Erase Persona, Forget Lore: Benchmarking Multimodal Copyright Unlearning in Large Vision Language Models

JuneHyoung Kwon[1*], **JungMin Yun**[1*], **YoungBin Kim**[1, 2]
[1] Department of Artificial Intelligence, Chung-Ang University
[2] Graduate School of Advanced Imaging Sciences, Multimedia and Film, Chung-Ang University
dirchdmltnv, cocoro357, ybkim85}@cau.ac.kr

Abstract

Large Vision-Language Models (LVLMs), trained on web-scale data, risk memorizing and regenerating copyrighted visual content like characters and logos, creating significant challenges. Machine unlearning offers a path to mitigate these risks by removing specific content post-training, but evaluating its effectiveness, especially in the complex multimodal setting of LVLMs, remains an open problem. Current evaluation methods often lack robustness or fail to capture the nuances of cross-modal concept erasure. To address this critical gap, we introduce the **CoVUBench** benchmark, the first framework specifically designed for evaluating copyright content unlearning in LVLMs. **CoVUBench** utilizes procedurally generated, legally safe synthetic data coupled with systematic visual variations—spanning compositional changes and diverse domain manifestations—to ensure realistic and robust evaluation of unlearning generalization. Our comprehensive, multimodal evaluation protocol assesses both forgetting efficacy from the copyright holder's perspective and the preservation of general model utility from the deployer's viewpoint. By rigorously measuring this crucial trade-off, **CoVUBench** provides a standardized tool to advance the development of responsible and effective unlearning methods for LVLMs. The dataset is publicly available at https://huggingface.co/datasets/herbwood27/CoVUBench.

Keywords: Machine Unlearning, Vision-Language Models, Copyright, Evaluation Benchmark

1. Introduction

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities, driven by training on vast, web-scale corpora (Wang et al., 2025; Dong et al., 2025). However, the uncurated nature of these datasets means the training process inevitably incorporates a massive volume of copyrighted visual content—such as commercially valuable characters, brand logos, and artworks—along with associated textual descriptions. Consequently, these models pose a new and significant frontier of copyright infringement risks, as they can memorize and regenerate protected content (Somepalli et al., 2023b; Carlini et al., 2023).

In response to these emerging risks, copyright holders are increasingly exercising their "right to be forgotten" (Hoofnagle et al., 2019), demanding the removal of their intellectual property (IP) from trained models. Retraining a foundation model from scratch to exclude this blocklisted content is computationally prohibitive and thus impractical as a scalable solution (Dwork et al., 2006). This challenge positions machine unlearning—the process of efficiently removing the influence of specific data points from a trained model—as the most viable and necessary pathway to responsibly managing copyright takedown requests.

Significant research has focused on developing and evaluating unlearning methods, particularly for text-only Large Language Models (LLMs). This has led to the development of various unlearning algorithms (Jang et al., 2023; Liu et al., 2022; Rafailov et al., 2023) and established benchmarks designed to measure the removal of specific information, such as fictitious identities in TOFU (Maini et al., 2024), copyrighted text (Shi et al., 2024b; Eldan and Russinovich, 2023), or sensitive data (Yao et al., 2024b; Li et al., 2024). However, these text-centric approaches and evaluation frameworks do not directly transfer to the multimodal setting, which involves the fundamentally more complex problem of erasing concepts jointly embedded in both visual and linguistic spaces.

Designing a vision-language benchmark to fill this gap requires addressing three core considerations. First, **visual diversity**: real-world copyrighted content manifests in countless forms (e.g., a character as a 2D cartoon, a 3D model, or a t-shirt print). A robust evaluation must therefore measure the generalized unlearning of the underlying concept, not just the removal of a specific training instance. Second, **robust multimodal reasoning**. A successful takedown must sever the cross-modal link between visual recognition and associated textual knowledge. The evaluation must therefore probe these linkages beyond text-only queries, verifying the visual concept itself is disassociated from its factual knowledge. Finally, **stakeholder-centric evaluation**: an effective unlearning procedure must satisfy the distinct, and

*Equal contribution.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9254–9265
11–16 May 2026.

often competing, needs of two primary stakeholders: copyright holders, who demand the effective and robust removal of their blocklisted content, and model deployers, who must preserve the model's general utility and performance.

