|S|}\sum_{s\in S} c_{s,\mathrm{txt}},$$
$$\mathrm{Acc}_{\mathrm{any}} = \frac{1}{|S|}\sum_{s\in S} \mathbb{I}[c_{s,\mathrm{img}} \vee c_{s,\mathrm{txt}}], \qquad \mathrm{Acc}_{\mathrm{all}} = \frac{1}{|S|}\sum_{s\in S} \mathbb{I}[c_{s,\mathrm{img}} \wedge c_{s,\mathrm{txt}}]. \tag{16}$$

The forget-set and utility-side accuracies are then

$$\mathrm{Acc}_F = \frac{\mathrm{Acc}_{\mathrm{img}} + \mathrm{Acc}_{\mathrm{txt}} + \mathrm{Acc}_{\mathrm{any}}}{3}, \qquad \mathrm{Acc}_R = \frac{\mathrm{Acc}_{\mathrm{img}} + \mathrm{Acc}_{\mathrm{txt}} + \mathrm{Acc}_{\mathrm{all}}}{3}. \tag{17}$$

We use Acc_F on the forget set and Acc_R on both the retain and real-person sets. Thus a forget-set example is penalized if either view still answers correctly, whereas a utility example receives full credit only when both views remain correct.

ROUGE-L aggregation. Let $r_{s,\mathrm{img}}$ and $r_{s,\mathrm{txt}}$ be the per-example ROUGE-L scores (Lin, 2004) for the image-conditioned and text-only generations. Define the self-weighted and harmonic means

$$W(x,y) = \frac{x^2+y^2}{x+y}, \qquad H(x,y) = \frac{2xy}{x+y}, \tag{18}$$

with either expression set to zero when $x + y = 0$. The reported generation scores are computed per example and then averaged:

$$\mathrm{RL}_F = \frac{1}{|S_F|}\sum_{s\in S_F} W(r_{s,\mathrm{img}}, r_{s,\mathrm{txt}}), \qquad \mathrm{RL}_R(S) = \frac{1}{|S|}\sum_{s\in S} H(r_{s,\mathrm{img}}, r_{s,\mathrm{txt}}), \tag{19}$$

where S is the retain or real-person set. The self-weighted mean emphasizes the larger residual score on the forget set, while the harmonic mean decreases when either utility view deteriorates.

UC and UT. Writing $\mathrm{Acc}^{\mathrm{cls}}$ and $\mathrm{Acc}^{\mathrm{cloze}}$ for the two accuracy tasks, the reported aggregates are

$$\mathrm{UC} = \frac{1}{3}\left(\mathrm{Acc}_F^{\mathrm{cls}} + \mathrm{Acc}_F^{\mathrm{cloze}} + \mathrm{RL}_F\right), \tag{20}$$

$$\mathrm{UT} = \frac{1}{6}\sum_{S\in\{S_R,S_P\}}\left(\mathrm{Acc}_R^{\mathrm{cls}}(S) + \mathrm{Acc}_R^{\mathrm{cloze}}(S) + \mathrm{RL}_R(S)\right), \tag{21}$$

where S_R and S_P denote the retain and real-person sets, respectively.

Modality Imbalance Coefficient. Let $\mathcal{K}$ index the evaluated tasks and let $q_{k,m}^{(0)} > 0$ be the original-model performance on task $k \in \mathcal{K}$ under modality $m \in \mathcal{M}$, and let $d_{k,m} = 1 - q_{k,m}/q_{k,m}^{(0)}$ denote the relative reduction after unlearning. For $|\mathcal{M}| \geq 2$, define

$$\mathrm{MIC} = \frac{\sum_{k\in\mathcal{K}}\sum_{m<n}|d_{k,m} - d_{k,n}|}{(|\mathcal{M}|-1)\sum_{k\in\mathcal{K}}\sum_{m}|d_{k,m}|}, \tag{22}$$

with MIC reported as "–" when the denominator is zero. MIC measures normalized pairwise disagreement across modality-specific changes and, when defined, satisfies $0 \leq \mathrm{MIC} \leq 1$; lower values indicate better alignment. We also use the signed mean reduction $S = (|\mathcal{K}||\mathcal{M}|)^{-1}\sum_{k\in\mathcal{K}}\sum_{m} d_{k,m}$ to distinguish aligned unlearning from the degenerate case in which no modality changes.

D BASELINE METHODS

Let $\mathcal{L}_{\mathrm{CE}}(\mathcal{D};\theta)$ denote the mean autoregressive cross-entropy on a dataset $\mathcal{D}$:

$$\mathcal{L}_{\mathrm{CE}}(\mathcal{D};\theta) = -\frac{1}{|\mathcal{D}|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}}\sum_{t=1}^{|\mathbf{y}|}\log \pi_\theta(y_t \mid \mathbf{x}, \mathbf{y}_{<t}). \tag{23}$$

All optimization-based baselines use the same forget-set and retain-set splits as MT-NPO.

