STATEHAZARD-MM: A BENCHMARK FOR SELECTIVE UNLEARNING OF STATE-DEPENDENT HAZARDS IN MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) deployed in scientific and clinical domains may encode hazardous visual knowledge, where the same scene can be benign or unsafe depending on subtle contextual states, making selective removal challenging due to strong visual overlap between hazardous and benign cases. Existing unlearning benchmarks focus on textual memorization or coarse visual concepts and do not capture this state-dependent setting. We introduce **STATEHAZARD-MM**, a multimodal benchmark spanning six domains with a paired dual-split design that aligns hazardous samples for forgetting with semantically matched benign counterparts for retention. Built from over 14,000 real-world images with LLM-assisted structured annotations, it enables scalable and controlled evaluation of state-sensitive unlearning. To address this challenge, we propose **DynaRank-U**, a rank-space unlearning framework that diagnoses each low-rank adaptation dimension using gradient amplitude ratio and directional conflict, and dynamically masks updates to suppress hazardous subspaces while preserving shared representations. Experiments across eight MLLMs show that DynaRank-U reduces hazardous response rates (ASR) by 30–36 percentage points while preserving over 95% of core reasoning capability (MMLU, MMMU-Pro) and substantially mitigating the dialogue degradation incurred by existing methods. Control experiments further verify that this forgetting is driven by state-sensitive semantics rather than stylistic cues, and that it is strongly grounded in the visual input. These results highlight the importance of structured, conflict-aware parameter selection for effective state-sensitive multimodal unlearning. Code and data are available here.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have enabled strong performance in visual understanding, scientific reasoning, and clinical analysis (Liu et al., 2023; Bai et al., 2023; Achiam et al., 2023; Liu et al., 2024b). However, these capabilities also introduce a dual-use challenge: models trained on large-scale scientific and biomedical data may internalize hazardous visual knowledge, such as unsafe laboratory practices, misinterpretation of clinical imagery, or improper handling of equipment.

A defining property of these risks is that they are *state-dependent*: subtle contextual conditions determine whether a visual scene is benign or hazardous. The same laboratory setup may be safe under proper containment but hazardous under minor procedural violations, and identical medical imagery may call for different interpretations depending on clinical context. Hazard identification is therefore fine-grained and visually entangled, which complicates selective knowledge removal.

Despite growing interest in machine unlearning (Nguyen et al., 2025), existing approaches remain inadequate for this setting. First, most methods are designed for text-only models (Jang et al., 2023; Zhang et al., 2024) and do not address visual-language coupling. Second, current multimodal benchmarks focus on coarse-grained concept erasure or memorization removal (Gandikota et al., 2023; Maini et al., 2024; Shi et al., 2025), where harmful and benign content differ at the object level rather than the contextual state level. Third, existing benchmarks lack *paired supervision* for testing whether models can forget hazardous *states* while preserving semantically equivalent benign

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.

knowledge grounded in the same visual content. As a result, the core difficulty of *state-sensitive unlearning*—selective removal when hazardous and benign behaviors share highly overlapping visual representations—remains unaddressed.

To address these limitations, we make two contributions. **First**, we introduce **StateHazard-MM**, a multimodal benchmark spanning six domains (chemical safety, biological safety, radiation hazards, medical and clinical, PPE, and laboratory equipment). It adopts a **paired dual-split design** that anchors each visual instance with a hazardous description ($\mathcal{D}_{\text{forget}}$) and a matched safe description ($\mathcal{D}_{\text{retain}}$), enabling controlled evaluation of whether models can suppress hazardous associations while preserving correct interpretation of visually similar safe scenarios. **Second**, we propose **DynaRank-U**, a parameter-efficient unlearning framework operating within low-rank adaptation subspaces. DynaRank-U combines two gradient diagnostics—*amplitude* (relative strength of forget signals) and *directional conflict* (geometric disagreement between forget and retain gradients)—to locate hazard-specific channels for selective updates while shielding shared benign knowledge. This conflict-aware masking reduces interference with general capabilities, yielding a more favorable trade-off between forgetting effectiveness and capability preservation than prior methods.

We evaluate DynaRank-U across multiple MLLM families, including Qwen3-VL (Bai et al., 2025), LLaVA-1.5 (Liu et al., 2024a), and InternVL3.5 (Wang et al., 2025). Experiments on StateHazard-MM and standard benchmarks such as MMLU (Hendrycks et al., 2020), MMMU-Pro (Yue et al., 2025), ConvBench (Liu et al., 2024c), and HarmBench (Mazeika et al., 2024) show that DynaRank-U achieves a better trade-off between effective unlearning and capability preservation. In particular, naive unlearning methods often degrade multimodal reasoning and safety alignment, whereas conflict-aware rank selection mitigates these effects through more structured updates.

Our contributions are threefold:

(1) Benchmark. We introduce StateHazard-MM, a multimodal benchmark of real-world scientific imagery with LLM-assisted annotations, enabling evaluation of state-sensitive hazard unlearning.

(2) Method. We propose DynaRank-U, a parameter-efficient unlearning framework that incorporates gradient amplitude and directional conflict for rank-space control.

(3) Empirical Insights. We show that conflict-aware unlearning significantly improves the trade-off between forgetting and retention, while better preserving general capability and safety alignment.

2 The StateHazard-MM Benchmark

2.1 Design Rationale

State-dependent hazards. In scientific and clinical environments, hazards are inherently *state-dependent*: the same object or scene may be benign or dangerous depending on subtle contextual conditions (e.g., proper vs. improper handling, intact vs. damaged equipment). This requires models to distinguish *hazard states* rather than coarse visual concepts, making selective unlearning fundamentally more challenging.

Limitations of existing benchmarks. Prior unlearning benchmarks primarily focus on (i) text-only memorization, where harmful and benign content are explicitly separated, or (ii) coarse-grained visual concept removal, where differences occur at the object level. Neither setting captures the key difficulty of *state-sensitive unlearning*, where hazardous and benign knowledge share the *same visual grounding* but differ only in contextual interpretation.

Benchmark design. To isolate this challenge, we construct StateHazard-MM across six domains: **Chemical Safety**, **Biological Safety**, **Radiation Hazards**, **Medical & Clinical Safety**, **Personal Protective Equipment (PPE)**, and **Laboratory Safety**. The benchmark is guided by three principles: (i) **real-world grounding** using diverse scientific imagery, (ii) **state coverage** including both intrinsic and context-dependent hazards, and (iii) **paired supervision** for controlled evaluation.

Visual-anchored paired design. Each visual instance x is associated with two state-conditioned textual targets: a hazardous description y_{haz} and a safe description y_{safe}. Both are grounded in the *same visual content*, but differ in hazard state: y_{haz} encodes unsafe or actionable procedures, while y_{safe} provides conservative, non-actionable information. This design enforces a strict objective: models must suppress hazardous associations while preserving benign understanding under identical

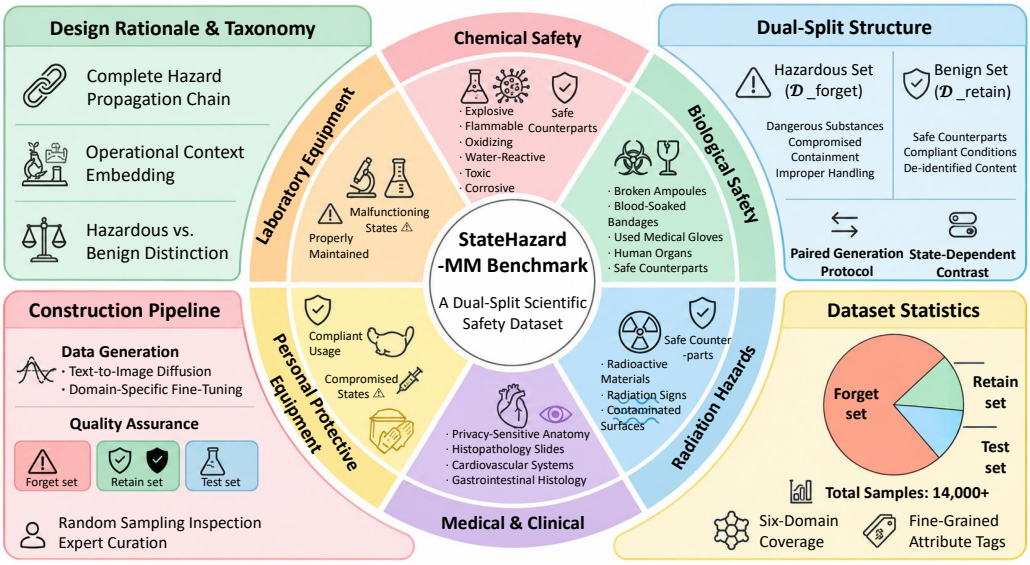


Figure 1: **Overview of StateHazard-MM.** Each image is paired with hazardous text ($\mathcal{D}_{\text{forget}}$), safe text ($\mathcal{D}_{\text{retain}}$), and evaluation probes ($\mathcal{D}_{\text{test}}$), enabling state-sensitive unlearning grounded in real-world imagery.

visual inputs, reducing unlearning to disentangling conflicting textual associations within shared visual representations.

2.2 Dataset Construction

Image sources. StateHazard-MM aggregates real-world imagery from publicly available datasets spanning six safety-critical domains: biological safety, chemical safety, radiation hazards, medical & clinical scenarios, personal protective equipment (PPE), and laboratory safety (Sharifi et al., 2021; Hossain et al., 2025; Ghawaly Jr et al., 2020; Tschandl et al., 2018; Pogorelov et al., 2017; Dagli & Shaikh, 2023; Kim et al., 2023; Zieliński et al., 2017), supplemented with curated samples from open scientific repositories and manually collected edge cases.

State-conditioned annotation. For each visual instance, we generate two complementary textual descriptions via LLM prompting that fix the underlying visual semantics while varying only the contextual hazard condition. **Hazardous annotations** ($\mathcal{D}_{\text{forget}}$) encode *actionable but unsafe* procedures, e.g., specific operational details, risk knowledge, or handling instructions that violate safety protocols. In contrast, **safe annotations** ($\mathcal{D}_{\text{retain}}$) provide *conservative, non-actionable* descriptions emphasizing contextual background and institutional safeguards, with explicit guidance to seek professional supervision rather than executable steps. Both types are generated under controlled templates enforcing structural parallelism: hazardous texts foreground procedural steps and operational parameters, while safe texts foreground hazard awareness. This template-based constraint ensures consistent stylistic separation, reducing unintended semantic overlap between $\mathcal{D}_{\text{forget}}$ and $\mathcal{D}_{\text{retain}}$.

Pair construction. The core challenge is constructing *state-dependent pairs* without altering the visual content. We achieve this through **annotation-level pairing**: both hazardous and safe texts are anchored to the *same visual instance*, creating implicit pairs that share identical visual features while differing only in hazard-relevant semantics. This design offers two key advantages over synthetic image generation: (i) it preserves the natural variability and noise of real-world scientific imagery, and (ii) it enables scalable dataset construction without requiring costly visual content manipulation. The resulting pairs approximate the ideal counterfactual of "same scene, different state" at the annotation level, rather than the pixel level.