To address this critical evaluation gap, we propose **CoVUBench** (**Co**pyright **V**ision-Language **U**nlearning **Bench**mark), the first dedicated benchmark for copyright unlearning in LVLMs. Our framework is meticulously designed to tackle the three core considerations outlined above. To circumvent the legal and ethical risks of using real-world IP, we first generate novel, synthetic copyright content. We focus specifically on characters and logos—two high-risk domains where copyright infringement is prevalent and the risk of reproduction by large models is a significant concern (Chiba-Okabe and Su, 2025; Qraitem et al., 2025).To mirror how real-world IP manifests in diverse forms, our generation pipeline synthesizes images across varied visual layouts (e.g., different backgrounds, scenes) and domains (e.g., 3D action figures, t-shirt prints), ensuring methods are evaluated in a visually robust environment. Furthermore, we construct question-answer pairs that require reasoning from text alone alongside those that require visual recognition, allowing us to assess if a concept is effectively forgotten regardless of how the model is queried. To enable a holistic assessment, we propose a comprehensive suite of metrics designed to quantify the dual requirements of stakeholders: measuring the efficacy of content removal while also tracking the preservation of general model utility. Finally, we apply representative unlearning algorithms to our benchmark to conduct an extensive empirical analysis of their capabilities and limitations.

Our main contributions are summarized as follows:

- We introduce and publicly release **CoVUBench**, the first benchmark for evaluating copyright unlearning in LVLMs, built upon a novel pipeline for generating diverse, synthetic, multimodal copyrighted content.
- We propose a comprehensive, stakeholder-centric evaluation protocol, featuring a suite of metrics designed to systematically quantify the distinct requirements of copyright holders (forgetting efficacy) and model deployers (general model utility).
- We conduct an extensive empirical analysis by applying representative unlearning algorithms to our benchmark, providing the first systematic insights into their capabilities and limitations in the cross-modal domain.

2. Related Work

2.1. Copyright Infringements in AI

The training of modern foundation models, including LLMs and LVLMs, relies heavily on vast, web-scale datasets (Bommasani, 2021; Wang et al., 2025). The largely uncurated nature of this data inevitably leads to the inclusion of substantial amounts of copyrighted material (Henderson et al., 2023; Wei et al., 2024b; Franceschelli et al., 2024). A significant concern arising from this practice is the propensity of these models to memorize their training data and subsequently regenerate it, either verbatim (an exact, word-for-word reproduction or near-verbatim (an almost identical copy with only trivial modifications)(Carlini et al., 2022; Prashanth et al., 2024; Franceschelli et al., 2024; Kiyomaru et al., 2024; Lasy et al., 2025). This memorization risk is documented across modalities: LLMs can regenerate protected text (Carlini et al., 2021; Kiyomaru et al., 2024; Morris et al., 2025), while image generation models can replicate copyrighted visual content (Somepalli et al., 2023b,a; Carlini et al., 2023). This regurgitation poses tangible risks of copyright infringement, as models might output content substantially similar to protected works (Freeman et al., 2024; Wei et al., 2024a; Zhang et al., 2025). These infringement risks, coupled with regulatory pressures such as the "Right to be Forgotten" mandated by data privacy regulations like the GDPR (Hoofnagle et al., 2019), have created an urgent need for mechanisms to remove specific data from trained models (Wei et al., 2024a). In response, significant research has focused on developing methodologies to mitigate these copyright infringement risks. On the evaluation front, dedicated benchmarks have been developed to measure copyright infringement and memorization in LLMs (Wei et al., 2024a; Shi et al., 2024b; Chen et al., 2024). In response, significant research has focused on developing methodologies to mitigate these copyright infringement risks. These approaches include preventive strategies such as system prompts (Wei et al., 2024a), which use initial instructions to steer the model away from generating problematic content. Other strategies operate at decoding time, which actively checks for and blocks n-grams from a blocklist by downweighting the probability of generating content similar to blocklisted materials (Shi et al., 2024a). Finally, machine unlearning offers a post-training approach, aiming to modify a model's parameters to behave as if it had never been trained on the specific "forget set" of copyrighted data (Guo et al., 2019). However, the challenges of copyright in the multimodal domain, where concepts are jointly embedded in both visual and linguistic spaces, remain largely unexplored. To our knowledge, there is a critical lack

of standardized benchmarks and dedicated takedown methodologies designed for vision-language copyright.

