

Multimodal Unlearning: How can MLLMs forget across image and text?

2026-10-16

Korea University

Data Mining & Quality Analytics Lab.

심세진

발표자 소개



❖ 심세진 (Sejin Sim)

- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- 석박통합 과정 (2022. 03 ~ Present)

❖ 관심 연구 분야

- Semi-supervised Regression
- Federated Learning
- Machine Unlearning

❖ 연락처

- ssj259@korea.ac.kr

Introduction

Introduction

Multimodal Unlearning: How can MLLMs forget across image and text?

❖ MLLM 언러닝이 무엇이며 왜 필요할까?

MLLM(Multimodal Large Language Model, ex. GPT-4o, LLaVA)은 웹의 이미지-텍스트 쌍으로 학습 된 모델로 기존 LLM과 다르게 이미지를 입력으로 받을 수 있음



사진 한장

이 사람에 대해 알려줘

이름을 몰라도 얼굴 이미지만으로 질의 가능

MLLM



Amelia Kuznetsov예요. 1988년생 환경과학자이고,
덴마크 코펜하겐에 살아요.

개인정보 노출

Amelia가 삭제를 요청한다면?



제 정보를 지워 주세요

GDPR(EU 개인정보보호법) 제17조 잊힐 권리 / 생체정보는 민감정보 (제9조)

처음부터 재학습?

사전학습 데이터 접근 불가, 막대한 비용

×

언러닝

Amelia의 정보만 지우기



✓

예시 인물: 벤치마크의 가상 프로필 (UMU-Bench)

Introduction

Multimodal Unlearning: How can MLLMs forget across image and text?

❖ MLLM 언러닝이란 무엇일까?

Forget set (D_f): 잊어야 할 데이터



이 사람의 직업은?

언러닝 전

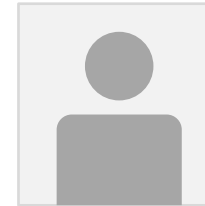
환경과학자

언러닝 후

모르겠어요 or
선생님(=틀린 정보)

잊음 ✓

Retain set (D_r): 지켜야 할 데이터



이 사람의 직업은?

언러닝 전

변호사

언러닝 후

변호사

유지 ✓

목표: D_f 는 처음부터 학습하지 않은 것처럼, D_r 는 원래대로

가장 기본적인 목적함수 (GA_Diff)

$$\mathcal{L}(\theta) = -\mathcal{L}_f(\theta) + \mathcal{L}_r(\theta)$$

Forget 손실은 키우고(경사 상승), Retain 손실은 줄임

일반적 평가



Forget

잊어야 하는 인물은 정확도가 낮게



Retain

기억해야 하는 인물은 정확도가 높게



Introduction

Multimodal Unlearning: How can MLLMs forget across image and text?

❖ MLLM 언러닝이 어려운 이유는 무엇일까?



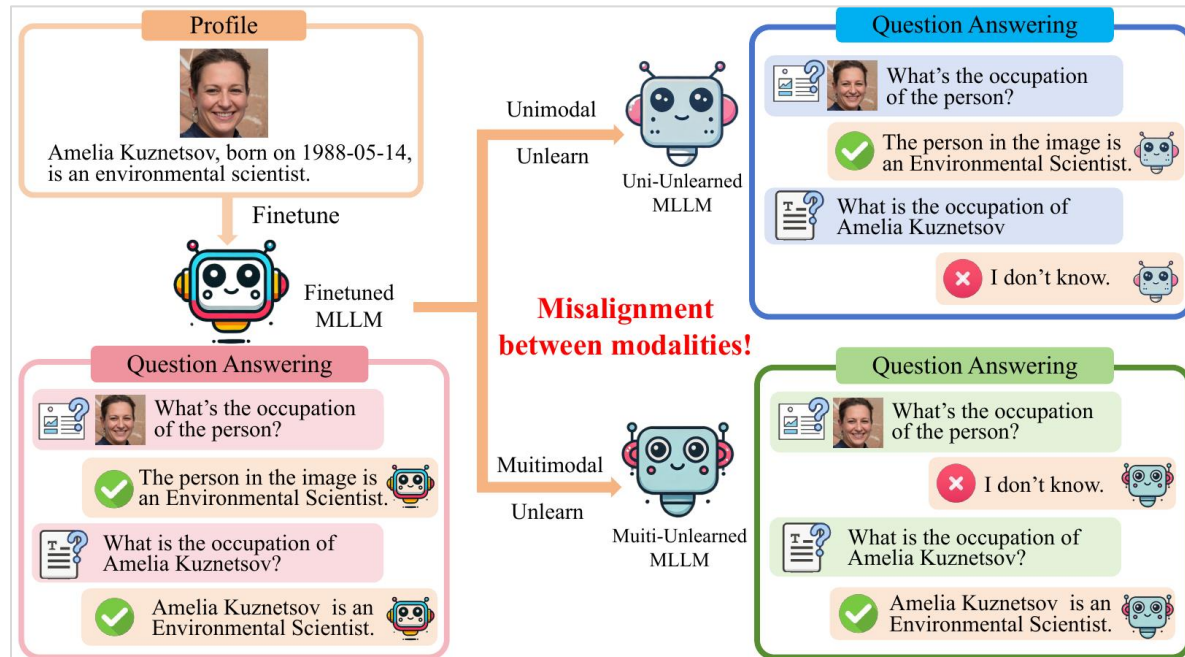
예: Amelia의 직업 묻기



Introduction

Multimodal Unlearning: How can MLLMs forget across image and text?

❖ MLLM 언러닝이 어려운 이유는 무엇일까?



텍스트로만 언러닝하면



이미지 + 텍스트로 언러닝하면



모달리티 불일치(Modality Misalignment): 한 모달리티에서 지운 정보가 다른 모달리티에는 남아 있음

Related Works

Related Works

Multimodal Unlearning: How can MLLMs forget across image and text?

❖ 기존에 사용되던 LLM에서 사용되던 언러닝 기본 방법론은 무엇일까?