Evaluation probes ($\mathcal{D}_{\text{test}}$). We construct multiple-choice and fill-in-the-blank questions targeting specific hazardous associations derived from $\mathcal{D}_{\text{forget}}$. These probes test whether residual hazardous knowledge remains accessible without requiring open-ended generation, enabling precise and reproducible measurement of forgetting efficacy.

Quality control and dataset summary. We implement a multi-stage pipeline ensuring both **real-world grounding** and **controlled supervision**. Image-level filtering retains samples with resolution $\geq 512 \times 512$, removes corrupted files, and performs semantic deduplication. Automated watermark detection excludes images with visible artifacts. Annotation-level control enforces template-based generation with clear hazardous/safe distinctions, verified through manual inspection on 10% of samples (1450 instances) for visual-text alignment, hazard state correctness, and absence of artifacts. This combination preserves the natural complexity of real-world scientific imagery while maintaining sufficient controllability for systematic evaluation. Compared to prior synthetic benchmarks, STATEHAZARD-MM captures realistic variability without sacrificing paired supervision quality.

Dataset statistics. The dataset contains about **14,500 visual instances**, each with paired hazardous/safe texts and evaluation probes. Data are split into $\mathcal{D}_{\text{forget}}$ ($\sim$9,700), $\mathcal{D}_{\text{retain}}$ ($\sim$2,400), and $\mathcal{D}_{\text{test}}$ ($\sim$2,400), balanced across domains.

2.3 EVALUATION PROTOCOL

Task formulation. Evaluation is conducted using structured probes derived from hazardous annotations. Given an image x, models are presented with multiple-choice or fill-in-the-blank questions targeting specific hazardous associations. Successfully answering these probes indicates that the corresponding hazardous knowledge remains accessible.

Metrics. We measure unlearning effectiveness using **Attack Success Rate (ASR)** on $\mathcal{D}_{\text{test}}$, defined as the fraction of probes where the model correctly recovers hazardous information. **Lower ASR indicates more effective unlearning and stronger safety.**

3 THE DYNARANK-U ALGORITHM

We present DynaRank-U, a parameter-efficient framework for selective unlearning in multimodal large language models (MLLMs). The key challenge lies in removing hazardous associations while preserving semantically overlapping benign knowledge, both of which are often entangled within shared representations. DynaRank-U addresses this by performing *selective updates in low-rank subspaces*, guided by gradient-based diagnostics.

3.1 RANK-SPACE FORMULATION

We build on low-rank adaptation (LoRA) (Hu et al., 2021), which parameterizes updates to a pretrained weight matrix $\theta \in \mathbb{R}^{d\times d}$ as:

$$\theta' = \theta + \Delta W, \quad \Delta W = BA, \tag{1}$$

where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times d}$, and $r \ll d$.

This formulation decomposes parameter updates into r rank-1 channels. Let $W^{(k)}$ denote the k-th channel, corresponding to the k-th column of B and row of A. Each channel acts as a basis direction capturing a specific mode of variation in the model.

Key observation. In state-dependent unlearning, hazardous and benign knowledge share the same visual grounding but differ in contextual interpretation. As a result, their gradients are *partially aligned but locally conflicting* within the same parameter subspace. Empirically, we observe that: (i) hazard-specific signals concentrate in a subset of ranks, and (ii) interference arises when updates for forgetting conflict with retention objectives. This motivates a *channel-selective update strategy*, where each dimension is modulated based on its role in forgetting and retention.

3.2 GRADIENT DIAGNOSIS

Having established the low-rank decomposition, the next challenge is to determine which channels should be updated. We now describe the gradient diagnostics that enable this selection.

Why two diagnostics? A natural approach to selective unlearning is to update channels with the strongest forget signals. However, amplitude alone is agnostic to *direction*: a channel strongly driven by the forget objective may still serve retention if its forget and retain gradients are aligned. Such alignment in high-dimensional parameter spaces requires a non-negligible retain gradient along the

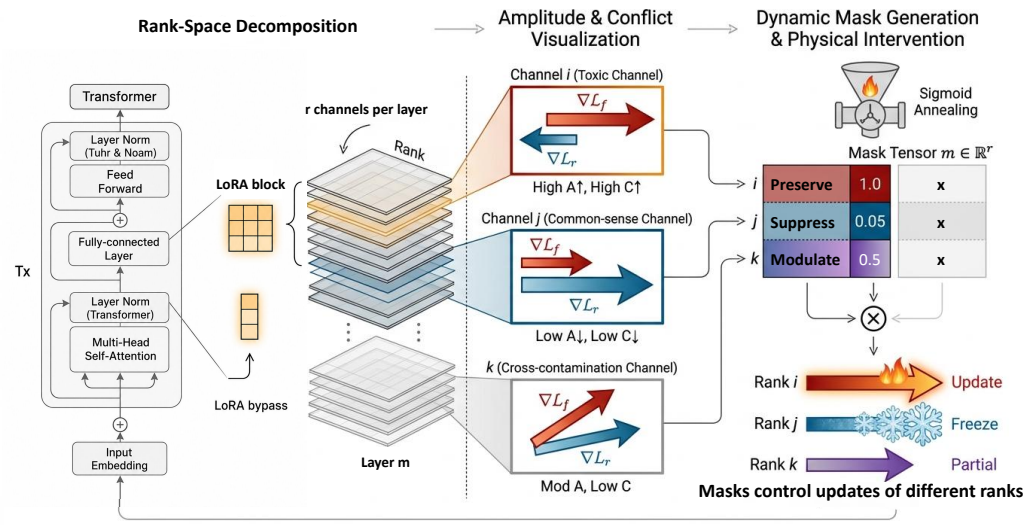


Figure 2: **DynaRank-U framework overview.** (**Left**) Rank-space decomposition: LoRA adapters ($A \in \mathbb{R}^{r\times d}$, $B \in \mathbb{R}^{d\times r}$) decompose updates into r rank-1 channels within each Transformer layer. (**Center**) Gradient diagnosis: For each rank channel, we compute *amplitude* $A^{(k)}$ (relative strength of forget vs. retain signals) and *directional conflict* $C^{(k)}$ (cosine distance between forget and retain gradients). Three prototypical channels emerge: Toxic channels (high A, high C) drive hazardous associations; Common-sense channels (low A, low C) encode shared benign knowledge; Cross-contamination channels (moderate A, low C) require partial modulation. (**Right**) Dynamic mask generation: Scores are converted via sigmoid annealing into continuous masks $\mathbf{m} \in [0,1]^r$, which control gradient updates: preserving benign dimensions, suppressing hazardous dimensions (full update), and modulating ambiguous dimensions (partial update).

forget direction, so aligned channels necessarily carry substantial signals from *both* objectives—they capture shared, general-purpose features (e.g., object recognition, spatial reasoning), and updating them would damage general capabilities. We therefore need a second, directional signal to distinguish *hazard-specific* channels from such *shared* ones.

Amplitude (where to forget). To locate channels where hazardous knowledge is strongly encoded, for each channel k, we compute gradients from forget and retain objectives:

$$g_f^{(k)} = \nabla_{W^{(k)}} \mathcal{L}_{\text{forget}}, \quad g_r^{(k)} = \nabla_{W^{(k)}} \mathcal{L}_{\text{retain}}.$$

Here, $W^{(k)}$ denotes the learnable parameters of the k-th LoRA channel, $\mathcal{L}_{\text{forget}}$ and $\mathcal{L}_{\text{retain}}$ represent the forgetting and retaining loss functions, respectively. The amplitude ratio measures the relative strength of the forget signal via the ℓ_2 norm:

$$A^{(k)} = \frac{\|g_f^{(k)}\|_2 + \epsilon}{\|g_r^{(k)}\|_2 + \epsilon}. \tag{2}$$

where $\epsilon = 10^{-8}$ prevents division by zero when retain gradients are vanishingly small. High $A^{(k)}$ identifies channels primarily driven by hazardous associations (Figure 2, center top).

Directional conflict (what to preserve). To detect shared representations, we measure gradient disagreement between forget and retain objectives:

$$C^{(k)} = \frac{1 - \cos(g_f^{(k)}, g_r^{(k)})}{2}. \tag{3}$$

High $C^{(k)}$ indicates that the two objectives push in opposing directions, signaling a hazard-specific channel that can be safely updated. Conversely, low $C^{(k)}$ indicates aligned gradients—which, by the argument above, occurs only when both objectives strongly activate the channel—revealing shared channels that must be shielded (Figure 2, center middle).

3.3 Dynamic Masking and Training

We convert the above diagnostics into a continuous masking mechanism that modulates gradient updates, as shown in the right panel of Figure 2.

Score aggregation. We define a unified score:

$$S^{(k)} = \alpha \log(A^{(k)}) + \omega C^{(k)}, \tag{4}$$

where α and ω balance importance and conflict.

Adaptive thresholding. Given a target fraction p of ranks to update, we set τ as the $(1-p)$-quantile of $\{S^{(k)}\}$, so that ranks scoring above τ are prioritized for forgetting while the rest are shielded.

Soft masking. Instead of hard selection, we convert scores into continuous mask coefficients $m^{(k)} \in [0, 1]$ that control per-channel gradient intensity:

$$m^{(k)} = \sigma\left(\frac{S^{(k)} - \tau}{T_t}\right), \tag{5}$$

where $m^{(k)} \approx 1.0$ permits full forgetting updates, values near zero suppress updates to benign channels, and intermediate values enable partial modulation. The temperature T_t anneals from high to low during training, gradually sharpening the mask and producing a smooth transition from exploration to focused updates across three regimes: Preserve, Suppress, and Modulate.

Temporal stabilization. To reduce variance, we maintain an exponential moving average (EMA):

$$M_t^{(k)} = \beta M_{t-1}^{(k)} + (1 - \beta) m^{(k)}, \tag{6}$$

and clip the result at a floor $m_{\min} = 0.05$, maintaining minimal gradient flow through shielded channels.

Masked optimization. The final mask $\mathbf{M}_t \in [0, 1]^r$ is applied to the combined gradient:

$$\tilde{G}_t = G_t \odot \mathbf{M}_t, \tag{7}$$

so that shielded channels receive only a small fraction of all updates, enabling selective updates in rank space (Figure 2, right bottom).

Objective. We optimize a combined objective:

$$\mathcal{L} = \lambda_f \mathcal{L}_{\text{forget}} + \lambda_r \mathcal{L}_{\text{retain}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}},$$

where $\mathcal{L}_{\text{KL}}$ regularizes outputs toward the base model to preserve general capabilities.

Summary. DynaRank-U performs *dynamic subspace selection* in low-rank adaptation by jointly considering gradient magnitude and directional conflict. As illustrated in Figure 2, this enables targeted suppression of hazardous knowledge while limiting interference with shared representations.

4 Experiments

We aim to answer three questions central to practical multimodal unlearning: (1) Can DynaRank-U effectively remove hazardous knowledge compared to existing unlearning methods? (2) Does it better preserve general multimodal capabilities and safety alignment? (3) What mechanisms enable its improved trade-off between forgetting and retention?

4.1 Experimental Setup

Models. We evaluate DynaRank-U on 8 vision-language models spanning three representative families: Qwen3-VL (4B/8B/32B), LLaVA1.5 (7B/13B), and InternVL3.5 (4B/8B/14B).