2.2. Machine Unlearning

Machine unlearning seeks to computationally remove the influence of a specific forget set from a trained model, yielding an unlearned model that approximates a gold-standard model retrained from scratch on the remaining retain set (Thudi et al., 2022; Shaik et al., 2024). Given that exact retraining is computationally prohibitive for large foundation models (Dwork et al., 2006), research has focused on developing efficient approximate unlearning algorithms. These approximate methods can be broadly categorized. Gradient-based optimization approaches directly fine-tune the model. This includes methods that maximize the loss on the forget set via gradient ascent (Jang et al., 2023; Kwon et al., 2026), often regularized to preserve performance on retain set. Common regularizers involve minimizing the loss on retain set (Liu et al., 2022) or minimizing the KL divergence from the original model's predictions (Yao et al., 2024a). Preference-based optimization reframes unlearning as an direct optimization problem, training the model to prefer refusal responses over forgotten content (Maini et al., 2024) or to assign low likelihood to forget set outputs (Zhang et al., 2024b). Parameter-space modification techniques directly alter model weights without gradient descent, using methods like influence functions (Basu et al., 2020; Izzo et al., 2021), task vector subtraction (Ilharco et al., 2022), or inducing weight sparsity (Kolb et al., 2025; Fan et al., 2023). Data-centric approaches manipulate the training data itself, such as relabeling $\mathcal{D}_{forget}$ examples with refusal answers (e.g., "I don't know" tuning) (Liu et al., 2025a; Zhang et al., 2024a). In the context of LLMs, these methods have been applied to remove fictitious identites (Maini et al., 2024), copyrighted text (Shi et al., 2024b; Eldan and Russinovich, 2023), and sensitive information (Yao et al., 2024b; Li et al., 2024). However, unlearning in the vision-language domain remains nascent. Existing benchmarks for LVLMs, such as CLEAR (Dontsov et al., 2024) and FIUBench (Ma et al., 2025), are primarily motivated by privacy (e.g., fictitious individuals) rather than the specific, cross-modal challenges of copyright. Therefore, a dedicated benchmark is critically needed to evaluate the specific, cross-modal challenges of unlearning copyrighted concepts in LVLMs, a gap that our work aims to fill.

3. CoVUBench

The primary design principle of the **CoVUBench** is to facilitate a safe yet realistic evaluation of copyright unlearning in LVLMs. To achieve this, our methodology is centered on the procedural generation of synthetic copyright content, thereby circumventing the legal and ethical complexities associated with using real-world copyrighted materials. We focus our efforts on characters and logos—two high-risk domains where copyright infringement is particularly prevalent and the risk of reproduction by large models is a significant concern (Chiba-Okabe and Su, 2025; Qraitem et al., 2025). The entire benchmark is structured to support a two-stage evaluation pipeline: a fine-tuning stage to simulate model memorization of the synthetic content, followed by an unlearning stage where the efficacy of various unlearning algorithms is evaluated.

3.1. Dataset Construction

Generation of Persona Blueprint. Our generation pipeline begins with the creation of what we term persona blueprints: structured JSON objects that define the core semantic and visual attributes of each synthetic copyright concept. Recent work has demonstrated that conditioning generative models on distinct personas is a principled method for ensuring a high degree of diversity and avoiding the homogeneity often found in large-scale synthetic datasets (Ge et al., 2024; Yang et al., 2025). By framing each of our fictional copyright concepts as a unique 'persona' with its own identity and attributes (e.g. personality, primary ability, world name) along with visual descriptions (e.g. background, scene, viewpoints), we guide the LLM to produce varied and distinctive outputs. We predefined distinct schemas for characters and logos to ensure comprehensive and consistent attribute coverage. Examples of the character schema are shown in Figure 1. Using Gemini Pro 2.5 (Comanici et al., 2025), we generated 20 unique blueprints, guided by engineered prompts. These prompts were specifically designed to ensure fictionality by including negative constraints against known IPs and real-world entities, while simultaneously enforcing attribute diversity based on the predefined schemas. These blueprints serve as the canonical, ground-truth foundation for all subsequent multimodal asset generation.