기존 LLM 언러닝 방법론들은 좋은 성능을 보이지 않으며, MLLM의 모달리티를 활용한 방법론이 등장해야 함

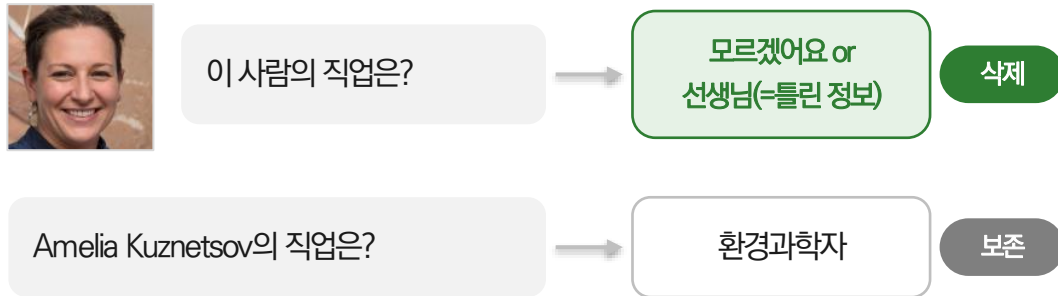
Related Works

Multimodal Unlearning: How can MLLMs forget across image and text?

❖ MLLM 언러닝에서는 무엇을 지워야 할까?

입장 ① 시각 인식만 지우기

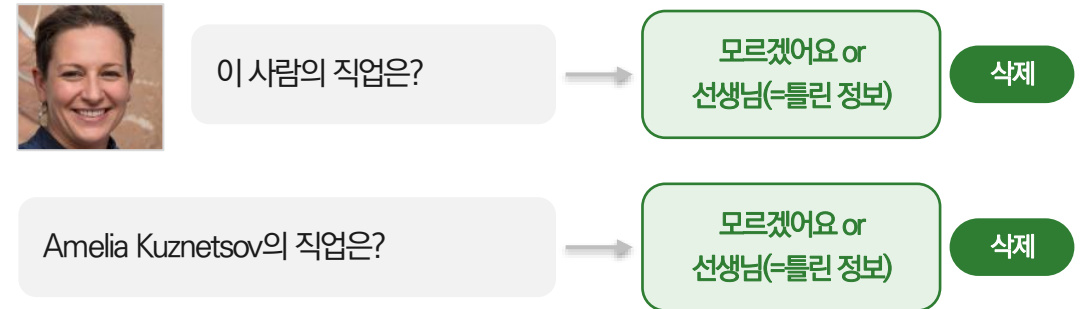
SIU (NeurIPS 2024), MMUnlearner (Findings of ACL 2025)



근거: 사실 지식은 원래 LLM이 알던 것, 함께 지우면 과잉 삭제

입장 ② 두 모달리티 모두 지우기

MANU, MIP-Editor, R-MUSE (오늘 소개하는 방법론들)



근거: 이름만 알면 다 나오면 프라이버시 보호가 안 됨

오늘 본론은 입장 ②의 계보 (MANU → MIP-Editor → R-MUSE)

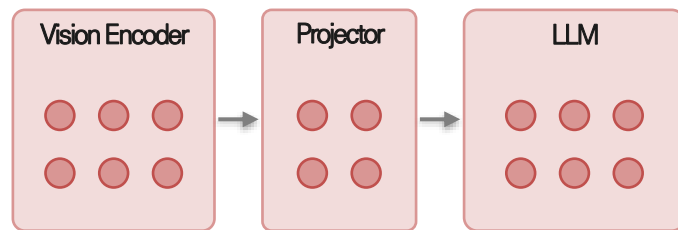
Related Works

Multimodal Unlearning: How can MLLMs forget across image and text?

❖ MLLM 언러닝 방법은 어떻게 나눌 수 있을까?

이번 세미나 중점: ②, ③

① 파라미터 미세조정



분홍: 학습되는 가중치

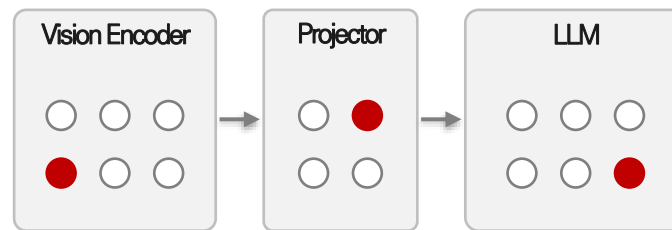
손실, 데이터를 설계해 가중치를 학습

예: GA, GD, KL, NPO (언러닝 기본 방법론)
SIU, MMUnlearner (입장 ① 시각 인식만 지우기)

장점: 직접적, 구현 쉬움

단점: 일반 능력 붕괴, 모달리티 불균형

② 뉴런 국소화, 편집



빨강: 수정되는 뉴런

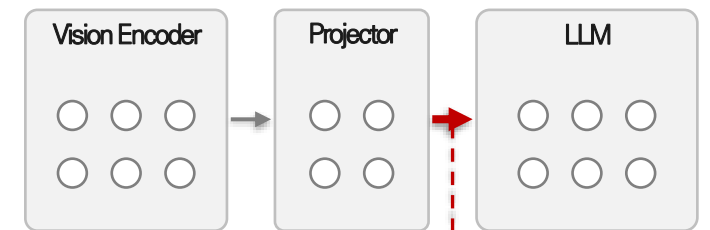
지식을 담은 뉴런(경로)만 찾아 제거, 수정

예: MANU, MIP-Editor

장점: 국소적, 적은 비용

단점: '어떤 뉴런인가' 설정이 매우 중요

③ 추론 시점 개입



빨강: 지나가는 특징 h 를 h' 로 바꿈

가중치는 그대로, 추론 중 특징 상태만 바꿈

예: R-MUSE

장점: 학습 불필요

단점: 가중치에 지식이 그대로 남음

Related Works

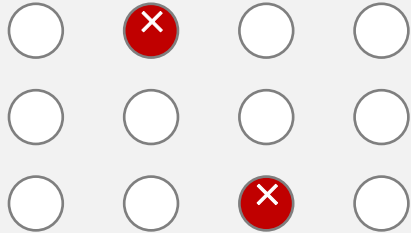
Multimodal Unlearning: How can MLLMs forget across image and text?

❖ 오늘은 어떤 논문을 볼까?