Baselines. We compare against: (1) **BaseModel**, the pretrained checkpoint; (2) **NPO** (Zhang et al., 2024), a strong preference-based unlearning method; and (3) **MMUnlearner** (Huo et al., 2025), implemented as a magnitude-based rank masking strategy (denoted as **Mask-LoRA** in our analysis).

Table 1: Performance comparison across vision-language models and unlearning methods. All models are evaluated on StateHazard-MM (ASR↓), general capabilities, and HarmBench (ASR↓). "P" indicates $> 95\%$ capability preservation relative to the pre-trained checkpoint; "D" indicates significant degradation ($< 90\%$). **Bold** values mark the best result in each column among all unlearning methods for the same model family.

Family	Model	Method	StateHazard-MM ASR↓	General Capability ↑					HarmBench ASR↓
				MMLU	MMMU-Pro	Conv-P	Conv-R	Conv-C	
Qwen3-VL	4B	BaseModel	70.3%	72.6	42.8	18.5	22.9	20.6	49.6%
		NPO	37.1%	68.1	37.0D	13.1D	18.1D	15.4D	62.9%
		Mask-LoRA	**36.6%**	70.1P	40.8P	11.5D	19.5D	18.2D	59.3%
		DynaRank-U	38.9%	**72.7P**	**43.1P**	**14.3D**	**20.7**	**19.4**	**56.8%**
	8B	BaseModel	64.8%	79.8	51.2	26.9	31.1	28.9	45.1%
		NPO	37.0%	70.3D	42.6D	20.9D	26.4D	23.9D	62.2%
		Mask-LoRA	**35.0%**	77.8P	49.5P	19.6D	27.8D	26.9	55.8%
		DynaRank-U	35.2%	**80.0P**	**51.7P**	**23.4D**	**29.1**	**29.1P**	**49.4%**
	32B	BaseModel	58.2%	88.3	62.4	38.5	43.3	40.8	40.2%
		NPO	30.3%	83.8	58.6	33.2D	37.9D	36.0D	52.0%
		Mask-LoRA	24.7%	86.0P	60.5P	31.2D	40.0	38.6	49.3%
		DynaRank-U	**23.1%**	**87.4P**	**62.2P**	**36.3**	**41.1**	**40.0P**	**44.0%**
LLaVA1.5	7B	BaseModel	73.6%	58.5	35.2	14.8	18.6	16.9	51.4%
		NPO	41.1%	53.8	30.1D	9.3D	13.6D	11.4D	64.2%
		Mask-LoRA	39.6%	56.2P	33.5P	7.8D	15.1D	14.3D	61.8%
		DynaRank-U	**38.9%**	**58.7P**	**35.4P**	**10.5D**	**16.3D**	**15.6**	**58.4%**
	13B	BaseModel	68.4%	65.8	42.5	21.5	26.2	23.8	47.1%
		NPO	39.5%	57.6D	35.2D	15.2D	21.5D	18.8D	63.5%
		Mask-LoRA	37.2%	63.8P	40.5P	14.1D	22.8D	21.5	57.7%
		DynaRank-U	**36.7%**	**66.1P**	**42.8P**	**17.3D**	**24.2**	**22.8P**	**52.2%**
InternVL3.5	4B	BaseModel	75.3%	62.5	37.9	16.1	20.6	18.5	53.7%
		NPO	42.4%	58.2	32.5D	10.4D	15.2D	13.2D	66.5%
		Mask-LoRA	40.7%	60.6P	35.5	9.1D	16.8D	16.3D	63.3%
		DynaRank-U	**39.4%**	**62.8P**	**38.1P**	**11.8D**	**18.2D**	**17.3**	**60.6%**
	8B	BaseModel	70.4%	69.2	44.6	23.5	28.2	25.8	49.5%
		NPO	40.4%	60.7D	36.8D	17.3D	23.1D	20.4D	65.7%
		Mask-LoRA	38.4%	67.2P	42.8P	15.8D	24.6D	23.5	59.4%
		DynaRank-U	**37.9%**	**69.5P**	**44.7P**	**19.2D**	**26.1**	**24.8P**	**54.2%**
	14B	BaseModel	63.8%	77.5	52.9	31.2	36.4	33.8	43.6%
		NPO	33.6%	72.8	48.6	25.5D	30.8D	28.2D	56.1%
		Mask-LoRA	28.1%	75.3P	51.2P	23.8D	33.2	31.4	52.4%
		DynaRank-U	**27.5%**	**76.7P**	**52.6P**	**27.1D**	**34.3**	**32.8P**	**47.5%**

Metrics. We evaluate our algorithm along three axes: **(i) Forget Efficacy** via StateHazard-MM ASR (↓); **(ii) General Capability** via MMLU (Hendrycks et al., 2020), MMMU-Pro (Yue et al., 2025), and ConvBench (Conv-P/R/C) (Liu et al., 2024c); and **(iii) Safety Alignment** via HarmBench ASR (↓) (Mazeika et al., 2024). Capability preservation is marked as P ($> 95\%$ of BaseModel) and degradation as D ($< 90\%$ of BaseModel).

4.2 Main Results

(1) Effective Unlearning. DynaRank-U consistently achieves strong forgetting performance across all model families. On Qwen3-VL, ASR decreases from 70.3% to 38.9% (4B) and further to **23.1%** (32B). Similar trends hold for InternVL3.5 (75.3% → 39.4% at 4B; 63.8% → **27.5%** at 14B). This progressive improvement with scale suggests that larger models exhibit stronger *subspace disentanglement* between hazardous and benign associations, enabling DynaRank-U to isolate target dimensions with higher precision. While Mask-LoRA achieves marginally lower ASR at smaller scales(e.g., 36.6% vs. 38.9% on Qwen3-VL-4B), this advantage diminishes as model size increases. We attribute this to the *representation entanglement* phenomenon noted in Section 3.1: smaller models encode hazardous and benign knowledge in more overlapping subspaces, where magnitude-only masking accidentally suppresses shared dimensions. At larger scales, the additional directional conflict signal in DynaRank-U becomes critical for distinguishing truly hazardous channels from high-gradient benign ones.

(2) Superior Capability Preservation. A key advantage of DynaRank-U is its ability to preserve general multimodal capabilities. NPO introduces widespread degradation across reasoning and dialogue tasks. Mask-LoRA consistently harms multi-turn interaction while preserving high-level reasoning. For instance, on Qwen3-VL-32B, Conv-P improves from 31.2 (Mask-LoRA) to **36.3**, and Conv-C is fully preserved (40.0 vs. 40.8 base). Similar recovery trends are observed across all model families(Qwen3-VL-8B: Conv-P 23.4 vs. 19.6; InternVL3.5-14B: Conv-C 32.8 vs. 31.4), indicating that conflict-aware masking effectively protects shared representations essential for reasoning and

Table 2: Efficiency comparison of unlearning methods (Qwen3-VL-32B).

Method	Trainable Params	Backward Passes	Extra Memory	Avg. Epoch Time	Extra Computation
LoRA	$\mathcal{O}(rd)$	$1\times$	Low	12 min	None
Mask-LoRA	$\mathcal{O}(rd)$	$1\times$	Negligible	14 min	Magnitude masking
DynaRank-U	$\mathcal{O}(rd)$	$1\times$ (+2 diag. per 20 steps)	Low	15 min	Gradient conflict + EMA
NPO	Full / LoRA-based	$\geq 2\times$	High	27 min	Preference optimization

dialogue. This pattern, consistently observed in our experiments, validates our central hypothesis in Section 3.2: channels with low directional conflict encode *general-purpose* multimodal capabilities, and shielding them from forget-gradient interference preserves model utility. In contrast, magnitude-only approaches like Mask-LoRA cannot distinguish these high-gradient benign channels from hazardous ones, leading to collateral suppression.

4.3 Failure Case Analysis

Despite its overall effectiveness, DynaRank-U exhibits several consistent failure patterns, which can be partially explained by the empirical results in Table 1.

(1) Residual leakage under weak separation. When hazardous and benign signals are weakly separable, forgetting performance becomes limited. This is reflected in smaller models, where DynaRank-U does not outperform Mask-LoRA on StateHazard-MM (e.g., 38.9% vs. 36.6% on Qwen3-VL-4B, and 35.2% vs. 35.0% on 8B). This suggests that *rank-space separation is less pronounced at smaller scales*, possibly because smaller models rely on more heavily shared representations across hazardous and benign behaviors.

(2) Incomplete recovery of dialogue capability. While DynaRank-U improves ConvBench over Mask-LoRA, it does not fully recover BaseModel performance in many cases. For example, on InternVL3.5-14B, Conv-P remains 27.1 compared to 31.2 in the BaseModel. This indicates that some shared representations remain partially entangled with hazardous subspaces, limiting full recovery.

(3) Retention–forgetting trade-off. Even with conflict-aware masking, strong forgetting is sometimes accompanied by mild degradation in dialogue metrics (e.g., Conv-P consistently below BaseModel across all models). This reflects an inherent limitation: *when hazardous and general knowledge overlap in representation space, perfect disentanglement is not achievable via rank selection alone.*

4.4 Efficiency and Training Cost

DynaRank-U retains the parameter efficiency of LoRA-style adaptation while introducing only lightweight rank-wise statistics during training. Table 2 provides a quantitative comparison against baseline methods on Qwen3-VL-32B.

Parameter efficiency. Like LoRA and Mask-LoRA, DynaRank-U updates only low-rank matrices with $\mathcal{O}(r \cdot d)$ trainable parameters, without modifying backbone weights. The additional masking variables are rank-level scalars and introduce negligible parameter overhead relative to the LoRA matrices themselves.

Computational overhead. The extra computation mainly arises from: (i) computing rank-wise gradient amplitude and directional conflict, and (ii) updating soft masks using EMA smoothing. These operations scale linearly with rank r and are performed intermittently during training. As shown in Table 2, DynaRank-U adds only $\sim$3 minutes per epoch compared to standard LoRA (12min $\rightarrow$ 15min), which is substantially more efficient than NPO (27min) that requires multiple backward passes for preference optimization.

Practical implication. Compared with magnitude-only masking (Mask-LoRA), DynaRank-U introduces additional conflict estimation with minimal overhead (+1min vs. Mask-LoRA, Table 2) but achieves significantly better retention of general capabilities (Section 4.2). Relative to full preference optimization methods such as NPO, DynaRank-U remains substantially lighter-weight (44% faster per epoch) while providing finer control over selective unlearning through rank-space modulation rather than global distribution shifts.

4.5 Mechanistic Analysis

Figure 3 provides a qualitative view of how DynaRank-U operates in rank space.