Original. The Original row is the frozen checkpoint before unlearning. It provides the reference utility and the pre-unlearning probabilities used by NPO and the relative-change metrics.

Gradient Ascent (GA). GA directly increases the language-model loss on the forget set. Under loss minimization, its objective is

$$\mathcal{L}_{\mathrm{GA}} = -\mathcal{L}_{\mathrm{CE}}(\mathcal{D}_F;\theta). \tag{24}$$

This aggressively suppresses the target responses but does not explicitly protect retained behavior.

Gradient Difference (GD). GD adds ordinary retain-set training to forget-set ascent:

$$\mathcal{L}_{\mathrm{GD}} = -\mathcal{L}_{\mathrm{CE}}(\mathcal{D}_F;\theta) + \lambda_{\mathrm{GD}}\mathcal{L}_{\mathrm{CE}}(\mathcal{D}_R;\theta). \tag{25}$$

The retain term counteracts utility degradation induced by GA.

Negative Preference Optimization (NPO). NPO treats the original target response as a negative completion and lowers its likelihood relative to a frozen reference model. We use the sequence-level objective in Eq. 2; unlike MT-NPO, it applies one nonlinear gate to the complete response and has no explicit modality allocator.

Preference Optimization (PO). PO replaces each target response in the forget set with a refusal response such as "I do not know" and trains jointly on these refusal examples and the retain set. If $\mathcal{D}_F^{\mathrm{idk}}$ denotes the refusal-target dataset, its objective is

$$\mathcal{L}_{\mathrm{PO}} = \mathcal{L}_{\mathrm{CE}}(\mathcal{D}_F^{\mathrm{idk}};\theta) + \lambda_{\mathrm{PO}}\mathcal{L}_{\mathrm{CE}}(\mathcal{D}_R;\theta). \tag{26}$$

Modality-Aware Neuron Unlearning (MANU). MANU is a pruning-based multimodal baseline (Liu et al., 2025b). It first identifies neurons that are especially important to the targeted knowledge under different modalities and then selectively clips the chosen neurons. This two-stage procedure removes modality-specific target representations without performing the token-level loss allocation used by MT-NPO.

E ADDITIONAL RESULTS ACROSS FORGET-SET SCALES

E.1 UMU-BENCH AT LARGER FORGET RATIOS

Table 5: UMU-Bench results at the 10% forget ratio. Lower forget-set metrics and UC are better; higher utility metrics and UT are better. Bold and underlined denote the best and second-best unlearning methods within each backbone.

Method	Forget set ↓				Retain set ↑			Real-person set ↑			UT ↑
	Cls.	Cloze	Gen.	UC	Cls.	Cloze	Gen.	Cls.	Cloze	Gen.	
LLaVA-1.5-7B											
Origin	0.8233	0.9100	0.8564	0.8632	0.6781	0.8078	0.7415	0.4063	0.2865	0.1751	0.5159
GA	0.5767	0.5733	0.5483	0.5661	0.3989	0.4119	0.1586	0.2185	0.2853	0.0278	0.2502
GD	0.5667	0.5933	0.6515	0.6038	0.3863	0.4367	0.0507	0.3627	0.2794	0.0231	0.2565
NPO	0.5800	0.5633	0.4887	0.5440	0.4319	0.4241	0.0791	0.3529	**0.2884**	0.0134	0.2650
PO	0.6600	0.6167	0.5389	0.6052	0.4570	**0.4470**	0.1647	0.3573	0.2808	0.0754	0.2971
MANU	0.6133	0.5533	0.4486	0.5384	0.3989	0.3515	0.2949	0.3562	0.2664	0.0997	0.2946
MT-NPO	**0.4933**	**0.2833**	**0.4480**	**0.4082**	**0.4974**	0.4026	**0.4207**	**0.4107**	0.2854	**0.1133**	**0.3550**
Qwen3-VL-8B											
Origin	0.7467	0.6300	0.4105	0.5957	0.6178	0.5204	0.3125	0.5000	0.2767	0.1406	0.3947
GA	0.6167	0.4867	0.2501	0.4511	0.4333	0.3967	0.0101	0.3551	0.2593	0.0044	0.2431
GD	0.6433	0.3700	0.2476	0.4203	0.5367	0.3467	0.0217	**0.4728**	0.2081	0.0107	0.2661
NPO	0.6600	0.4800	0.2141	0.4514	0.5033	0.4167	0.0598	0.4466	0.2505	0.0048	0.2803
PO	0.6067	0.3533	0.2381	0.3994	0.5867	0.4133	0.1282	0.3769	0.2375	0.0182	0.2935
MANU	0.6167	0.4500	**0.1881**	0.4183	0.4267	0.2300	0.1669	0.3725	0.2582	0.1267	0.2635
MT-NPO	**0.5933**	**0.2833**	0.2016	**0.3594**	**0.6221**	**0.4881**	**0.2580**	0.4649	**0.2639**	**0.1316**	**0.3714**