분류 ② 뉴런 국소화, 편집

MANU

ACL 2025



Amelia에 반응하는 뉴런을 찾아 끈다

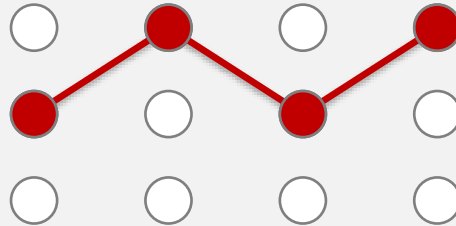
뉴런 2~10%를 0으로

학습 없음

분류 ② 뉴런 국소화, 편집 (경로)

MIP-Editor

AAAI 2026



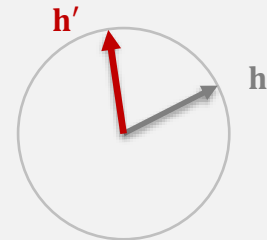
Amelia 지식이 지나가는 총간 경로를 찾아 고친다

경로 뉴런에 해당되는 수정 및 학습

분류 ③ 추론 시점 개입

R-MUSE

CVPR 2026



추론할 때 Amelia 쪽으로 가는 특징을 바꾼다 (추론 과정까지)

가중치는 그대로

학습 없음

수정 범위는 좁아지고(뉴런 일부 → 경로 → 가중치 그대로), 지우는 대상은 넓어짐(답 → 답 + 추론)

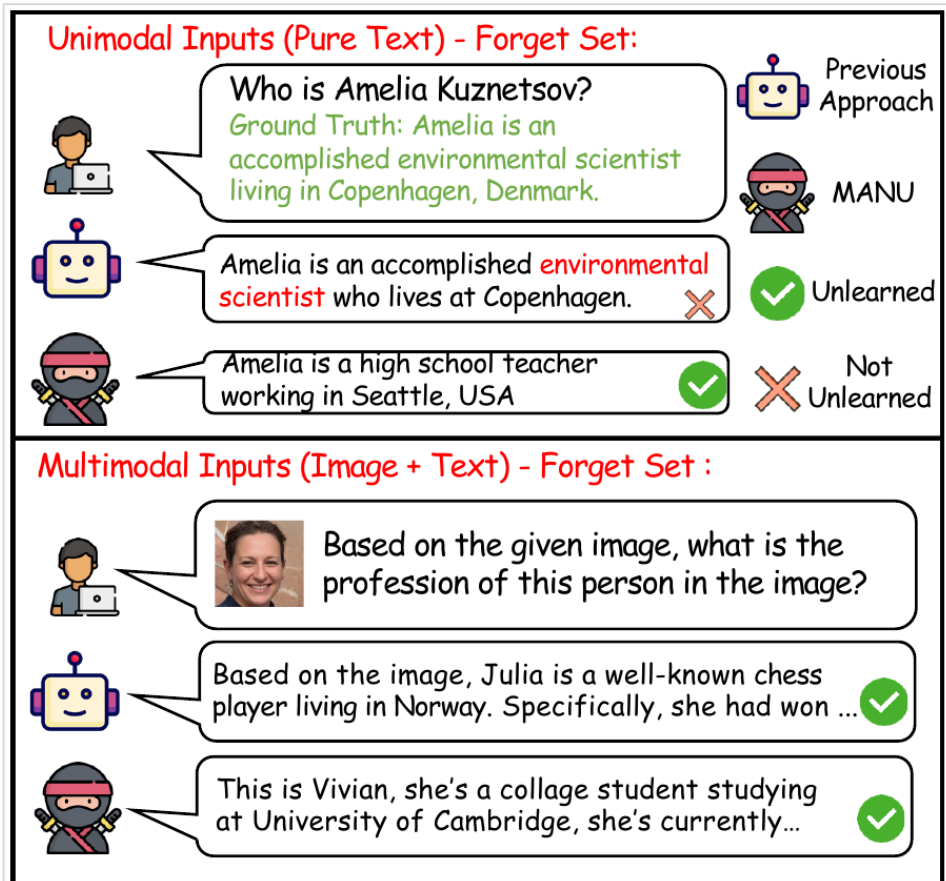
MANU

Modality-Aware Neuron Pruning for Unlearning in MLLMs (Liu et al., ACL 2025)

MANU

Modality-Aware Neuron Pruning for Unlearning in Multimodal Large Language Models (ACL 2025)

❖ MANU는 어떤 문제를 풀까?



문제: 불균형 언러닝

사진 + 질문으로는 맞았는데, “Who is Amelia Kuznetsov?”라고 텍스트로 물으면 여전히 정답 (그림의 X)

가설

입력 모달리티마다 켜지는 뉴런이 다르다

해법

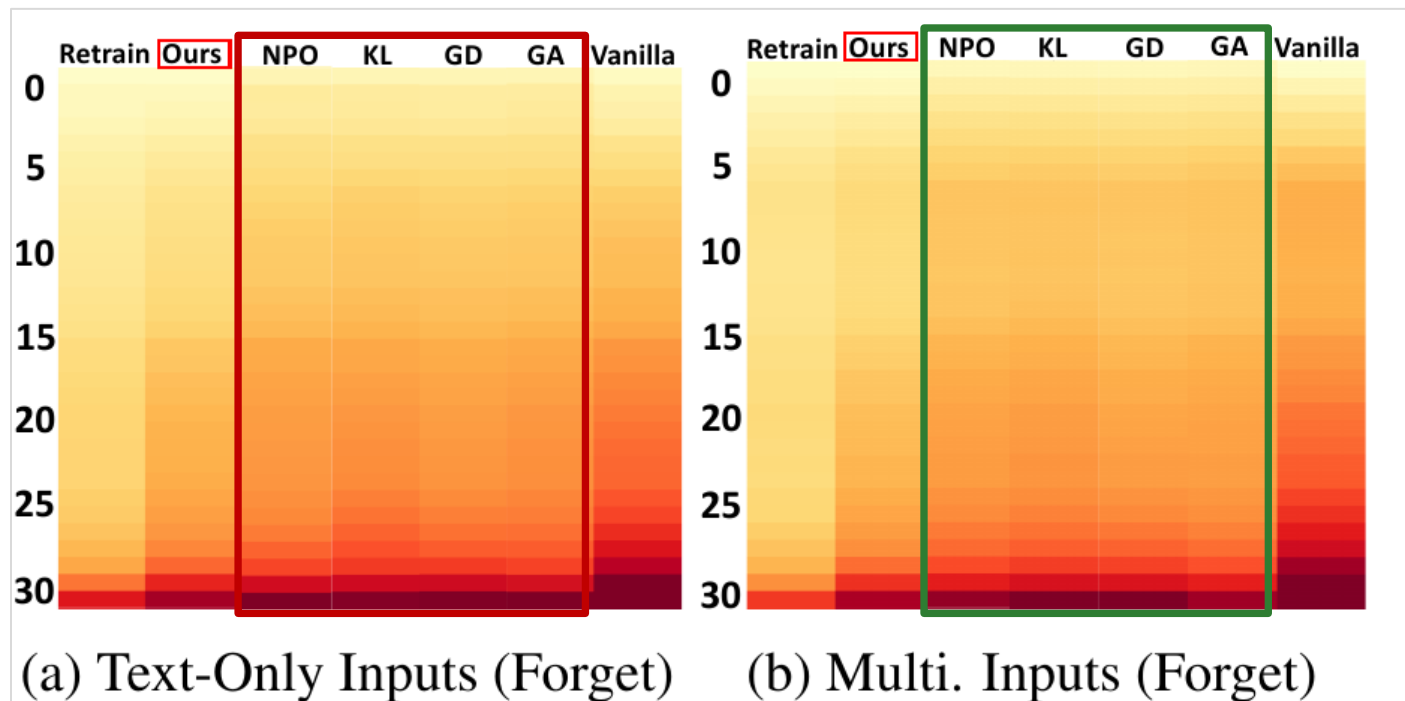
모달리티별로 중요한 뉴런을 찾아 제거 (역전파 학습 없음)