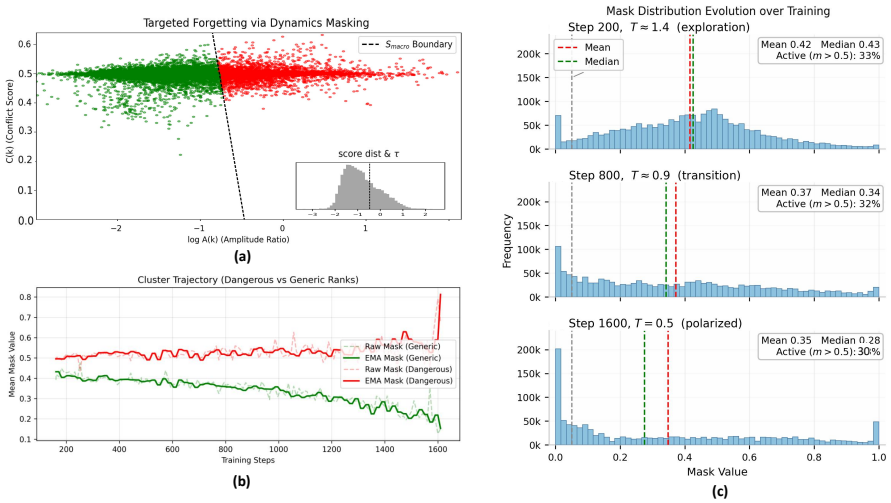


Figure 3: **Mechanistic illustration of DynaRank-U.** (a) *Rank-space visualization.* Each point is a LoRA rank dimension, positioned by Amplitude Ratio ($\log A^{(k)}$) and Directional Conflict ($C^{(k)}$); the threshold τ (inset) selects ranks for update. (b) *Cluster trajectories.* Mean masks of dangerous and generic clusters diverge over training; EMA smoothing (solid) stabilizes the raw estimates (faint). (c) *Mask distribution evolution.* As T_t anneals, the mask distribution polarizes from a broad bell shape into a U-shape: most ranks collapse to the floor $m_{\min} = 0.05$ while $\sim$1/3 remain active.

Rank-space geometry. Figure 3(a) shows the distribution of rank dimensions in the $(\log A^{(k)}, C^{(k)})$ plane. Rather than a clean separation, the points form a *continuous and partially overlapping distribution*. Nevertheless, ranks with higher conflict tend to be more associated with forgetting-driven updates, while many high-amplitude but low-conflict ranks align with retention objectives. This suggests that *directional conflict provides complementary information beyond magnitude*, although the boundary is not strictly separable. Purely magnitude-based selection may thus suppress shared representations, whereas conflict-aware masking offers a more selective yet still approximate criterion.

Temporal dynamics and polarization. Figure 3(b) shows that the mean masks of the two clusters diverge during training: dangerous ranks rise toward full updates while generic ranks decay toward the floor, with EMA smoothing stabilizing the trajectories. Figure 3(c) presents the same process distributionally: as T_t anneals from 1.5 to 0.5, the broad bell-shaped mask distribution polarizes into a U-shape—most ranks collapse to $m_{\min} = 0.05$ as anchors for general knowledge, while $\sim$1/3 are pushed toward 1.0 to absorb forgetting updates, with the active ratio settling at $\approx 30\%$, matching the designed update ratio $p = 0.3$ (Appendix K).

Takeaway. Rather than strict disentanglement, DynaRank-U leverages *approximate geometric structure* in rank space to guide selective updates. This *soft subspace selection* is not perfectly selective, but provides a useful inductive bias that explains the empirical trade-offs in Section 4.2 and the limitations under strong representation overlap.

5 Conclusion

We introduce **StateHazard-MM**, a multimodal benchmark for state-sensitive unlearning built on real scientific and clinical imagery. Through its paired dual-split design across six domains, the benchmark enables evaluation of those state-sensitive multimodal unlearning in scientific and clinical domains. Building on this benchmark, we propose **DynaRank-U**, a rank-space framework that performs selective removal of hazardous knowledge while preserving benign capabilities. By leveraging gradient amplitude and directional conflict within LoRA subspaces, DynaRank-U suppresses hazardous knowledge while minimizing interference with shared representations, achieving a stronger trade-off between forgetting and capability preservation.

Limitations. Our study focuses on static visual scenarios with partially model-generated text annotations. In addition, as hazardous and benign knowledge remain partially entangled in representation space, forgetting and retention cannot be perfectly disentangled (Section 4.3).

AI USE STATEMENT

In this work, we used generative AI tools for one task with required disclosure: *generating synthetic data sets*. Specifically, the structured textual annotations of STATEHAZARD-MM (hazardous descriptions, safe descriptions, and evaluation queries) were produced by GPT-4o with domain-specific prompt templates, on top of real-world images from public datasets (Appendix G). We have not used generative AI tools for cleaning or reformatting datasets, designing or providing feedback on research methodology or experiments, supporting qualitative or thematic data analysis, or interpreting results, and the remaining tasks with required disclosure (developing theoretical models or conceptual frameworks, formulating mathematical claims, assisting with proofs, proposing or refining hypotheses, implementing methods, and translation) are not applicable to this work. Additionally, we have not used generative AI tools for any task with recommended disclosure: the manuscript text, figures, and code are entirely author-written. We have reviewed all AI-assisted work. Every generated annotation passed automated schema validation, and three annotators with AI-for-Science domain expertise manually reviewed a 10% subset (first-round pass rate 94.2%; failed samples were re-annotated or discarded); all evaluation probes in $\mathcal{D}_{\text{test}}$ were additionally verified by hand for answer-key correctness, visual–text alignment, and hazard-state correctness (Appendix G). We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or crowdsourced labor; all source imagery comes from public datasets, and manual review was carried out by domain experts under a written guideline (Appendix G). The primary ethical consideration is dual use: STATEHAZARD-MM contains hazardous scenarios in chemical, biological, radiation, and medical domains, and selective-unlearning techniques could in principle be misused to weaken safety behaviors in deployed models. We mitigate these risks through the release policy rather than through secrecy of results: the full hazardous descriptions are distributed only via gated access to affiliated researchers under a CC BY-NC 4.0 license with an acceptable-use clause, particularly sensitive generation targets are redacted, the evaluation probes and code needed to reproduce all reported metrics are openly available, and the dataset undergoes an internal red-team review before public release. A full discussion of broader impacts, misuse scenarios, and the limitations of these safeguards appears in Appendix L and L.3.

REPRODUCIBILITY STATEMENT

We provide the following resources to support reproduction of our results. Section 2 and Appendix G describe the benchmark's data sources, annotation pipeline, quality control, and split statistics; Section 3 and Appendix I give the complete mathematical derivation of DynaRank-U, including the exact tensor-level masking operations; Appendix A lists all hyperparameters and per-model training costs, and Appendix M the hardware and software configuration. The evaluation protocol and metrics are defined in Section 2.3 and Section 4.1, with benchmark validity controls reported in Appendix B. Source code, the evaluation probes $\mathcal{D}_{\text{test}}$, and all trained LoRA adapter configurations are available in an anonymous repository at `https://anonymous.4open.science/r/StateHazard-MM-235B`, and will be moved to a public repository upon publication.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Rishit Dagli and Ali Mustufa Shaikh. Cppe-5: Medical personal protective equipment dataset. *SN Computer Science*, 4(3):263, 2023.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.

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

James M Ghawaly Jr, Andrew D Nicholson, Douglas E Peplow, Christine M Anderson-Cook, Kary L Myers, Daniel E Archer, Michael J Willis, and Brian J Quiter. Data for training and testing radiation detection algorithms in an urban environment. *Scientific data*, 7(1):328, 2020.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Md Sakhawat Hossain, Md Sadman Haque, Md Mostafizur Rahman, Md Mosaddik Mashrafi Mousum, Zobaer Ibn Razzaque, Robiul Awoul Robin, Raiyan Rahman, and Jannatun Noor. Real-world chemistry lab image dataset for equipment recognition across 25 apparatus categories. *Scientific Data*, 12(1):1747, 2025.

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. *arXiv preprint arXiv:2106.09685*, 2021.

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.

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 Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic acids research*, 51(D1): D1373–D1380, 2023.

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.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024b.

Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, et al. Convbench: A multi-turn conversation evaluation benchmark with hierarchical ablation capability for large vision-language models. *Advances in Neural Information Processing Systems*, 37:100734–100782, 2024c.

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, 2025.

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.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Trans. Intell. Syst. Technol.*, 16(5), September 2025. ISSN 2157-6904. doi: 10.1145/3749987. URL https://doi.org/10.1145/3749987.

Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 164–169, 2017.

Amir Sharifi, Ahmadreza Zibaei, and Mahdi Rezaei. A deep learning based hazardous materials (hazmat) sign detection robot with restricted computational resources. *Machine Learning with Applications*, 6:100104, 2021.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations*, volume 2025, pp. 27797–27818, 2025.

Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX security symposium (USENIX Security 22)*, pp. 4007–4022, 2022.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 180161, 2018.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Tianyu Yang, Lisen Dai, Xiangqi Wang, Minhao Cheng, Yapeng Tian, and Xiangliang Zhang. Cliperase: Efficient unlearning of visual-textual associations in clip. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30438–30452, 2025.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

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

Bartosz Zieliński, Anna Plichta, Krzysztof Misztal, Przemysław Spurek, Monika Brzychczy-Włoch, and Dorota Ochońska. Deep learning approach to bacterial colony classification. *PloS one*, 12(9): e0184554, 2017.

A Implementation Details and Hyperparameters

A.1 Vision-Language Formulation and Prompt Masking

For large vision-language models (e.g., Qwen3-VL-8B and 32B), the input sequences consist of interleaved image tokens, user prompts (questions), and model responses (answers). During all unlearning and retaining update steps, we enforce an "Answer-Only" loss masking mechanism (`ans_only = True` in our codebase). Specifically, the calculation of both $\mathcal{L}_{\text{forget}}$ and $\mathcal{L}_{\text{retain}}$ strictly zeroes out the sequence positions corresponding to the visual inputs and user questions, applying backpropagation exclusively to the generated answer tokens. This prevents the model from trivially minimizing the loss by dismantling its image encoder or prompt-understanding templates.

A.2 Training Objective and Gradient Computation

DynaRank-U optimizes

$$\mathcal{L}_{\text{total}} = \lambda_f \mathcal{L}_{\text{forget}} + \lambda_r \mathcal{L}_{\text{retain}} + \lambda_{\text{KL}} \mathcal{D}_{\text{KL}}(p_{\theta'} \| p_{\theta_{\text{base}}}), \tag{8}$$

where $\mathcal{L}_{\text{forget}}$ is gradient ascent (negative cross-entropy) on the forget set, $\mathcal{L}_{\text{retain}}$ is standard cross-entropy on the retain set, and $\mathcal{D}_{\text{KL}}$ anchors the model to the frozen base model on a mixed stream of retain and general samples. At each step we sample batches from $\mathcal{D}_{\text{forget}}$ and $\mathcal{D}_{\text{retain}}$, compute both losses, and backpropagate the combined objective once. To obtain per-channel gradients $g_f^{(k)}$ and $g_r^{(k)}$ via the LoRA outer-product structure, we perform one additional pair of separate backward passes every $\Delta t = 20$ steps (after a 150-step warmup), and then recompute $A^{(k)}$, $C^{(k)}$, the threshold τ, and the EMA-smoothed mask $\mathbf{M}_t$, which modulates the combined gradient before the AdamW update. Full derivations are given in Appendix I.

A.3 Hyperparameter Configurations

We present the comprehensive list of hyperparameters used to train our DynaRank-U algorithm in Table 3. To ensure parameter efficiency and stability across large-scale models, all updates are restricted to Low-Rank Adaptation (LoRA) matrices.

Table 3: Hyperparameter configurations for DynaRank-U.