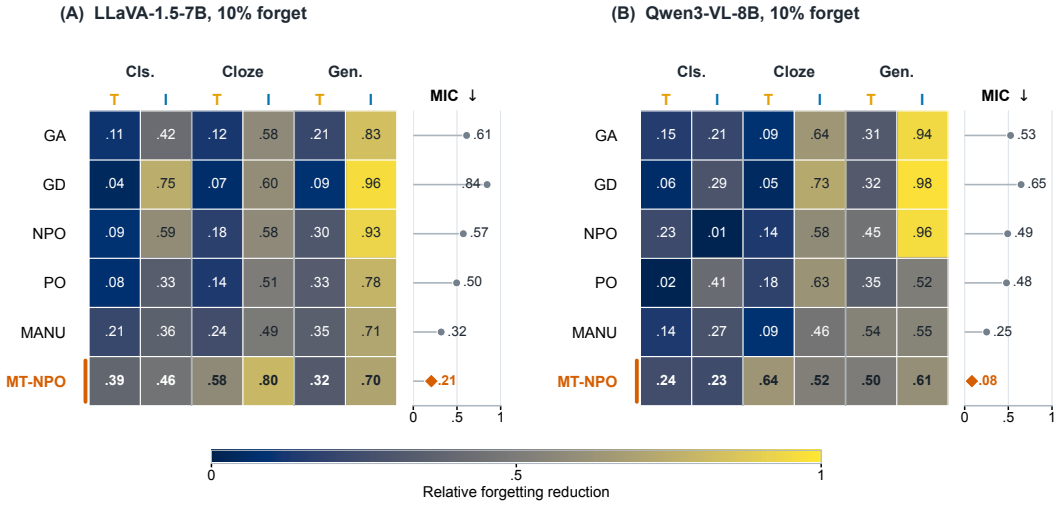

Figure 5: View-specific unlearning reduction and MIC on UMU at the 10% forget ratio (T: text; I: image-conditioned); lower MIC is better.

E.2 CLEAR AT LARGER FORGET RATIOS

Table 6: Additional CLEAR generation results at 5% and 10% forget ratios. MIC_F is computed from forget-set ROUGE-L and BLEU; lower forget metrics and MIC_F and higher retain metrics are better. Bold and underlined denote the best and second-best unlearning methods within each setting.

Method	LLaVA-1.5-7B					Qwen3-VL-8B				
	F R-L ↓	F BLEU ↓	R R-L ↑	R BLEU ↑	MIC_F ↓	F R-L ↓	F BLEU ↓	R R-L ↑	R BLEU ↑	MIC_F ↓
5% forget										
Origin	0.8847	0.8076	0.8238	0.6884	–	0.8514	0.8369	0.8012	0.6315	–
GA	0.4147	0.1674	0.3816	0.2318	0.137	0.3628	0.0964	0.2826	0.1128	0.156
GD	0.4416	0.2148	0.4978	0.3194	0.187	0.3691	0.1127	0.4685	0.2016	0.130
NPO	0.4527	0.1846	0.4417	0.3092	0.113	0.4062	0.0963	0.3987	0.2149	0.134
PO	0.4741	0.2384	0.5653	0.3414	0.099	0.3826	0.1624	**0.5681**	0.2517	0.199
MANU	0.3587	0.1132	0.5734	0.3291	0.037	0.3118	0.0826	0.5012	0.2419	0.064
MT-NPO	**0.3429**	**0.0692**	**0.6178**	**0.3581**	**0.024**	**0.2897**	**0.0581**	0.5514	**0.2761**	**0.050**
10% forget										
Origin	0.9132	0.7914	0.8072	0.6961	–	0.8896	0.7983	0.7741	0.6598	–
GA	0.3862	0.1197	0.4175	0.1863	0.049	0.3371	0.1248	0.3264	0.0839	0.199
GD	0.4789	0.1892	0.4931	0.3247	0.114	0.4018	0.1316	0.4167	0.2458	0.202
NPO	0.4093	0.2081	0.5324	0.2675	0.096	0.3594	0.1697	0.4491	0.1726	0.107
PO	0.5298	0.2616	0.5612	**0.3778**	0.118	0.4817	0.2113	**0.5179**	0.3014	0.197
MANU	0.4026	0.1228	0.5267	0.3515	0.041	0.3696	0.0842	0.4568	0.2861	0.053
MT-NPO	**0.3047**	**0.0714**	**0.5716**	0.3648	**0.029**	**0.2613**	**0.0794**	0.5037	**0.3186**	**0.051**