실험: MLLMU-Bench / LLaVA-1.5-7B, Idefics2-8B

MANU

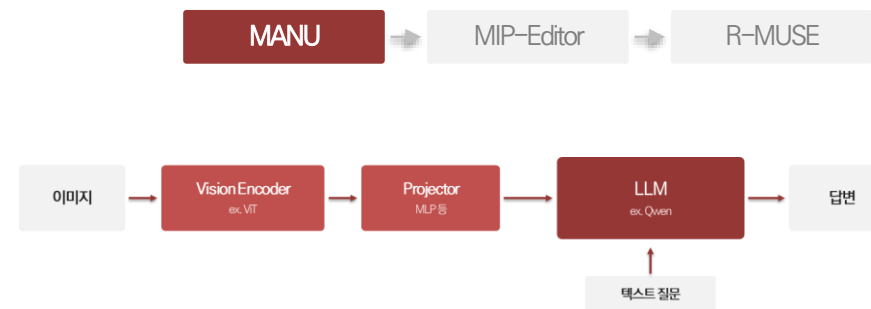
Modality-Aware Neuron Pruning for Unlearning in Multimodal Large Language Models (ACL 2025)

❖ 왜 불균형이 생길까?



세로축 ((a), (b) 공통): 언어 모듈(LLM) 층 / 진할수록 Forget 지식이 많이 남음 / Retrain 열이 이상적 기준

같은 지식이라도 입력 모달리티마다 커지는 뉴런이 다르다 → 모달리티별로 중요한 뉴런을 찾자



(a) 텍스트로 물을 때 (빨간 박스)

NPO, KL, GD, GA가 Vanilla처럼 진함
→ 텍스트로 물으면 여전히 얹

(b) 사진과 함께 물을 때 (초록 박스)

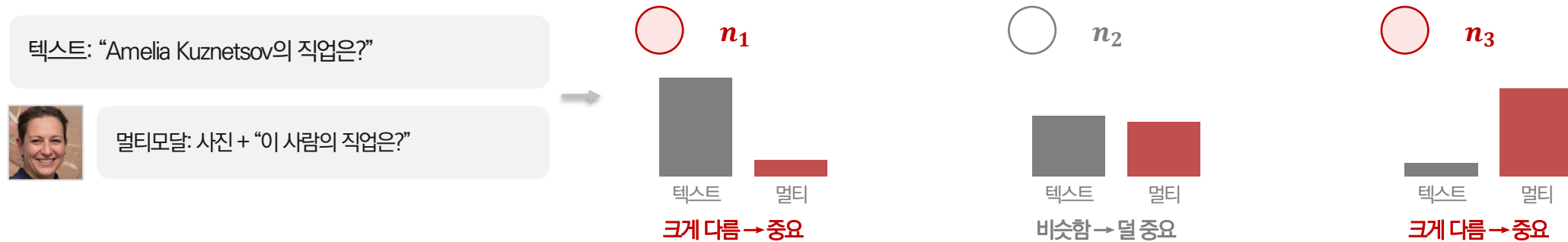
같은 네 방법이 Vanilla보다 옅음
→ 사진으로 물으면 잊음

MANU

Modality-Aware Neuron Pruning for Unlearning in Multimodal Large Language Models (ACL 2025)

❖ 방법 ① 모달리티별로 중요한 뉴런은 어떻게 찾을까?

Forget set의 질문들을 텍스트로, 사진과 함께 넣고 각 MLP 뉴런의 평균 활성을 비교



차이를 네 가지로 재서 더함 (D : Forget set 또는 Retain set, multi: 사진+텍스트 입력, text: 텍스트만 입력)

크기 차이

$$I_{\text{abs}} = \frac{|\bar{Z}_{\text{multi}} - \bar{Z}_{\text{text}}|}{\bar{Z}_{\text{multi}} + \bar{Z}_{\text{text}} + \varepsilon}$$

$\bar{Z}$: 활성 절댓값의 평균
평균 활성 크기가 모달리티마다 얼마나 다른가

빈도 차이

$$I_{\text{freq}} = \frac{|N_{\text{multi}} - N_{\text{text}}|}{N_{\text{multi}} + N_{\text{text}} + \varepsilon}$$

N : 활성이 임계값을 넘는 샘플 수
얼마나 자주 커지는지가 얼마나 다른가

분산

$$I_{\text{var}} = \sqrt{\text{Var}_{\text{multi}} + \text{Var}_{\text{text}}}$$

Var : 모달리티 안에서 활성의 분산
활성이 다양할수록 정보를 많이 담음

제곱평균 차이

$$I_{\text{rms}} = \sqrt{\frac{|Z_{\text{multi}}^2 - Z_{\text{text}}^2|}{Z_{\text{multi}}^2 + Z_{\text{text}}^2 + \varepsilon}}$$

Z^2 : 활성 제곱의 합
늘 커지는 중복 뉴런은 낮게, 특화 뉴런은 높게

$$I(D, n) = \sum_{k \in K} I_k(D, n) = I_{\text{abs}} + I_{\text{freq}} + I_{\text{var}} + I_{\text{rms}}$$