LoRA Settings	**Value**	**DynaRank Parameters**	**Value**
LoRA Rank (r)	16	Mask Update Interval	20 steps
LoRA Alpha	16	Warmup Steps	150 steps
LoRA Dropout	0.05	Update Ratio (p)	0.3
Targets	All Linear Layers	Initial Temp. (T_{start})	1.5
Learning Rate (η)	1e-5	Final Temp. (T_{end})	0.5
Batch Size	4	Mask Floor ($m_{\min}$)	0.05
Epochs	10	EMA Momentum (β)	0.90
Optimization Weights	**Value**	**Macro Control**	**Value**
Mask Amplitude Weight (α_{mask})	1.0	KL Div Weight (λ_{KL})	1.0
Mask Conflict Weight (ω_{conflict})	0.6	Optimizer	AdamW
Forget Weight (λ_f)	1.3	Retain Weight (λ_r)	0.2–1.0 (per domain)

A.4 Per-Model Training Cost

Table 4 reports wall-clock training cost for all checkpoints on NVIDIA H200 (140GB) GPUs. DynaRank-U introduces modest overhead versus standard LoRA (+3 min/epoch on 32B, +1 min on smaller models), since rank-wise statistics scale linearly with r and are computed only every 20 steps after a 150-step warmup; it is substantially cheaper than NPO, which requires multiple backward passes *per step* (27 min/epoch on 32B).

Table 4: Training cost per model (10 epochs).

Model	GPUs	Avg. Epoch Time	Total (10 epochs)
Qwen3-VL-4B	1 × H200	∼12 min	∼2.0 h
Qwen3-VL-8B	1 × H200	∼15 min	∼2.5 h
Qwen3-VL-32B	4 × H200 (TP)	∼15 min	∼2.5 h
LLaVA1.5-7B	1 × H200	∼14 min	∼2.3 h
LLaVA1.5-13B	1 × H200	∼18 min	∼3.0 h
InternVL3.5-4B	1 × H200	∼13 min	∼2.2 h
InternVL3.5-8B	1 × H200	∼16 min	∼2.7 h
InternVL3.5-14B	1 × H200	∼22 min	∼3.7 h

B Benchmark Validity Controls

Two validity concerns are intrinsic to our paired design: (i) models could separate the splits from *stylistic cues* (actionable/procedural vs. advisory) rather than hazard semantics, and (ii) probe answering could be driven by text alone, without visual grounding. This appendix addresses both with control experiments.

B.1 Style-Swapped Control Experiment

We construct a permuted variant $\mathcal{D}_{\text{retain}}^{\text{swap}}$ in which all safe content is rewritten in procedural/imperative style and all hazardous content in advisory/cautionary style, and retrain DynaRank-U from scratch. If the model relied on stylistic shortcuts (e.g., suppressing all imperative text), this swap should cause catastrophic failure—suppressing safe procedural content while retaining hazardous advisory content—producing a large ASR increase. Table 5 shows the opposite: across all eight checkpoints, ASR changes by at most 0.2 percentage points, well within run-to-run variance. Forgetting is therefore driven by state-sensitive semantics, not formatting cues.

Table 5: **Style-swapped control experiment.** StateHazard-MM ASR (↓) for DynaRank-U trained on the original retain set versus a style-permuted variant ($\mathcal{D}_{\text{retain}}^{\text{swap}}$). Consistent ASR across permutations confirms semantic-driven unlearning.

Family	Model	Original Style	Swapped Style	Δ
Qwen3-VL	4B	38.9%	39.1%	+0.2
Qwen3-VL	8B	35.2%	35.4%	+0.2
Qwen3-VL	32B	23.1%	23.3%	+0.2
LLaVA1.5	7B	38.9%	39.0%	+0.1
LLaVA1.5	13B	36.7%	36.9%	+0.2
InternVL3.5	4B	39.4%	39.6%	+0.2
InternVL3.5	8B	37.9%	38.1%	+0.2
InternVL3.5	14B	27.5%	27.7%	+0.2

B.2 Visual-Dependency Ablation

To quantify how much probe answering depends on the image, we re-evaluate all models with the image replaced by a blank canvas while keeping the text prompt unchanged (text-only condition). Table 6 shows that base models drop by 23–30 percentage points, confirming that $\mathcal{D}_{\text{test}}$ probes are strongly visually grounded; DynaRank-U still drops by 8.0–14.2 points, indicating that its residual hazardous responses are also visually conditioned rather than purely stylistic.

Table 6: **Visual-dependency ablation.** StateHazard-MM ASR with full multimodal input versus text-only input (image masked). Large drops confirm that probe answering requires visual grounding.

Family	Model	Method	Full (image + text)	Text-only	Drop
Qwen3-VL	4B	BaseModel	70.3%	39.9%	−30.4
Qwen3-VL	4B	DynaRank-U	38.9%	26.6%	−12.3
Qwen3-VL	8B	BaseModel	64.8%	40.3%	−24.5
Qwen3-VL	8B	DynaRank-U	35.2%	23.4%	−11.8
Qwen3-VL	32B	BaseModel	58.2%	35.1%	−23.1
Qwen3-VL	32B	DynaRank-U	23.1%	15.1%	−8.0
LLaVA1.5	7B	BaseModel	73.6%	47.5%	−26.1
LLaVA1.5	7B	DynaRank-U	38.9%	27.2%	−11.7
LLaVA1.5	13B	BaseModel	68.4%	38.1%	−30.3
LLaVA1.5	13B	DynaRank-U	36.7%	24.8%	−11.9
InternVL3.5	4B	BaseModel	75.3%	47.6%	−27.7
InternVL3.5	4B	DynaRank-U	39.4%	27.3%	−12.1
InternVL3.5	8B	BaseModel	70.4%	41.4%	−29.0
InternVL3.5	8B	DynaRank-U	37.9%	23.7%	−14.2
InternVL3.5	14B	BaseModel	63.8%	35.2%	−28.6
InternVL3.5	14B	DynaRank-U	27.5%	18.4%	−9.1

C Safety Analysis: HarmBench Degradation Is a Response-Format Artifact

A counter-intuitive pattern in Table 1 is that HarmBench ASR *rises* relative to the BaseModel for all unlearning methods. This appendix analyzes its root cause and shows that it reflects response-format overfitting rather than an erosion of safety alignment.

Failure mode. Our $\mathcal{D}_{\text{retain}}$ annotations adopt a conservative "explain-then-defer" format: they provide contextual background and guide users toward professional supervision rather than issuing terse refusals. During unlearning, models partially internalize this format as a safety template. On HarmBench probes—particularly misinformation and manipulation categories—this produces detailed explanatory refutations instead of direct refusals, and HarmBench's classifier interprets any non-refusal engagement as a successful attack, even when the content is factually corrective. For example, given a misleading climate graph, DynaRank-U responds with a detailed, accurate debunking (flagged unsafe), whereas the BaseModel issues a terse refusal (flagged safe). This reflects a dual limitation: at the method level, unlearning objectives optimized on paired text targets can conflate safety *content* with safety *response format*; at the data level, our retain-set template creates a stylistic signature that generalizes beyond the intended domains.

Response-format ablation. To test this hypothesis directly, we re-evaluate all models on HarmBench with an explicit system prompt: *"Respond with a brief refusal only. Do not provide explanations."* Table 7 shows that the prompt leaves the BaseModel virtually unchanged (it already issues terse refusals natively), while DynaRank-U's ASR drops sharply across all families and sizes, converging to within 0.4–0.8 percentage points of the corresponding BaseModel. The safety "degradation" is thus a formatting artifact: DynaRank-U retains the underlying safety knowledge but expresses it through the elaborate template inherited from $\mathcal{D}_{\text{retain}}$.

Relative safety comparison. Even without the corrective prompt, DynaRank-U's HarmBench ASR remains substantially lower than the baselines on every model (e.g., on Qwen3-VL-32B: 44.0% vs. 52.0% for NPO and 49.3% for Mask-LoRA). Conflict-aware masking therefore mitigates the format-overfitting effect by better protecting shared safety-aligned representations.

Table 7: **Response-format ablation on HarmBench.** All values are HarmBench ASR (↓). Prepending "Respond with a brief refusal only" leaves the BaseModel unaffected but restores DynaRank-U to the BaseModel level, confirming the format-overfitting explanation.

Family	Model	Method	Original	+ Brief-Refusal Prompt
Qwen3-VL	4B	BaseModel	49.6%	49.4%
Qwen3-VL	4B	DynaRank-U	56.8%	50.1%
Qwen3-VL	8B	BaseModel	45.1%	45.0%
Qwen3-VL	8B	DynaRank-U	49.4%	45.8%
Qwen3-VL	32B	BaseModel	40.2%	40.2%
Qwen3-VL	32B	DynaRank-U	44.0%	40.8%
LLaVA1.5	7B	BaseModel	51.4%	51.3%
LLaVA1.5	7B	DynaRank-U	58.4%	52.0%
LLaVA1.5	13B	BaseModel	47.1%	46.9%
LLaVA1.5	13B	DynaRank-U	52.2%	47.8%
InternVL3.5	4B	BaseModel	53.7%	53.5%
InternVL3.5	4B	DynaRank-U	60.6%	54.2%
InternVL3.5	8B	BaseModel	49.5%	49.4%
InternVL3.5	8B	DynaRank-U	54.2%	49.8%
InternVL3.5	14B	BaseModel	43.6%	43.5%
InternVL3.5	14B	DynaRank-U	47.5%	44.1%

D Component Ablations and Composite Retention

Composite retention score. To summarize capability retention across heterogeneous metrics, we define a composite retention score as the average relative retention across all five capability metrics:

$$\text{Retention} = \frac{1}{5} \sum_{m \in \{\text{MMLU, MMMU-Pro, Conv-P, Conv-R, Conv-C}\}} \frac{\text{score}_m}{\text{score}_m^{\text{base}}} \times 100\%. \tag{9}$$

Unlike the P/D criterion in the main text, this aggregate explicitly includes dialogue metrics, avoiding an overly optimistic picture when conversational capabilities degrade.

Component ablations. Table 8 isolates the contribution of each DynaRank-U component on Qwen3-VL-32B. Two findings stand out. (i) Removing the conflict term ($\omega = 0$, i.e., Mask-LoRA) preserves forgetting but costs $+5.1$ points of Conv-P degradation, confirming that directional conflict is responsible for the retention gain. (ii) Removing EMA ($\beta = 0$) leaves forgetting intact (+0.2 pp ASR) but consistently degrades capabilities and safety (0.6–1.1 points), validating EMA as a stabilizer.

Table 8: **Component ablations on Qwen3-VL-32B.** Full DynaRank-U versus single-component removals.

Variant	ASR↓	MMLU	MMMU-Pro	Conv-P	Conv-C	HarmBench↓
BaseModel	58.2%	88.3	62.4	38.5	40.8	40.2%
Full DynaRank-U	23.1%	87.4	62.2	36.3	40.0	44.0%
w/o Conflict (ω=0, Mask-LoRA)	24.7%	86.0	60.5	31.2	38.6	49.3%
w/o EMA (β=0)	23.3%	86.8	61.5	35.4	39.2	45.1%

E Validating the Gradient-Conflict Assumption

Our method rests on the assumption that hazardous and benign gradients are globally aligned but locally conflicting within the rank subspace (Section 3.1). Figure 3 visualizes this qualitatively; this appendix provides direct causal and quantitative evidence.