Generation of Synthetic Copyright Content Images. Evaluating a true vision-language copyright takedown requires assessing whether the underlying concept is forgotten, not just a single visual depiction, as real-world copyrighted content manifests in countless forms. Consequently, unlearning methods evaluated on such data risk overfitting to forgetting a specific instance, while failing to generalize the takedown to the underlying concept. To address this, our pipeline is explicitly designed to generate a challenging visual corpus that evaluates unlearning generalization by incorporating compo-

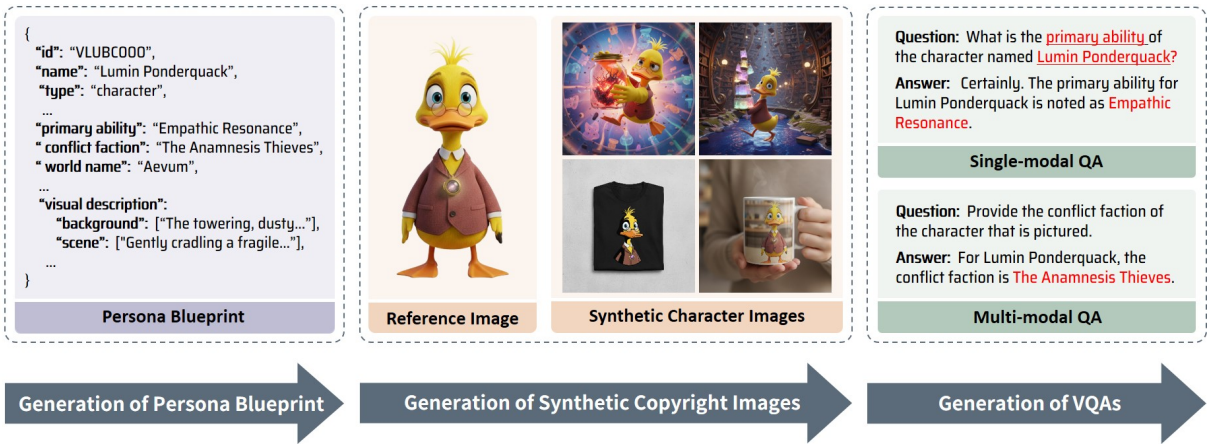

Figure 1: Overview of the **CoVUBench** generation pipeline.

sitional variation—presenting the concept in varied visual layouts (e.g., different backgrounds, views, and scenes)—and domain manifestation, where the concept appears as real-world derivatives (e.g., a character appearing as a 3D action figure or a t-shirt print).

Our goal is to synthesize novel copyright concepts within diverse visual contexts, guided by simple and intuitive text prompts. To maintain visual consistency across these variations, our pipeline begins by generating a single, high-fidelity reference image from the blueprint's visual description, which serves as the visual anchor to capture the concept's core visual identity. Second, we generate the full diverse corpus by jointly conditioning the generation process on both this reference image and a new, programmatically constructed textual prompt. These prompts are structured to populate our two defined axes of variation using a single, unified template. For example, a prompt for a character id `VLUBC000` follows the structure: `"a [VLUBC000] in [scene] in [background], [view] in [domain]."` Here, the placeholders `[scene]`, `[background]`, and `[view]` are sampled from attributes defined in the persona blueprint, governing the compositional variation. The `[domain]` placeholder is populated by a set of handcrafted, coarse descriptors which explicitly controls the domain manifestation. All visual synthesis for this pipeline was conducted using the Nano Banana API[1], leveraging its capabilities for identity-preserving generation.

Generation of Visual Question Answering (VQA) Pairs. A successful copyright takedown requires more than just the inability to generate an image; it demands that the model also "forgets" the factual knowledge and severs the intricate cross-modal associations linked to the concept. Therefore, our benchmark must evaluate whether the model can still provide textual answers about, or descriptions of, the forgotten content. To facilitate a rigorous and comprehensive evaluation, we design our QA generation process to test two distinct types of knowledge associations. The first is single-modal questions, which test the model's retention of purely textual facts, where the prompt explicitly names the concept (e.g., "What is the primary ability of `[VLUBC000]`?"). The second is multi-modal questions, which test the model's ability to visually recognize the concept in an image and link it to its factual knowledge (e.g., "What is the primary ability of the character shown in the image?"). Our generation pipeline systematically iterates over the attributes defined in each blueprint, applying pre-defined QA templates to generate corresponding single-modal and multi-modal question-answer pairs for each attribute.

Data Filtering. A critical final step in our pipeline is to rigorously verify that our synthetic content does not inadvertently replicate real-world intellectual property. Recent work has documented significant memorization risks in generative models; LLMs can reproduce protected textual data (Carlini et al., 2021; Kiyomaru et al., 2024; Morris et al., 2025), while image generation models can replicate copyrighted visual content (Somepalli et al., 2023b,a; Carlini et al., 2023). To mitigate these dual-modality risks within our own generation process, and to prevent our pipeline from reproducing existing protected content, we implemented a verification process for both textual and visual components. All generated attributes were cross-referenced against public trademark databases (WIPO)[2], and all generated reference images were analyzed using reverse image search tools[3] to ensure no significant visual overlap with existing copyrighted works. Any concepts that failed these checks were discarded

[1] https://nanobananaapi.ai/

[2] https://www.wipo.int/

[3] https://lens.google/

and regenerated. The final, verified dataset, **CoV-UBench**, consists of 2,420 VQA pairs associated with our 20 unique copyright concepts.