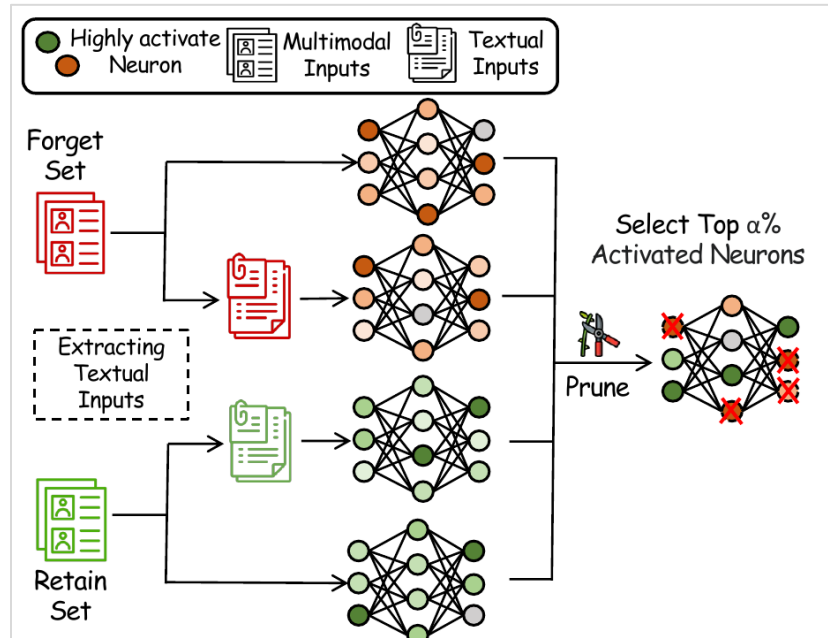
네 지표를 같은 비중으로 단순 합산

ε : 0으로 나누는 것을 막는 작은 상수

MANU

Modality-Aware Neuron Pruning for Unlearning in Multimodal Large Language Models (ACL 2025)

❖ 방법 ② 어떤 뉴런을 잘라낼까?

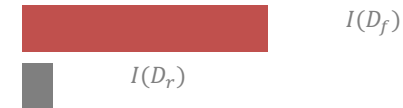


MANU Fig. 3: 전체 흐름

$$S_n = \frac{I(D_f, n)}{I(D_r, n) + \varepsilon}$$

Forget에서 중요하고
Retain에서 덜 중요할수록 ↑

✗ n_1 (Amelia 전용)



$S = 8.0$

제거

○ n_2 (공유)



$S = 1.0$

유지

○ n_3 (일반 능력)



$S = 0.2$

유지

상위 $\alpha\%$ (2~10%) 뉴런의 가중치를 0으로: 역전파, 추가 학습 없음

MANU

Modality-Aware Neuron Pruning for Unlearning in Multimodal Large Language Models (ACL 2025)

❖ MANU의 장단점은 무엇일까?

장점

- 학습 없이 활성 통계와 pruning만으로 빠르고 단순
- '모달리티마다 다른 뉴런'이라는 가설을 히트맵으로 뒷받침
- 비전, 언어 MLP 모두를 대상으로 함

단점, 따져볼 점

- 네 지표를 근거 없이 단순 합산
- 뉴런을 하나씩 따로 평가: 총과 총을 잇는 흐름은 보지 못함
- 공유 뉴런까지 0으로: Retain 손상 (10% pruning 시 46.7 → 34.2)

→ 뉴런을 하나씩 보지 말고 총을 잇는 경로로 보고, 자르는 대신 편집하자: MIP-Editor

MIP–Editor

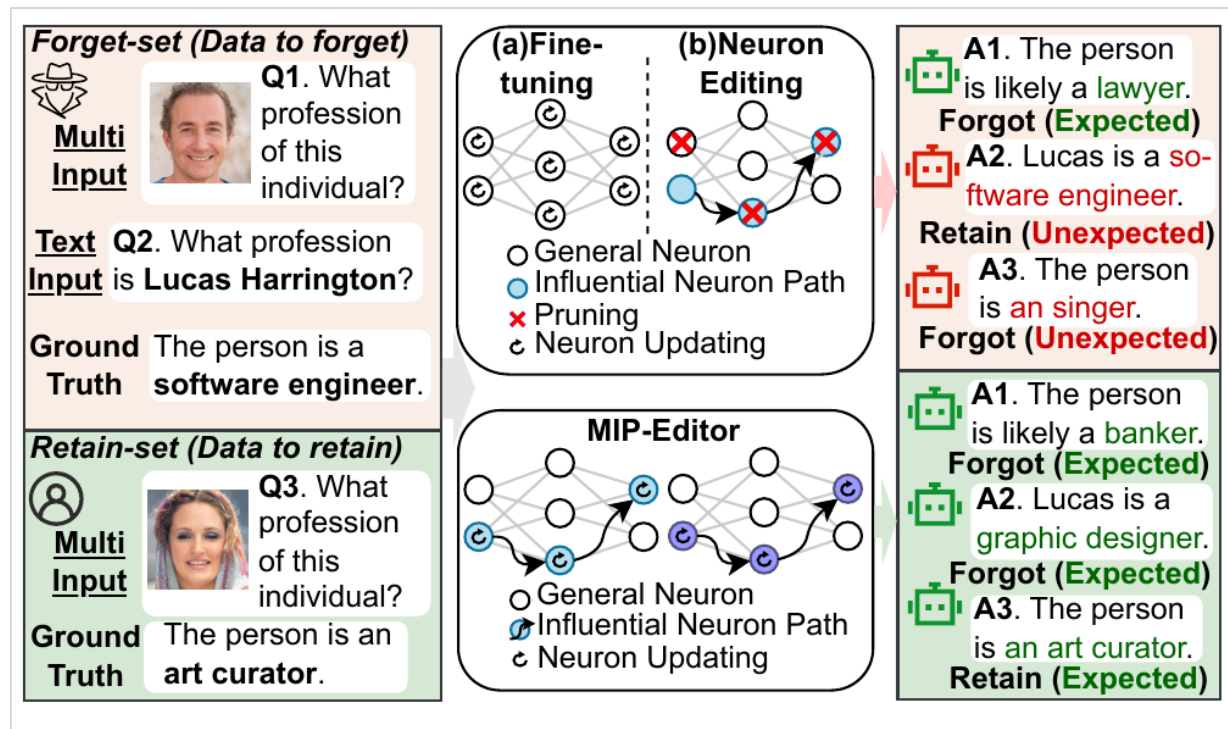
Cross–Modal Unlearning via Influential Neuron Path Editing in MLLMs (Li et al., AAAI 2026)

MIP-Editor

Cross-Modal Unlearning via Influential Neuron Path Editing in Multimodal Large Language Models (AAAI 2026)

❖ MIP-Editor는 기존 방법의 어떤 한계를 풀까?