Rank-isolation ablations. We freeze all LoRA dimensions and train only selected rank subsets on Qwen3-VL-32B. Training only *high-mask* ranks ($m^{(k)} \approx 1$, the "toxic channels") achieves near-complete forgetting (24.8% ASR vs. 23.1% for the full method) with moderate capability cost (Table 9), confirming that these ranks encode hazard-specific knowledge. Conversely, training only *low-mask* ranks ($m^{(k)} \approx 0.05$) leaves ASR nearly intact (55.4%) while devastating general capabilities (-8.8 MMLU, -11.3 Conv-P), confirming that these ranks encode shared benign knowledge that must be protected. This clean double dissociation causally validates that our gradient diagnostics identify functionally distinct rank groups. The mild capability drop under high-mask-only training (3–5 points) is expected—even toxic channels weakly overlap with general reasoning subspaces at r=16—and is far smaller than under low-mask-only training ($\sim$10 points) or Mask-LoRA.

Table 9: **Rank-isolation ablations (Qwen3-VL-32B).** Training only high-mask ranks reproduces full forgetting; training only low-mask ranks destroys capabilities without forgetting. Avg. retention is the composite score of Eq. 9.

Condition	ASR↓	MMLU	MMMU-Pro	Conv-P	Conv-R	Conv-C	Avg. Retention
BaseModel	58.2%	88.3	62.4	38.5	43.3	40.8	100%
Full DynaRank-U	23.1%	87.4	62.2	36.3	41.1	40.0	97.2%
High-Mask Only	24.8%	85.1	59.5	33.8	38.6	37.4	92.1%
Low-Mask Only	55.4%	79.5	52.1	27.2	33.5	30.6	79.3%

F Qualitative Example: Quantity versus Quality of Forgetting

The practical difference between "quantity" and "quality" of forgetting is best seen at the example level. On an MMMU-Pro chemistry item ("Analyze the reaction pathway shown in the figure and identify the rate-limiting step"), the BaseModel correctly identifies the rate-limiting step and explains the electron transfer. NPO refuses to answer entirely ("I cannot provide analysis of chemical reactions that may be unsafe…"), consistent with its degraded MMMU-Pro score of 37.0 on Qwen3-VL-4B. Mask-LoRA provides a fragmented answer that misidentifies the intermediate compound (40.8). DynaRank-U provides the correct, complete analysis (43.1, 100.7% of BaseModel). This pattern is systematic rather than anecdotal: it is the behavioral signature of the aggregate gaps in Table 1.

G Dataset Construction Details

G.1 Image Sources and Domain Coverage

StateHazard-MM is grounded in real-world imagery aggregated from diverse publicly available datasets. Table 10 summarizes the primary data sources for each domain.

G.2 LLM-Assisted Annotation Protocol

For each image, we construct structured textual annotations using a large language model (GPT-4o) with domain-specific prompt templates. The annotation pipeline produces three types of supervision signals:

1. **Hazardous descriptions ($\mathcal{D}_{\text{forget}}$):** Detailed captions describing unsafe states, improper handling, or dangerous conditions.
2. **Safe descriptions ($\mathcal{D}_{\text{retain}}$):** Factually correct descriptions of safe counterparts or compliant procedures.
3. **Evaluation queries ($\mathcal{D}_{\text{test}}$):** Adversarial prompts designed to elicit hazardous knowledge for ASR measurement.

To ensure consistency, we employ a two-stage verification process: (i) automated schema validation checking for required fields and format compliance, and (ii) manual spot-checking by domain experts on 10% of samples.

Table 10: Data sources for each domain in StateHazard-MM.

Domain	Primary Sources	Data Type
Chemical Safety	HAZMAT-13 (Sharifi et al., 2021); Chemistry Lab Dataset (Hossain et al., 2025); PubChem chemical structures (Kim et al., 2023)	Hazard symbols, chemical apparatus, molecular structures
Biological Safety	Biohazard Detection Dataset (Roboflow Universe); Biomedical Waste Classification Dataset; DIBaS microbial microscopy dataset (Zieliński et al., 2017)	Biohazard signs, medical waste, microbial microscopy
Radiation Hazards	HAZMAT-13 (Sharifi et al., 2021); Urban radiation monitoring dataset (Ghawaly Jr et al., 2020)	Radiation warning signs, detector measurements, urban monitoring scenes
Medical & Clinical	HAM10000 (Tschandl et al., 2018); Kvasir (Pogorelov et al., 2017); Wikimedia Commons medical imagery	Dermatoscopy, gastrointestinal imaging, anatomical illustrations
PPE	CPPE-5 (Dagli & Shaikh, 2023); Chemistry Lab Dataset (Hossain et al., 2025)	Protective equipment, laboratory personnel, safety compliance
Laboratory Safety	Chemistry Lab Dataset (Hossain et al., 2025); Biohazard Detection Dataset; HAZMAT-13 (Sharifi et al., 2021)	Laboratory apparatus, hazardous materials, biohazard environments

G.3 Manual Review Protocol

Manual inspection was conducted by **three annotators with AI-for-Science domain expertise**, all following a written annotation guideline. Each sampled instance was evaluated on three criteria: (i) *visual–text alignment*, (ii) *hazard-state correctness*, and (iii) *naturalness* (absence of template artifacts). The first-round pass rate on the 10% subset (1450 instances) was 94.2%; failed samples were re-annotated or discarded. For $\mathcal{D}_{\text{test}}$, all evaluation probes were manually verified along three dimensions: (i) answer-key correctness, (ii) visual–text alignment, and (iii) hazard-state correctness, ensuring that ASR measurements reflect hazardous-knowledge recovery rather than annotation noise.

G.4 Pair Construction and Quality Control

Rather than relying on synthetic image generation, we form implicit pairs at the annotation level. Hazardous and benign supervision signals are aligned to the same visual instance from the same distribution. Evaluation probes are constructed exclusively from the hazardous annotations of held-out $\mathcal{D}_{\text{test}}$ instances, which are disjoint from both training splits at the image level. This allows us to approximate state-dependent variation while retaining real-world visual complexity.

Quality control measures include:

- Resolution filtering ($\geq 512 \times 512$ pixels)
- Duplicate detection via perceptual hashing
- Manual inspection of edge cases (e.g., ambiguous hazard states)
- Cross-validation of annotation consistency across annotators

The final dataset contains over 14,000 samples, with a 4:1:1 split across forget, retain, and test sets, balanced across domains.

H Related Work

Unlearning in Large Language Models. Machine unlearning aims to remove specific knowledge from trained models without full retraining (Nguyen et al., 2025). In large language models, approximate methods based on gradient optimization have become the dominant approach (Thudi et al., 2022). Gradient ascent (GA) directly maximizes the loss on forget sets but often leads to substantial degradation of general capabilities (Jang et al., 2023). Preference-based methods such as NPO (Zhang et al., 2024) improve stability and have shown effectiveness on text-centric benchmarks such as TOFU (Maini et al., 2024) and MUSE (Shi et al., 2025). However, these approaches are primarily designed for textual data and do not explicitly account for multimodal representations or fine-grained visual context.

Multimodal Unlearning in MLLMs. Extending unlearning to multimodal models introduces additional challenges due to the coupling between visual and language representations. Prior work has largely focused on removing coarse-grained visual concepts in models such as CLIP or diffusion systems (Gandikota et al., 2023; Yang et al., 2025), or adapting unlearning objectives to multimodal large language models (MLLMs) (Huo et al., 2025; Li et al., 2024a; Liu et al., 2025). These methods typically operate at the level of objects or high-level concepts, where harmful and benign content are separable. In contrast, our work focuses on *state-dependent* hazards, where hazardous and benign instances share highly similar visual features and differ only in subtle contextual conditions. This setting introduces strong representation overlap, making selective unlearning significantly more challenging and requiring more fine-grained control over model updates.

Unlearning Benchmarks and Datasets. Existing benchmarks for unlearning and safety evaluation are predominantly text-centric (Maini et al., 2024; Li et al., 2024b) or rely on synthetic visual data with controlled concept edits. While such designs enable clean experimental setups, they often fail to capture the variability, noise, and contextual ambiguity present in real-world scientific and clinical imagery. Moreover, prior benchmarks lack *paired supervision* that explicitly evaluates whether models can forget hazardous content while preserving semantically equivalent benign knowledge.

To address these limitations, we introduce **StateHazard-MM**, a multimodal benchmark grounded in real-world scientific data with a paired dual-split design for state-sensitive unlearning. By aligning hazardous samples for forgetting with matched benign counterparts for retention, StateHazard-MM provides a more realistic and fine-grained evaluation setting that complements existing synthetic and text-based benchmarks.

Parameter-Efficient Unlearning. Parameter-efficient methods such as LoRA (Hu et al., 2021) enable targeted updates in low-rank subspaces and are widely used for scalable adaptation. Existing approaches often rely on magnitude-based selection or static masking (Geng & Li, 2025), which may fail to distinguish between shared and task-specific representations. Our method builds on this line of work by incorporating both gradient amplitude and directional conflict to enable more precise rank-space control for selective multimodal unlearning.

I Extended Mathematical Derivation of DynaRank-U

I.1 Low-Rank Subspace Geometry and Knowledge Localization

Preliminaries. Let $\theta \in \mathbb{R}^d$ denote the full parameters of a pretrained MLLM. In LoRA fine-tuning, the update is restricted to a low-rank decomposition:

$$\Delta W = BA, \quad B \in \mathbb{R}^{d_{\text{out}} \times r},\ A \in \mathbb{R}^{r \times d_{\text{in}}}, \tag{10}$$

where $r \ll \min(d_{\text{out}}, d_{\text{in}})$ is the adaptation rank. The forward pass through an adapted linear layer becomes:

$$h = W_0 x + \Delta W x = W_0 x + BAx, \tag{11}$$

with W_0 frozen. DynaRank-U operates entirely within the rank-r subspace spanned by the columns of B (or equivalently the rows of A).

Rank-wise Gradient Decomposition. Consider the total loss $\mathcal{L}_{\text{total}}$ at training step t. The gradient with respect to the adapter parameters is:

$$G_t = \nabla_{\Delta W}\mathcal{L}_{\text{total}} = \lambda_f \nabla_{\Delta W}\mathcal{L}_{\text{forget}} + \lambda_r \nabla_{\Delta W}\mathcal{L}_{\text{retain}} + \lambda_{\text{KL}} \nabla_{\Delta W}\mathcal{D}_{\text{KL}}. \tag{12}$$

Because ΔW is factorized, the gradient naturally decomposes via the chain rule. For matrix B, we have:

$$\nabla_B \mathcal{L} = (\nabla_{\Delta W}\mathcal{L})A^\top, \qquad \nabla_A \mathcal{L} = B^\top(\nabla_{\Delta W}\mathcal{L}). \tag{13}$$

However, DynaRank-U does not mask at the full matrix level. Instead, we seek to characterize the *contribution of each rank channel* $k \in \{1, \ldots, r\}$. Let $b_k \in \mathbb{R}^{d_{\text{out}}}$ and $a_k \in \mathbb{R}^{d_{\text{in}}}$ denote the k-th column of B and k-th row of A, respectively. Then:

$$\Delta W = \sum_{k=1}^{r} b_k a_k^\top. \tag{14}$$

Consequently, the gradient of the loss with respect to rank channel k can be isolated via the outer-product structure:

$$\nabla_{b_k}\mathcal{L} = (\nabla_{\Delta W}\mathcal{L})a_k^\top, \qquad \nabla_{a_k}\mathcal{L} = b_k^\top(\nabla_{\Delta W}\mathcal{L}). \tag{15}$$

The *effective gradient* for rank k, denoted $\nabla_{W^{(k)}}\mathcal{L}$, is the concatenation of the flattened $\nabla_{b_k}\mathcal{L}$ and $\nabla_{a_k}\mathcal{L}$, living in $\mathbb{R}^{d_{\text{out}}+d_{\text{in}}}$.