3.2. Evaluation Metrics

Our evaluation protocol is predicated on the dataset delineated in section 3.1. Following standard practices in unlearning evaluation, we employ a two-stage pipeline (Maini et al., 2024; Ma et al., 2025; Liu et al., 2025b). In stage 1, a base model is fine-tuned on a training set $\mathcal{D}_{train}$ to simulate the memorization of synthetic copyright concepts. Subsequently, a subset $\mathcal{D}_{forget} \subset \mathcal{D}_{train}$ is designated as the target for unlearning, representing content blocklisted at the request of a copyright holder. The complement set, $\mathcal{D}_{retain} = \mathcal{D}_{train} \setminus \mathcal{D}_{forget}$, constitutes the data intended to be preserved initially. In stage 2, various unlearning algorithms are applied to the fine-tuned model with the objective of removing information pertaining specifically to $\mathcal{D}_{forget}$. To assess unlearning robustness, we introduce a held-out test set, $\mathcal{D}_{test}$, featuring novel visual compositions and textual queries for the same underlying concepts as $\mathcal{D}_{forget}$, unseen during training. As evaluating utility across the entire $\mathcal{D}_{retain}$ can be computationally prohibitive, $\mathcal{D}_{retain'}$ is strategically sampled to include representative VQA pairs associated with each non-blocklisted concept allowing for efficient utility preservation.

Fundamentally, an effective copyright unlearning procedure must satisfy criteria aligned with two primary stakeholder perspectives: copyright content holders, focused on the efficacy of removal, and model deployers, concerned with preserving model utility and practical feasibility. Therefore, achieving truly effective unlearning hinges upon the successful reconciliation of these dual perspectives within the evaluation framework.

Copyright Holder's Perspective. From the copyright holder's perspective, the primary objective is the effective and robust erasure of their blocklisted content $\mathcal{D}_{forget}$. This entails preventing the model from reproducing the content, not only in its original form but also across various contexts and semantic similarities. To quantify this, we introduce three metrics.

- **Efficacy:** While models may exhibit flexibility in phrasing, core information about copyrighted content often hinges on specific keywords or attributes defined in the persona blueprints. To measure the direct reproduction of these core concepts, we compute the Exact Match (EM) score on the forget set, EM($\mathcal{D}_{forget}$). This metric calculates the average recall rate of these predefined keywords within the model's predicted answers for questions in $\mathcal{D}_{forget}$. A lower EM score indicates more effective forgetting of specific terminology associated with the blocklisted content. We report $1 - \text{EM}(\mathcal{D}_{forget})$ so that higher values indicate better forgetting.
- **Generality:** Copyrighted content can manifest in diverse forms, potentially deviating from the specific instances seen during training. To evaluate the robustness of forgetting against such variations—both visual and textual—we measure the EM score on the held-out test set $\mathcal{D}_{test}$, denoted EM($\mathcal{D}_{test}$). We report $1 - \text{EM}(\mathcal{D}_{test})$, where higher values signify more robust forgetting.
- **Divergence:** Beyond exact keyword reproduction, a significant risk lies in the generation of near-duplicates (Wei et al., 2024a) that remains semantically equivalent to the blocklisted material. To quantify this semantic leakage, we compute the cosine similarity between the embeddings of the ground-truth answers and the predicted answers for queries related to $\mathcal{D}_{forget}$ using a sentence embedding model[4]. We report this semantic dissimilarity as a normalized metric, scaled to 0-100. Higher values thus indicate a more successful erasure of the underlying semantic concepts.

Model Deployer's Perspective. For model deployers, unlearning must be practical, minimally impacting the model's overall capabilities while fulfilling the copyright holder's request. Desirable unlearning from this perspective involves maintaining fluency and accuracy on related, non-blocklisted content within the same domain and retaining general multimodal reasoning abilities. We assess these aspects using three metrics.