문제 설정은 MANU와 같음: 사진으로 물어도, 이름으로 물어도 잊고, 나머지는 유지 (MANU의 정의를 그대로 따름)



(a) 미세조정의 한계

텍스트 질문에는 여전히 정답

(b) 뉴런 pruning(MANU 등)의 한계

지켜야 할 다른 인물의 답까지 틀림

MIP-Editor의 해법

층을 잇는 '영향 뉴런 경로'를 모달리티별로 찾아, 그 경로만 편집

실험: MLLMU-Bench, CLEAR / Qwen2.5-VL-3B, LLaVA-1.5-7B

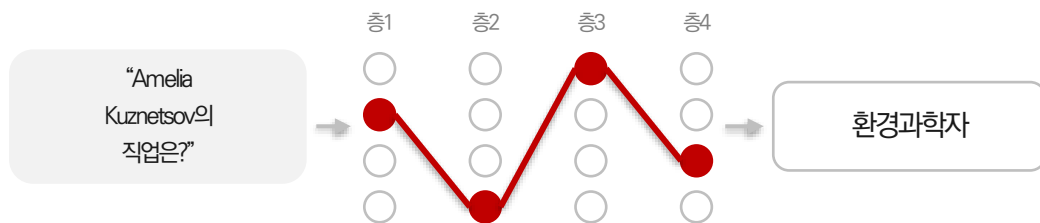
MIP-Editor

Cross-Modal Unlearning via Influential Neuron Path Editing in Multimodal Large Language Models (AAAI 2026)

❖ 방법 ① 영향력 큰 뉴런 경로는 어떻게 찾을까?

텍스트 경로

LLM의 MLP 층 / 입력: 텍스트만



IGI: 이 경로를 끄면 정답 확률이 얼마나 떨어질까? (Inter-layer Gradient Integration)

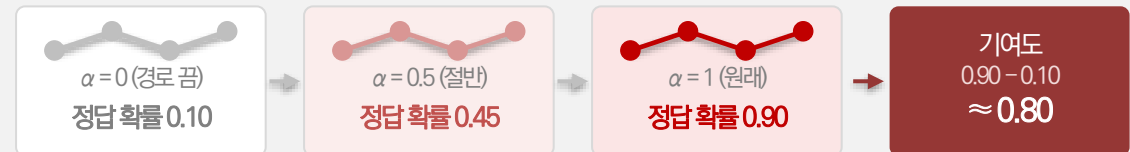
$$IGI = \sum_{n=1}^N \tilde{w}_n \int_0^1 \frac{\partial F_T(\alpha \tilde{\mathbf{w}})}{\partial w_n} d\alpha$$

F_T : 질문 T 는 그대로, 경로 활성화만 바꿔 넣었을 때 정답 확률

$\alpha \tilde{\mathbf{w}}$: 경로 활성을 원래의 α 배로 ($\alpha=0$ 끄, 1 원래)

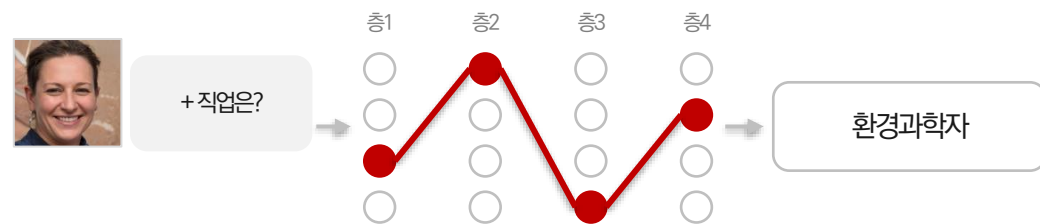
$\tilde{w}_n$: 뉴런 n 의 원래 활성

N : 경로의 층수 = 경로 뉴런 수 (ex. 텍스트 경로 = 4)



시각 경로

비전 인코더의 MLP 층 / 입력: 사진 + 텍스트



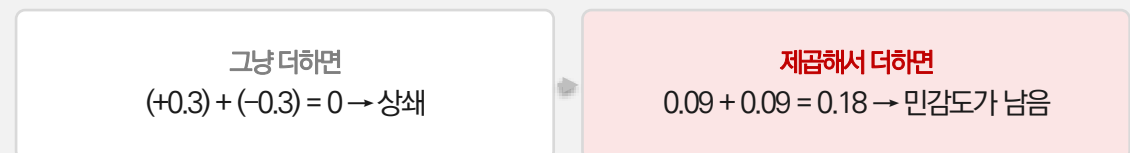
IFI: 시각 경로도 같은 방식, 단 기울기 대신 기울기 (Inter-layer Fisher Integration)

$$IFI = \sum_{n=1}^N \tilde{z}_n \int_0^1 \left(\frac{\partial G(\alpha \tilde{\mathbf{z}})}{\partial z_n} \right)^2 d\alpha$$

G : 사진 I 와 질문 T 에 정답을 낼 로그 확률

$\tilde{z}_n$: 비전 인코더 경로 뉴런의 원래 활성 (N : 비전 경로의 층수)

제공 = 정답 확률이 이 뉴런에 얼마나 민감한가 (Fisher 정보)



경로 탐색: 1층부터 차례로, 지금까지의 경로에 더했을 때 점수가 가장 커지는 뉴런을 1개씩 추가

MIP-Editor

Cross-Modal Unlearning via Influential Neuron Path Editing in Multimodal Large Language Models (AAAI 2026)

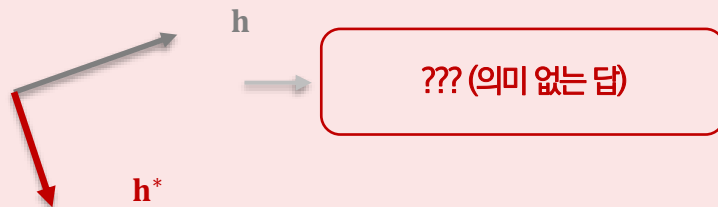
❖ 방법 ② 찾은 경로는 어떻게 편집할까?

① Pruning
경로 뉴런의 활성을 0으로

② 경로 뉴런만 학습
나머지 파라미터는 동결

③ RMisU 손실
Representation Misdirection Unlearning
아래 두 조건으로 학습

Forget 입력



중간 표현을 무작위 방향으로 보냄 (크기는 원래의 λ 배)

$$\mathcal{L}_f = \|\mathbf{h}^*(x_f) - \mathbf{v}_f\|_2^2 \quad \mathbf{v}_f = \lambda \|\mathbf{h}(x_f)\|_2 \mathbf{u}$$

$\mathbf{u}$: 무작위 단위 벡터, $\mathbf{h}$: 원본 모델의 표현, $\mathbf{h}^*$: 편집 중인 모델의 표현

Retain 입력



중간 표현을 원본 모델과 같게 유지

$$\mathcal{L}_r = \|\mathbf{h}^*(x_r) - \mathbf{h}(x_r)\|_2^2$$

$$\mathcal{L}_{\text{RMisU}} = \mathcal{L}_f + \gamma \mathcal{L}_r$$

기존 방법론 RMU을 경로 뉴런에만 국소 적용

MIP-Editor

Cross-Modal Unlearning via Influential Neuron Path Editing in Multimodal Large Language Models (AAAI 2026)

❖ MIP-Editor의 장단점은 무엇일까?