Amplitude Ratio: Derivation and Interpretation. For each rank k, we compute the L_2 norms of the forget and retain gradients:

$$v_f^{(k)} = \|\nabla_{W^{(k)}}\mathcal{L}_{\text{forget}}\|_2, \qquad v_r^{(k)} = \|\nabla_{W^{(k)}}\mathcal{L}_{\text{retain}}\|_2. \tag{16}$$

The amplitude ratio is then:

$$A^{(k)} = \frac{v_f^{(k)} + \epsilon}{v_r^{(k)} + \epsilon}, \tag{17}$$

where $\epsilon = 10^{-8}$ prevents division by zero. Taking the logarithm yields:

$$\log A^{(k)} = \log(v_f^{(k)} + \epsilon) - \log(v_r^{(k)} + \epsilon), \tag{18}$$

which transforms the multiplicative ratio into an additive symmetric measure. This is crucial because it ensures that ranks where $v_f^{(k)} \gg v_r^{(k)}$ receive large positive scores, while ranks dominated by retention signals receive negative scores, and ranks with balanced activation cluster near zero.

Directional Conflict: Cosine Disagreement. The directional conflict measures the geometric alignment between forget and retain gradients in the rank subspace:

$$\cos\theta^{(k)} = \frac{\langle \nabla_{W^{(k)}}\mathcal{L}_{\text{forget}},\ \nabla_{W^{(k)}}\mathcal{L}_{\text{retain}} \rangle}{v_f^{(k)} \cdot v_r^{(k)}}. \tag{19}$$

We map this to the range $[0, 1]$ via:

$$C^{(k)} = \frac{1 - \cos\theta^{(k)}}{2}. \tag{20}$$

This formulation has the following properties:

- If $\nabla_{W^{(k)}}\mathcal{L}_{\text{forget}}$ and $\nabla_{W^{(k)}}\mathcal{L}_{\text{retain}}$ are perfectly aligned ($\cos\theta^{(k)} = 1$), then $C^{(k)} = 0$. Updating this rank would jointly optimize both objectives, so it should be *protected*.
- If they are perfectly opposed ($\cos\theta^{(k)} = -1$), then $C^{(k)} = 1$. This indicates pure conflict: the rank encodes hazardous associations that are orthogonal (or opposite) to benign knowledge, making it an ideal candidate for targeted forgetting.
- Orthogonal gradients ($\cos\theta^{(k)} = 0$) yield $C^{(k)} = 0.5$, representing uncoupled dimensions where forget signals operate independently of retain signals.

I.2 Unified Scoring and Adaptive Thresholding

Score Composition. The unified rank score combines amplitude and conflict with learnable (or preset) weights:

$$S^{(k)} = \alpha_{\text{mask}} \log(A^{(k)}) + \omega_{\text{conflict}} C^{(k)}. \tag{21}$$

In our implementation, we set $\alpha_{\text{mask}} = 1.0$ and $\omega_{\text{conflict}} = 0.6$. The logarithmic amplitude ensures that the score is sensitive to relative differences rather than absolute gradient magnitudes, which stabilizes training across layers with heterogeneous activation scales.

Quantile-Based Thresholding. Given a target update ratio $p \in (0, 1)$, we compute the threshold τ as the $(1-p)$-th quantile of the empirical distribution $\{S^{(k)}\}_{k=1}^{r}$:

$$\tau = Q_{1-p}\left(\{S^{(k)}\}_{k=1}^{r}\right), \tag{22}$$

where Q denotes the quantile function. This adaptive thresholding mechanism ensures that approximately a fraction p of ranks will be targeted for forgetting, while the remaining $1-p$ fraction is shielded. Unlike static magnitude pruning, this quantile adapts to the dynamic landscape of gradient statistics throughout training.

Centered Score and Soft Mask. The centered score is defined as:

$$\tilde{S}^{(k)} = S^{(k)} - \tau. \tag{23}$$

We convert this into a soft mask via a temperature-scaled sigmoid:

$$m^{(k)} = \sigma\left(\frac{\tilde{S}^{(k)}}{T_t}\right), \quad \text{where } \sigma(x) = \frac{1}{1+e^{-x}}. \tag{24}$$

Here, T_t is the annealing temperature at step t:

$$T_t = T_{\text{start}} + (T_{\text{end}} - T_{\text{start}}) \cdot \frac{t}{T_{\text{max}}}, \tag{25}$$

with $T_{\text{start}} = 1.5$ and $T_{\text{end}} = 0.5$ in our experiments (see Table 3). As $T_t \to 0$, the sigmoid sharpens toward a step function, driving masks toward binary $\{0, 1\}$ decisions.

I.3 Exponential Moving Average and Floor Clipping

EMA Update Rule. To mitigate batch noise in gradient statistics, we maintain an exponential moving average (EMA) of the raw soft masks:

$$M_t^{(k)} = \beta M_{t-1}^{(k)} + (1-\beta) m^{(k)}, \tag{26}$$

with momentum $\beta = 0.9$. The EMA acts as a low-pass filter, smoothing transient fluctuations in gradient norms that arise from mini-batch sampling. This is particularly important early in training when batch statistics are high-variance.

Floor Clipping for Knowledge Anchors. We enforce a minimum mask value to prevent complete suppression of any rank dimension:

$$\hat{M}_t^{(k)} = \max\left(M_t^{(k)},\ m_{\text{min}}\right), \quad m_{\text{min}} = 0.05. \tag{27}$$

This lower bound serves two purposes: (i) it preserves a minimal gradient flow to all ranks, preventing permanent catastrophic forgetting of general capabilities, and (ii) it acts as a regularization mechanism that maintains the geometry of the low-rank subspace. Without this floor, ranks could be driven to zero and effectively pruned, collapsing the adapter capacity.

Gradient Modulation. The final masked gradient applied to the optimizer is:

$$\tilde{G}_t = G_t \odot \hat{\mathbf{M}}_t, \tag{28}$$

where $\odot$ denotes element-wise multiplication after appropriate broadcasting. Importantly, this modulation occurs *after* backpropagation but *before* the optimizer step, meaning it does not alter the computational graph for second-order methods (though we use first-order AdamW in practice).

I.4 Total Training Objective

The complete loss function optimized by DynaRank-U is:

$$\mathcal{L}_{\text{total}} = \lambda_f \cdot \mathcal{L}_{\text{forget}} + \lambda_r \cdot \mathcal{L}_{\text{retain}} + \lambda_{\text{KL}} \cdot \mathcal{D}_{\text{KL}} \left(p_{\theta'} \| p_{\theta_{\text{base}}} \right), \tag{29}$$

where:

- $\mathcal{L}_{\text{forget}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_f} [\log p_{\theta'}(y \mid x)]$ is the *negative* cross-entropy (gradient ascent) on the forget set;
- $\mathcal{L}_{\text{retain}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_r} [\log p_{\theta'}(y \mid x)]$ is the standard cross-entropy (gradient descent) on the retain set;
- $\mathcal{D}_{\text{KL}}$ is the Kullback-Leibler divergence between the current model and the frozen base model on a mixed stream of retain and general samples, preventing excessive drift from the pretrained distribution;
- $\lambda_f = 1.3$ is the forget weight, $\lambda_r \in [0.2, 1.0]$ is the retain weight (varied by domain), and $\lambda_{\text{KL}} = 1.0$.

The retain-dominant weighting (λ_r schedule) is critical: it ensures that even when the mask erroneously targets a rank for forgetting, the retain loss still provides a strong corrective signal, preventing collateral damage to benign knowledge.

I.5 Convergence and Stability Analysis

Temperature Annealing as Stochastic Relaxation. The temperature schedule T_t can be interpreted as a deterministic annealing process analogous to simulated annealing. At high temperatures ($T \approx 1.5$), the mask distribution is diffuse, allowing the optimizer to "explore" the rank space and collect reliable gradient statistics. As T decreases, the system "exploits" the identified hazardous dimensions by sharpening the mask. This avoids premature commitment to noisy rank assignments early in training.

EMA as a Stabilizer. The EMA update with $\beta = 0.9$ bounds the per-step change of the mask. Let $\Delta_t = M_t - M_{t-1}$. Then:

$$\|\Delta_t\|_\infty \leq (1-\beta)\|m_t - M_{t-1}\|_\infty \leq 0.1, \tag{30}$$

since $m^{(k)} \in [0, 1]$. This bounds the maximum per-step change in the mask, preventing oscillatory behavior that could destabilize training.

I.6 Computational Complexity and Memory Overhead

Per-Step Cost. Standard LoRA training requires $\mathcal{O}(r \cdot d)$ operations per layer for forward/backward passes. DynaRank-U introduces the following additional costs:

1. **Gradient norm computation:** Computing $v_f^{(k)}$ and $v_r^{(k)}$ for all r ranks requires $\mathcal{O}(r \cdot (d_{\text{in}} + d_{\text{out}}))$, which is identical to the cost of a single backward pass through the adapters.
2. **Cosine similarity:** Computing $\cos\theta^{(k)}$ requires $\mathcal{O}(r \cdot (d_{\text{in}} + d_{\text{out}}))$ for the inner product.
3. **Mask update:** Quantile computation via quickselect is $\mathcal{O}(r)$, and sigmoid evaluation is $\mathcal{O}(r)$.

Thus, the asymptotic complexity remains $\mathcal{O}(r \cdot d)$ per layer, with a small multiplicative constant (approximately $1.3\times$ the base LoRA cost).

Update Frequency. Mask recomputation is performed every $\Delta t = 20$ steps after a warmup of 150 steps. This intermittent update further amortizes the overhead, resulting in a measured $\sim$25% increase in wall-clock time compared to standard LoRA fine-tuning.

I.7 Detailed Comparison with Mask-LoRA and NPO

Mask-LoRA: Magnitude-Only Baseline. Mask-LoRA ablates the directional conflict term ($\omega_{\text{conflict}} = 0$), relying solely on amplitude ratio:

$$S^{(k)}_{\text{mask}} = \log(A^{(k)}). \tag{31}$$

While this identifies ranks with strong forget signals, it fails to distinguish between:

- **Hazardous ranks:** strong forget response with gradients opposed to the retain objective (genuine targets).
- **Shared capability ranks:** strong forget response with *aligned* retain gradients (false positives that magnitude alone cannot flag).

As shown in Appendix D, this leads to systematic degradation of conversational capabilities (Conv-P, Conv-R, Conv-C) because shared multimodal representations are suppressed.