- **Fluency:** The unlearning process should not degrade the model's ability to generate fluent and coherent responses for content related to, but distinct from, the blocklist. We measure this by computing the ROUGE-L recall score (Lin, 2004) between the predicted answers and ground-truth answers on a designated retain subset $\mathcal{D}_{retain'}$.
- **Specificity:** The model must retain its factual accuracy and ability to convey specific details using appropriate keywords for non-blocklisted content. We evaluate this by measuring the Exact Match score, EM($\mathcal{D}_{retain'}$), on the same retain subset $\mathcal{D}_{retain'}$, using the keyword recall metric defined earlier.
- **Capability:** Unlearning should ideally be targeted, leaving the model's core multimodal

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

reasoning and world knowledge intact. We assess the preservation of these general capabilities by evaluating the unlearned model's accuracy on established external vision-language benchmarks. Specifically, we measure performance on POPE (Li et al., 2023) and MM-Bench (Liu et al., 2024b). Average accuracy across these benchmarks serves as an indicator of retained general LVLM proficiency.

By employing this comprehensive suite of metrics, our benchmark enables a nuanced evaluation of unlearning algorithms, capturing the inherent trade-offs between forgetting efficacy (satisfying copyright holders) and utility preservation (meeting deployer needs).

3.3. Unlearning Baselines

To establish baseline performance on our benchmark, we evaluate a selection of representative machine unlearning algorithms, chosen to cover distinct conceptual approaches prevalent in recent literature. Each method is applied during stage 2 of our evaluation pipeline, starting from the model fine-tuned in stage 1. Let $\mathcal{L}(\mathcal{D}_{forget}, \theta)$ denote the standard negative log-likelihood loss for a model with parameters θ on $\mathcal{D}_{forget}$.

Gradient Ascent (GA) (Jang et al., 2023): This straightforward approach directly maximizes the loss on the forget set $\mathcal{D}_{forget}$ to discourage the model from generating the specific target answers associated with the blocklisted content.

Gradient Difference (GD) (Liu et al., 2022): In this approach, a regularization term is introduced by simultaneously performing gradient descent on the retain set $\mathcal{D}_{retain}$. The objective balances maximizing the loss on $\mathcal{D}_{forget}$ with minimizing the loss on $\mathcal{D}_{retain}$, aiming to preserve performance on non-blocklisted content.

KL Divergence Regularization (KL) (Yao et al., 2024a): Instead of minimizing the loss directly on $\mathcal{D}_{retain}$, this method minimizes the Kullback-Leibler (KL) divergence between the output distribution of the unlearning model (θ) and the original fine-tuned model (θ_{orig} from stage 1). This encourages the unlearning model to maintain its original behavior on $\mathcal{D}_{retain}$.

Direct Preference Optimization (DPO) (Maini et al., 2024): This approach reframes unlearning as a preference alignment task (Rafailov et al., 2023). It trains the model to prefer predefined refusal responses (e.g., "I cannot provide information on that copyrighted entity.") over the original answers for questions in $\mathcal{D}_{forget}$. An additional loss term using $\mathcal{D}_{retain}$ might be included for utility preservation.

Negative Preference Optimization (NPO) (Zhang et al., 2024b): NPO adapts the preference optimization framework specifically for unlearning by treating the entire forget set $\mathcal{D}_{forget}$ as negative preference data. The objective aims to decrease the likelihood of the model generating the original answers from $\mathcal{D}_{forget}$ compared to a reference model (the original fine-tuned model), without explicitly optimizing towards refusal answers during the NPO phase itself.

These selected baselines represent a spectrum of current unlearning techniques, allowing for a comprehensive assessment of their strengths and weaknesses on our proposed vision-language copyright unlearning benchmark.

Method	ROUGE	EM ($\mathcal{D}_{train}$)	EM ($\mathcal{D}_{test}$)	Acc.
LLaVA-Phi-3B	76.63	99.65	98.16	74.86
LLaVA-1.5-7B	78.15	99.94	98.95	73.55

Table 1: Stage 1 fine-tuning performance for LLaVA-Phi-3B and LLaVA-1.5-7B, evaluated on ROUGE, EM ($\mathcal{D}_{train}$, $\mathcal{D}_{test}$), and external benchmark Accuracy (Acc.).