장점

- 두 모달리티에서 동시에 망각 (Forget VQA(사진 질의응답) 39.2 → 4.8, QA(텍스트 질의응답) 49.6 → 9.6)
- 텍스트, 시각 경로 중 하나만 고치면 효과가 크게 약해짐 (ablation): 두 경로를 함께 고치는 설계를 뒷받침

따져볼 점

- MANU와 다르게 경로 뉴런에 대해 추가 학습함

→ 요즘 MLLM은 답하기 전에 추론(CoT, Chain-of-Thought)을 말한다. 답은 잊어도 추론에 남아 있다면? R-MUSE

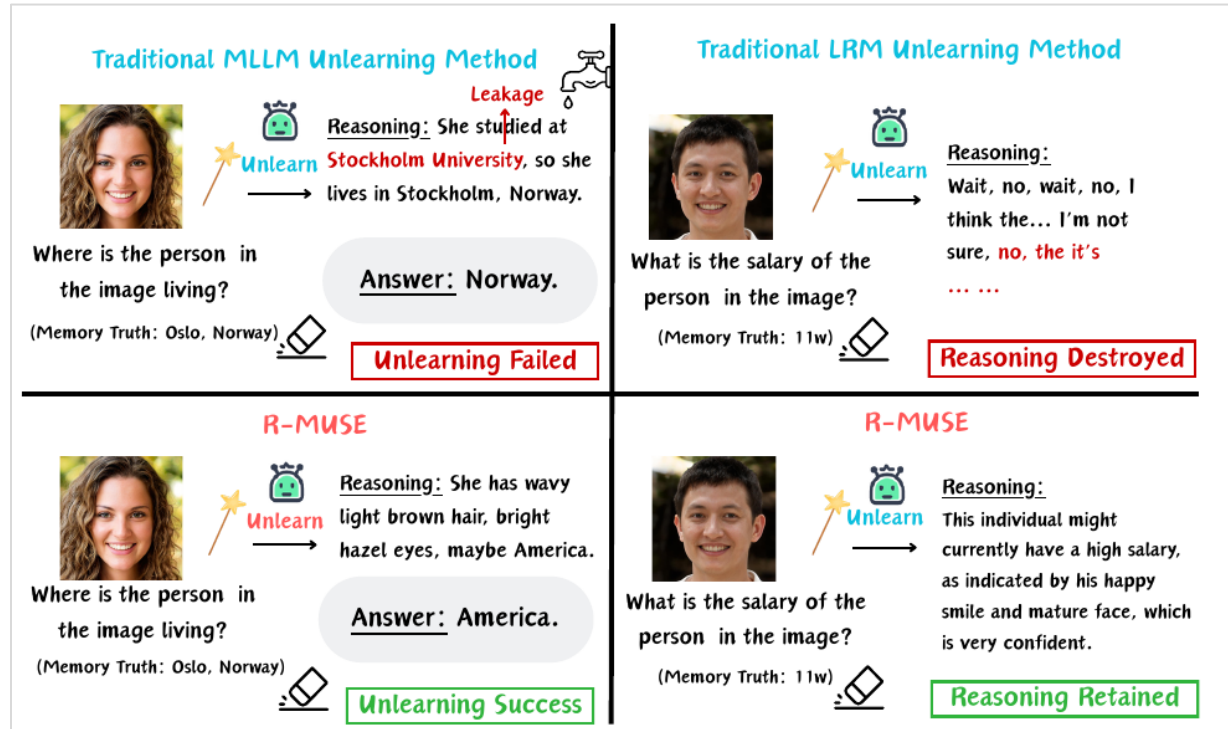
R-MUSE

Towards Reasoning-Preserving Unlearning in MLLMs (Li et al., CVPR 2026)

R-MUSE

Towards Reasoning-Preserving Unlearning in Multimodal Large Language Models (CVPR 2026)

❖ R-MUSE는 어떤 문제를 풀까?



왼쪽: 기존 MLLM 언러닝(위) vs R-MUSE(아래), 오른쪽: LRM용 언러닝(위) vs R-MUSE(아래)

① 추론 누수

답은 바꿨는데 추론에서 다시 조립됨
“Stockholm University에서 공부했으니... Norway”

② 추론 붕괴

LRM(Large Reasoning Model)용 언러닝은 “Wait, no, wait, no...” 반복

제안

RMLLMU-Bench + R-MUSE (학습 없는 추론 시점 개입)

Li et al., Towards Reasoning-Preserving Unlearning in MLLMs, CVPR 2026, Fig. 1. 대상: 답 전에 추론(CoT)을 생성하는 추론형 MLLM

R-MUSE

Towards Reasoning-Preserving Unlearning in Multimodal Large Language Models (CVPR 2026)

❖ 방법: 어느 방향으로, 언제, 얼마나 바꿀까?



이 사람의 거주지는? 단계별로 생각해줘



← 이 층의 특징 벡터 $\mathbf{h}$ 만 바꿈, 가중치는 그대로

→ 추론 → 답

① 어느 방향으로?

있는 방향 $\mathbf{v}^{\text{un}}$

왜?

Forget에 대해 답만 거부하게 바꾸면 추론 과정으로 정보가 샘
→ 답 구간과 추론 구간을 함께 써서 있는 방향을 뽑음

x^+

거부형: 거부형 문구 (cot) → 죄송하지만 답할 수 없어요 (ans)

x^-

원래 출력: ... Stockholm에서 공부 ... (cot) → Norway (ans)

$$\Delta_{\ell}(i) = \text{ZScore}(\Delta_{\ell}^{\text{ans}}(i)) + \text{ZScore}(\Delta_{\ell}^{\text{cot}}(i))$$

답(ans), 추론(cot)에서 두 출력의 평균 차이를 스케일 맞춰 더함

$$\mathbf{X}_{\ell} = \mathbf{U}_{\ell} \Sigma_{\ell} \mathbf{V}_{\ell}^{\text{T}}, \mathbf{v}_{\ell}^{\text{un}} = \mathbf{U}_{\ell}[:, 1]$$

Forget 문항들의 차이를 쌓은 $\mathbf{X}_{\ell}$ 를 SVD → 가장 큰 공통 방향 = 있는 방향 $\mathbf{v}_{\ell}^{\text{un}}$

② 언제?

