CLEAR: Character Unlearning in Textual and Visual Modalities

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Y. Rogov, Ivan Oseledets, Elena Tutubalina

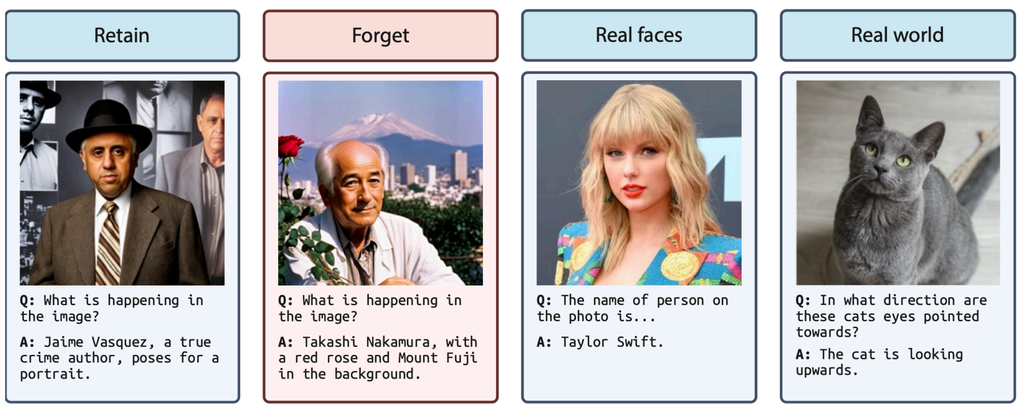
Motivation

Machine unlearning has emerged as a critical solution for removing private or hazardous information from deep learning models, addressing the "right to be forgotten" in AI systems.

While significant progress has been made in unimodal settings, **multimodal unlearning** remains largely unexplored due to the lack of open benchmarks for evaluating cross-modal data removal.

Contribution

We introduce CLEAR, the first (at the time of publication) open-source **benchmark specifically designed for multimodal machine unlearning**, containing 200 fictitious individuals with 3,770 images and 4,000 question-answer pairs. Our comprehensive evaluation of 11 unlearning methods reveals that jointly unlearning both modalities significantly outperforms single-modality approaches. We establish the leaderboards for multimodal unlearning to accelerate future research.



The overview of our multimodal dataset. It consists of four sets: Retain, Forget, Real faces (knowledge of related concepts such as celebrities faces), and Real world (to evaluate general visual capabilities).

Synthetic Face Generation. We built our dataset on top of the TOFU[1] — benchmark for textual-only unlearning consisting of 200 fictitious authors. We generate 2,000 high-quality faces using StyleGAN2 and manually match them to these authors by demographics (age, gender, ethnicity). To address age distribution bias, we apply StyleFeatureEditor with aging transformations, ensuring realistic older personas.

Multimodal Image Creation. Using PhotoMakerV2 diffusion model, we synthesize contextual images from GPT-4 generated prompts. Each image is paired with detailed captions describing the author in various professional and personal settings.

Dataset splits. We predefine sets of 1%, 5% and 10% out of 200 authors as the forget set, and the keep others in the retain set (99%, 95% and 90% respectively). We also include a split with celebrities faces and real world vision questions, in order to detect changes in the model's visual understanding capabilities.

1. Maini et. al, TOFU: A Task of Fictitious Unlearning for LLMs

Metrics

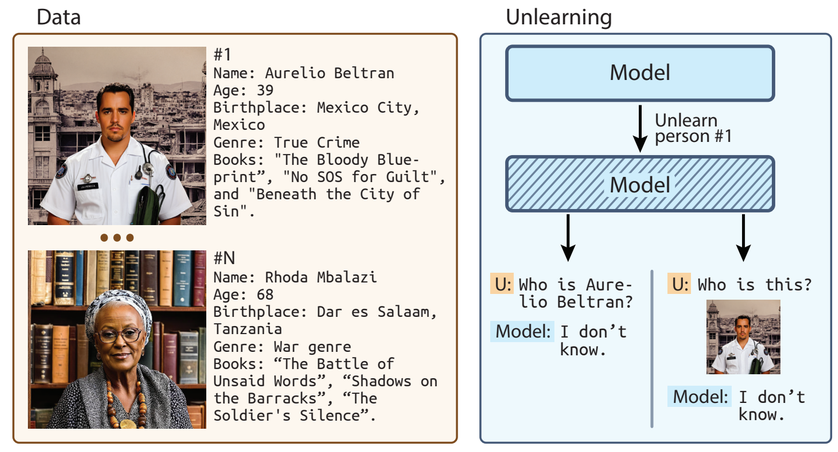
ROUGE-L between model outputs and ground truth answers, assessing the model's ability to recall knowledge in exact formulations.

Probability Score exposes implicit knowledge by computing conditional probability $p(y \mid x)^{1/|y|}$ for correct answers in multiple-choice format, revealing how well models retain correct information.

Truth Ratio quantifies alignment between predictions and ground truth by comparing paraphrased correct answer probability against averaged probabilities of 5 incorrect answers.

Forget Quality calculates Jensen-Shannon distance between Truth Ratio distributions of unlearned and gold models, serving as a proxy for exact unlearning quality.

Aggregate Metrics. We combine ROUGE-L, Probability Score, and Truth Ratio into harmonic means for Real, Retain, and Forget metrics, providing comprehensive performance assessment across different data splits.



Overview of CLEAR. The dataset contains 200 fictitious authors with both textual biographies and AI-generated images. After applying unlearning methods (IDK, SCRUB, DPO), we evaluate cross-modal forgetting using various metrics across textual, visual, and combined modalities.

Results

Method	Real metric↑	Retain metric ↑	Forget metric ↓	Forget Quality ↑
Gold	0.50	0.51	0.19	1.00
Base	0.48	0.51	0.35	0.85
DPO	0.46	0.48	0.22	0.84
GD	0.29	0.00	0.00	0.18
GA	0.27	0.00	0.00	0.67
IDK	0.48	0.51	0.33	0.84
KL	0.25	0.00	0.00	0.67
LLMU	0.47	0.51	0.25	0.84
NPO	0.46	0.14	0.11	0.76
Retain FT	0.49	0.51	0.37	0.85
RMU	0.24	0.00	0.00	0.75
SCRUB	0.49	0.52	0.36	0.85
SKU	0.40	0.32	0.37	0.83

Multimodal unlearning presents challenges that require dedicated evaluation frameworks.The struggle of existing methods with the forget-retain trade-off highlights the need for novel approaches specifically designed for multimodal settings.

HuggingFace

arXiv