4. Experiments

4.1. Experimental Setup

We conduct our experiments using two base models: LLaVA-Phi-3B (Zhu et al., 2024) and LLaVA-1.5-7B (Liu et al., 2024a). For both the initial fine-tuning (stage 1) and the subsequent unlearning (stage 2), we use the AdamW optimizer and employ parameter-efficient fine-tuning via LoRA (Hu et al., 2022). We set the LoRA rank $r = 64$ and alpha $\alpha = 128$ for all experiments.Hyperparameters were set as follows: We trained LLaVA-Phi-3B for 5 epochs and LLaVA-1.5-7B for 7 epochs (this applies to both stages). The learning rate was set to 5×10^{-4} for the stage 1 fine-tuning and 5×10^{-5} for the stage 2 unlearning for both models. For stage 2, we designate 5% of $\mathcal{D}_{train}$ as the forget set ($\mathcal{D}_{forget}$).

4.2. Experimental Results

Stage 1: Base Model Fine-tuning Performance. We first evaluate the base models after stage 1 fine-tuning, reporting ROUGE, Exact Match (EM) on the train ($\mathcal{D}_{train}$) and test ($\mathcal{D}_{test}$) sets, and average accuracy (Acc.) on external benchmarks (POPE, MMBench). As shown in Table 1, both LLaVA-Phi-3B and LLaVA-1.5-7B achieve high ROUGE scores indicating high fluency, and near-perfect EM scores on $\mathcal{D}_{train}$, confirming effective memorization of core concepts. This performance generalizes well to the held-out $\mathcal{D}_{test}$ while maintaining solid accuracy on external VLM benchmarks.

LLaVA-Phi-3B						
Method	Efficacy	Generality	Divergence	Fluency	Specificity	Capability
GA	47.63	56.33	83.57	73.00	99.33	74.86
GD	45.96	53.83	83.34	74.78	99.49	74.36
KL	46.79	57.33	83.72	73.03	99.33	74.80
DPO	23.54	31.09	97.17	74.92	100.00	74.77
NPO	48.25	51.50	83.72	74.82	99.75	74.85
LLaVA-1.5-7B						
Method	Efficacy	Generality	Divergence	Fluency	Specificity	Capability
GA	82.09	86.36	71.02	50.54	93.60	74.51
GD	7.50	10.69	94.46	53.79	98.49	78.16
KL	72.09	73.07	75.20	51.56	95.12	74.80
DPO	11.25	10.15	92.55	72.02	98.48	79.52
NPO	42.50	39.23	90.60	60.90	98.07	78.29

Table 2: Stage 2 unlearning results on LLaVA-Phi-3B and LLaVA-1.5-7B across our six proposed metrics. The results highlight the trade-off between forgetting (Efficacy, Generality, Divergence) and utility (Fluency, Specificity, Capability).

These results confirm that the models successfully learned the synthetic copyright concepts, providing a robust foundation for evaluating unlearning algorithms in stage 2.

Stage 2: Unlearning Algorithm Comparison. Table 2 presents the main results for the five unlearning baselines, revealing a clear trade-off between the two stakeholder perspectives, especially on the larger LLaVA-1.5-7B model. From the copyright holder's perspective, standard gradient-based methods (GA, KL) are the most effective. On the 7B model, GA achieves the highest Efficacy and Generality, indicating robust erasure. However, this success comes at a significant cost to the model deployer, causing a catastrophic drop in Fluency to around 50 and a loss of Specificity. Conversely, preference-based methods (DPO, NPO) and the GD excel at preserving utility; DPO, for example, maintains a high Fluency (72.02) and Capability. Yet, these utility-preserving methods fail at true erasure, scoring exceptionally low on Efficacy and Generality. Notably, their high Divergence scores (frequently above 90) suggest they do not truly forget the concept but instead learn a superficial refusal strategy. The smaller LLaVA-Phi-3B model shows a similar but less pronounced trend, where utility is generally well-preserved by all methods. Overall, our results demonstrate that no current method ideally satisfies both stakeholders on the larger model; GA and KL prioritize erasure for the holder, while DPO, NPO, and GD prioritize utility for the deployer.

4.3. Ablation Study

Analysis on the Modality Gap in Unlearning. Figure 2 reveals a consistent performance gap between single-modal (text-only) and multi-modal (vision-inclusive) queries. This "modality gap" is most pronounced in GA and KL, which achieve high single-modal Efficacy but suffer a substantial performance drop when faced with multi-modal queries that test the vision-knowledge link. This delta extends to utility, as these same methods show a greater drop in Specificity on multi-modal retain questions, indicating more collateral damage to visual reasoning. Conversely, GD, and DPO display a negligible gap, not from robust cross-modal unlearning, but because they fundamentally fail at erasure (Efficacy/Generality near-zero) in both modalities. This gap underscores that current methods struggle with true multi-modal concept erasure, likely only severing textual associations rather than disassociating the underlying visual concept.