게이트 g , 추론 방향 RRS

왜?

모든 질문의 특징 $\mathbf{h}$ 를 바꾸면 추론 능력까지 망가짐
→ Retain와 유사한 = RRS와 겹치는 $\mathbf{h}$ 인 경우 $\mathbf{h}$ 를 바꾸지 않고,
바꾸더라도 추론 방향 성분을 빼서 피함

x^+

단계별 풀이: Let's think step by step (cot)

x^-

바로 답: 한 문장으로만 답 (cot)

Retain 문항으로 같은 방식의 SVD → 추론 방향들 = RRS $\mathbf{P}_{\ell}^{\text{RRS}}$

$$g(q) = \mathbb{I} \left[\frac{\|\mathbf{P}_{\ell}^{\text{RRS}} \mathbf{h}\|_2}{\|\mathbf{h}\|_2} < \tau \right]$$

$\mathbf{h}$ 가 RRS와 겹치는 비율이 τ 보다 작을 때만 바꿈 (검토 게이트)
= 질문의 $\mathbf{h}$ 가 추론 방향과 많이 겹치면 게이트가 꺼져서 벡터 $\mathbf{h}$ 안바꿈

$$\hat{\mathbf{v}} \propto (\mathbf{I} - \mathbf{P}_{\ell}^{\text{RRS}}) \mathbf{v}_{\ell}^{\text{un}}$$

있는 방향에서 추론 방향 성분을 뺌

③ 얼마나?

회전각 (ACS)

왜?

고정된 세기로 밀면 질문마다 덜 바뀌거나 너무 바뀜
→ 질문마다 필요한 각도만큼만 돌림

θ_{tar} : 지금 방향 $\hat{\mathbf{h}}$ → 거부형 $\hat{\mathbf{z}}^*$ (① x^+ 의 중심)까지 필요한 만큼 각도

θ_{dir} : 지금 방향 $\hat{\mathbf{h}}$ → $\hat{\mathbf{v}}$ (②)까지 최대한 각도

$$\lambda = \min(\theta_{\text{tar}}, \theta_{\text{dir}}) / \theta_{\text{dir}}$$

최종 회전각: 필요한 만큼, 단 $\hat{\mathbf{v}}$ 를 넘지 않음

$$\tilde{\mathbf{h}} = r \cdot \text{slerp}(\hat{\mathbf{h}}, \hat{\mathbf{v}}; \lambda)$$

$\hat{\mathbf{h}}$ 에서 $\hat{\mathbf{v}}$ 까지 λ 비율 만큼 회전, 크기 r 은 그대로

예: $\theta_{\text{tar}} = 40^\circ, \theta_{\text{dir}} = 100^\circ \rightarrow \hat{\mathbf{v}}$ 쪽으로 40° 회전

게이트가 켜진 질문만, 특징 벡터를 $\hat{\mathbf{v}}$ 쪽으로 필요한 각도만큼 회전 (크기와 가중치는 그대로)

R-MUSE

Towards Reasoning-Preserving Unlearning in Multimodal Large Language Models (CVPR 2026)

❖ R-MUSE의 장단점은 무엇일까?

장점

- 학습 없이 Forget 정확도와 추론 누수를 함께 낮춤
추론 누수(RIL) 78.5% → 38.5% (비교 방법은 모두 51.2% 이상)
- 새 평가 지표 및 벤치마크 제시

따져볼 점

- 공개 가중치면 수정 코드만 떼어내도 원상복구
→ '삭제'보다 추론 시점 가드레일에 가까움

가중치는 그대로라면?



직업은?

수정 코드 켜

모르겠어요



직업은?

수정 코드 제거

환경과학자예요

Conclusion

Conclusion

Multimodal Unlearning: How can MLLMs forget across image and text?

❖ 세 편을 통해 무엇을 알 수 있을까?



느낀 점

무엇을 지울지부터 합의해야

입장 ①(시각 인식만)과 ②(두 모달리티 모두)는 같은 벤치마크에서도 성공 기준이 다름

'잊었다'의 기준은 무엇일까? 성능 만일까?

R-MUSE 대답은 언러닝 성공? or 가중치는 그대로이기 때문에 지식이 잔존?

지운 지식을 되살릴 수는 없나

R-MUSE는 수정 코드만 떼면 원상복구가 가능하며, 서비스 관점에서는 비용절감이지 않을까?

References

Multimodal Unlearning: How can MLLMs forget across image and text?

- [1] Liu, Z., Dou, G., Yuan, X., Zhang, C., Tan, Z., & Jiang, M. (2025). Modality-Aware Neuron Pruning for Unlearning in Multimodal Large Language Models. In ACL.
- [2] Li, K., Li, W., Wu, D., Yang, L., Bai, J., Jia, J., & Xue, J. (2026). Cross-Modal Unlearning via Influential Neuron Path Editing in Multimodal Large Language Models. In AAAI.
- [3] Li, H., Yu, M., Yao, J., Singh, P., Li, X., Wang, D., & Hu, L. (2026). Towards Reasoning-Preserving Unlearning in Multimodal Large Language Models. In CVPR.
- [4] Li, J., Wei, Q., Zhang, C., Qi, G., Du, M., Chen, Y., Bi, S., & Liu, F. (2024). Single Image Unlearning: Efficient Machine Unlearning in Multimodal Large Language Models. In NeurIPS.
- [5] Huo, J., Yan, Y., Zheng, X., Lyu, Y., Zou, X., Wei, Z., & Hu, X. (2025). MMUnlearner: Reformulating Multimodal Machine Unlearning in the Era of Multimodal Large Language Models. In Findings of ACL.
- [6] Wang, C., Li, Y., Feng, X., Chen, C., Zheng, X., & Yin, J. (2025). UMU-Bench: Closing the Modality Gap in Multimodal Unlearning Evaluation. In NeurIPS Datasets and Benchmarks Track.
- [7] Liu, Z., Dou, G., Jia, M., Tan, Z., Zeng, Q., Yuan, Y., & Jiang, M. (2025). Protecting Privacy in Multimodal Large Language Models with MLLMU-Bench. In NAACL.
- [8] Dontsov, A., Korzh, D., Zhavoronkin, A., Mikheev, B., Bobkov, D., Alanov, A., Rogov, O., Oseledets, I., & Tutubalina, E. (2025). CLEAR: Character Unlearning in Textual and Visual Modalities. In Findings of ACL.

Thank you