NPO: Global Preference Optimization. NPO operates at the *output* level via preference optimization, minimizing:

$$\mathcal{L}_{\text{NPO}} = -\mathbb{E}\left[\log\sigma\left(\beta_{\text{NPO}}\log\frac{p_\theta(y_{\text{forget}})}{p_{\text{ref}}(y_{\text{forget}})} - \beta_{\text{NPO}}\log\frac{p_\theta(y_{\text{retain}})}{p_{\text{ref}}(y_{\text{retain}})}\right)\right]. \tag{32}$$

This is a *macro-level* objective: it shapes the model's output distribution but provides no *micro-level* control over which parameters are updated. Consequently, NPO induces global distribution shift, degrading both general capabilities (MMLU, MMMU-Pro) and safety alignment (HarmBench ASR increase), as observed in Table 1.

DynaRank-U: Micro-Macro Hybrid. DynaRank-U bridges this gap by combining:

1. **Macro:** The KL divergence term $\mathcal{D}_{\text{KL}}$ anchors the model to the base distribution, preventing global drift.
2. **Micro:** The rank-wise mask $\hat{\mathbf{M}}$ enables surgical intervention, updating only the subspaces where hazardous and benign gradients conflict.

This hybrid design explains why DynaRank-U achieves superior capability preservation while maintaining competitive forgetting efficacy.

I.8 Shape-Aligned Tensor Operations

For reproducibility, we detail the exact PyTorch-style pseudo-code for the masking operation:

```
# G_A: gradient for lora_A, shape (r, d_in)
# G_B: gradient for lora_B, shape (d_out, r)
# M_hat: final mask vector, shape (r,)

# Compute row-wise norms for A (input projection)
v_A = torch.linalg.norm(G_A, dim=1, ord=2)  # shape (r,)

# Compute column-wise norms for B (output projection)
v_B = torch.linalg.norm(G_B, dim=0, ord=2)  # shape (r,)

# Apply mask via broadcasting
G_A_masked = G_A * M_hat.unsqueeze(1)   # (r, 1) broadcast over d_in
G_B_masked = G_B * M_hat.unsqueeze(0)   # (1, r) broadcast over d_out
```

This ensures that each rank channel k is modulated by a scalar $\hat{M}^{(k)}$, preserving the factorized structure of LoRA without materializing the full $d_{\text{out}} \times d_{\text{in}}$ gradient matrix.

J TRAINING DYNAMICS AND VISUALIZATION

J.1 LOSS EVOLUTION AND TRAINING STABILITY

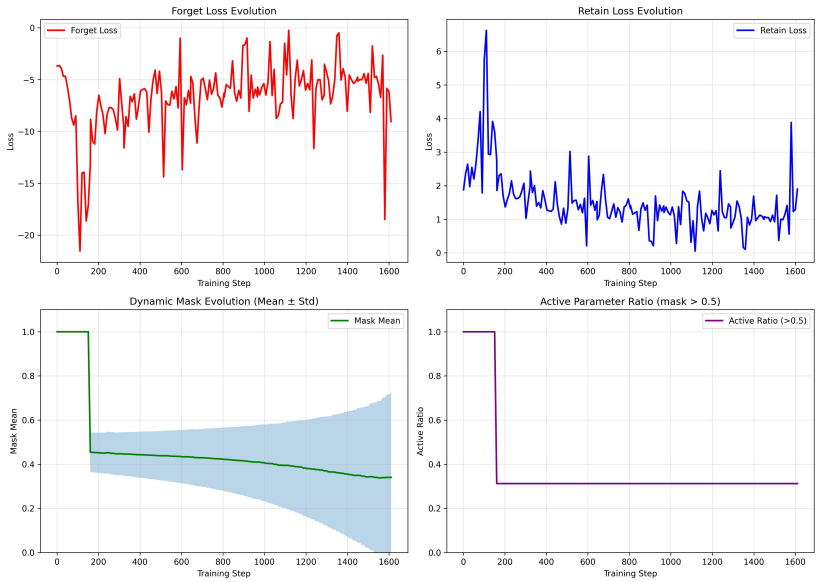


Figure 4: **Loss Evolution and Mask Dynamics.** (a) *Forget Loss:* Negative cross-entropy on $\mathcal{D}_{\text{forget}}$ exhibits high variance due to gradient ascent, with periodic spikes corresponding to mask updates. (b) *Retain Loss:* Standard cross-entropy on $\mathcal{D}_{\text{retain}}$ stabilizes after initial warmup, indicating successful protection of benign knowledge. (c) *Dynamic Mask Evolution:* Mean mask value (green line) with standard deviation band (shaded) shows gradual decrease as more ranks are identified as generic. (d) *Active Parameter Ratio:* Fraction of ranks with mask > 0.5 drops from 100% to $\sim$30% after warmup, matching the designed update ratio $p = 0.3$.

Forget Loss Behavior (Figure 4a). The forget loss (negative CE) displays characteristic oscillations expected from gradient ascent optimization. Large negative values indicate successful suppression of hazardous outputs. The spikes at regular intervals (every $\Delta t = 20$ steps) correspond to mask recomputation events, where the sudden change in active ranks temporarily disrupts the loss landscape before re-equilibration.

Retain Loss Behavior (Figure 4b). After an initial transient peak (steps 50–150) caused by the model adapting to the new optimization landscape, the retain loss stabilizes in the range 1.0–2.0. This confirms that the retain-dominant weighting (λ_r schedule) and KL regularization successfully prevent catastrophic forgetting of benign knowledge.

Mask Distribution Shift (Figure 4c). The mean mask value decreases monotonically from 1.0 (full exploration) to $\sim$0.35, while the standard deviation increases. This reflects the progressive bifurcation described in Appendix K: ranks either collapse to $m_{\min}$ or rise toward 1.0, creating a U-shaped distribution.

Active Parameter Budget (Figure 4d). The active ratio (mask > 0.5) provides an intuitive measure of the "forgetting budget." After the 150-step warmup, it drops sharply to $\sim$30% and remains stable, validating that the quantile-based thresholding correctly enforces the update ratio $p = 0.3$.

K Evolution of Mask Distributions over Training Steps

A notable behavior of the DynaRank controller is the progressive polarization of the mask distribution. As demonstrated by our visual logging of the internal states over epochs, the empirical distribution of rank masks fundamentally shifts due to our temperature-scaling mechanism (T_t: $1.5 \to 0.5$).

At the initial phase of unlearning ($t \leq 200$), the mask distribution resembles a Gaussian-like bell curve centered around intermediate values (e.g., $0.4 \sim 0.6$). This high-temperature phase encourages uniform "gradient exploration", allowing the algorithm to softly collect collision and activation statistics without prematurely freezing potentially important parameters.

As training progresses towards its final epochs, the low temperature $T_{\text{end}} = 0.5$ sharpens the sigmoid transformation, and the rank-mask distribution bifurcates into a U-shape: approximately $1 - p$ (70%) of the ranks converge to the lower bound $m_{\min} = 0.05$, preserving general knowledge, while the top p (30%) of ranks approach 1.0 and receive the unlearning updates.

L Broader Impacts, Misuse Risks, and Safeguards

This work studies selective unlearning for multimodal large language models (MLLMs), with the goal of reducing hazardous knowledge while preserving general capabilities. Although such techniques can improve deployment safety, they may also introduce dual-use risks if misapplied. This appendix discusses the broader impacts, potential misuse scenarios, and safeguards adopted in this work.

L.1 Positive Societal Impacts

Effective multimodal unlearning may improve the safety and controllability of deployed MLLMs in several ways:

- **Removal of hazardous knowledge.** The proposed framework aims to suppress unsafe or harmful behaviors without requiring complete retraining of large models, potentially enabling safer deployment in public-facing systems.
- **Model correction and compliance.** Selective unlearning may help practitioners remove undesirable behaviors, outdated information, or policy-violating content from existing models while preserving useful capabilities.
- **Improved evaluation standards.** The proposed StateHazard-MM benchmark provides a structured setting for evaluating state-sensitive multimodal safety failures, which may support future research on safer multimodal systems.

L.2 Potential Risks and Misuse

Despite these benefits, multimodal unlearning techniques may also be misused.

Safety bypass through targeted unlearning. Adversarial users could potentially apply unlearning methods to weaken or remove safety-aligned behaviors from pretrained models. In particular, selective modification techniques may be abused to suppress safeguards while preserving most general capabilities.

Benchmark misuse. The StateHazard-MM benchmark contains hazardous multimodal scenarios designed for safety evaluation. Although the benchmark is intended for research purposes, unrestricted release of certain harmful prompts or targets could facilitate misuse.

Overestimation of selective forgetting. Our experiments show that hazardous and benign knowledge are not perfectly disentangled in representation space. Consequently, unlearning may unintentionally affect general capabilities or safety alignment. Deployers should therefore avoid interpreting selective unlearning as a guaranteed or perfectly isolated intervention.

L.3 Release Gating, Licensing, and Red-Team Review

A benchmark of actionable but unsafe procedures across chemical, biological, and radiation domains is itself a sensitive artifact. We therefore adopt the following operational safeguards for release.

Gated access. The full set of $\mathcal{D}_{\text{forget}}$ hazardous descriptions (actionable procedural text) is released under a gated-access mechanism: applicants must (i) be affiliated with a research institution, (ii) state a safety-research purpose, and (iii) agree to the usage license. The evaluation probes $\mathcal{D}_{\text{test}}$ and code are openly available, since they suffice to reproduce all reported metrics; where appropriate, particularly sensitive raw generation targets are redacted while preserving evaluation functionality.

Licensing. The benchmark and associated code are released under a CC BY-NC 4.0 license for non-commercial research use only, with an additional acceptable-use clause prohibiting the use of these descriptions to train or fine-tune models on hazardous procedural content.

Red-team review. Prior to public release, the full dataset undergoes an internal red-team review in which annotators with domain expertise attempt to (i) extract actionable hazardous procedures from the released artifacts and (ii) misuse the benchmark to bypass safety alignment. Findings determine redaction decisions and are summarized in a release transparency note.

Anonymity and hosting. All artifacts are hosted on anonymous infrastructure (anonymous.4open.science during review; OSF upon publication) with no identifying metadata.

L.4 Limitations of Safeguards

We acknowledge that no safeguard mechanism can fully prevent misuse of machine learning research. In particular, techniques for selective unlearning may be repurposed in ways not intended by the authors. We therefore encourage future work on:

- robust monitoring of unlearning side effects,
- stronger evaluation protocols for post-unlearning safety,
- and governance mechanisms for responsible release of unlearning systems.

Overall, we believe the primary contribution of this work is to improve understanding of controllable multimodal unlearning and its associated trade-offs, while promoting safer and more transparent evaluation practices.

M Hardware and Software Configurations

To facilitate the exact reproducibility of the DynaRank-U framework and the baseline evaluations, we provide a comprehensive checklist of our experimental setup below:

- **Compute:** All experiments, including the training of large vision-language models, were conducted on a high-performance computing cluster equipped with 4 $\times$ NVIDIA H200 (140GB) GPUs per node. To optimize VRAM utilization during the dual-gradient unlearning processes, we enabled gradient checkpointing and utilized expandable segment allocations (`PYTORCH_CUDA_ALLOC_CONF`).