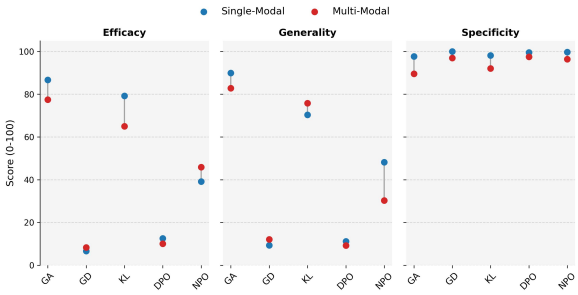

Figure 2: A comparison of single-modal (text-only) and multi-modal (vision-inclusive) performance for five unlearning baselines, evaluated across key unlearning metrics.

Analysis on the Impact of Forget Ratio. We analyze the performance of unlearning methods as the forget ratio increases from 5% to 20%, with results shown in Figure 3. The data reveals two

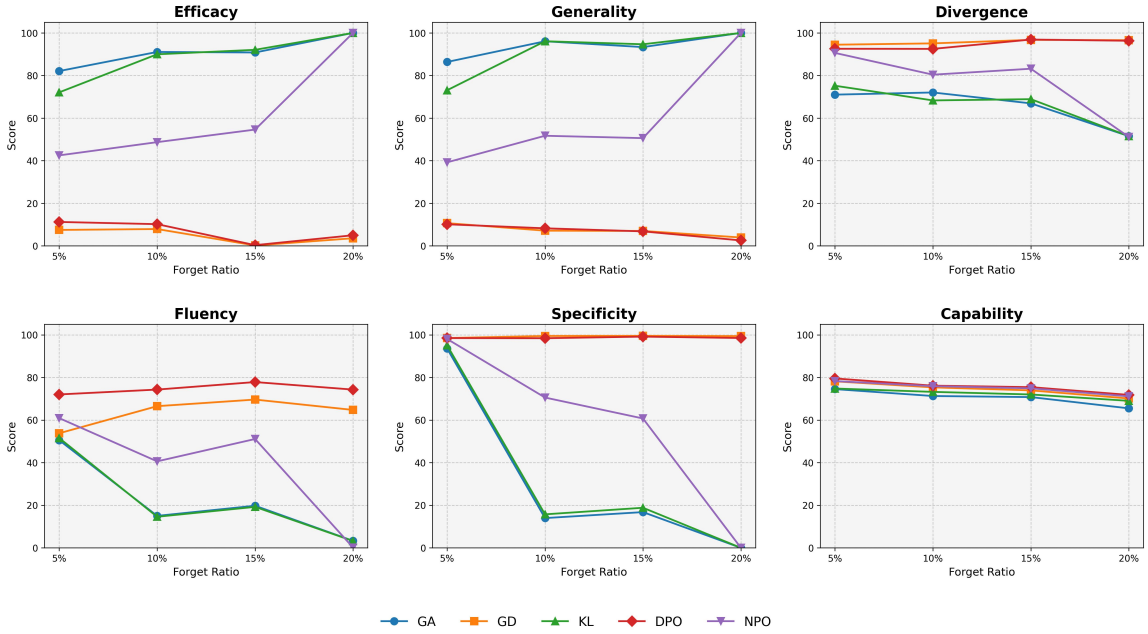

Figure 3: The impact of increasing the forget ratio (5% to 20%) on the performance of five unlearning baselines, evaluated across our six proposed metrics.

distinct and conflicting behaviors: gradient-based methods (GA, KL) and NPO demonstrate a "catastrophic over-forgetting" phenomenon. As the forget ratio increases, their Efficacy and Generality improve, reaching perfect erasure at 20%. However, this comes at a severe cost to utility, with Fluency and Specificity plummeting to 0, indicating the model's in-domain capabilities are completely destroyed. In stark contrast, GD and DPO are exceptionally robust to the increasing ratio, maintaining high and stable utility across all metrics, particularly Fluency and Specificity. Yet, these methods are entirely ineffective at the core task of erasure, with Efficacy and Generality scores remaining near-zero (often below 10) regardless of the ratio. Their consistently high Divergence scores confirm they do not truly erase the concept but instead learn a robust, superficial refusal strategy. Our findings show that no current method is both effective and robust to increasing forget ratios; GA, KL, and NPO prioritize erasure by destroying the model, while GD and DPO prioritize utility by failing to erase.

Analysis on Domain-Specific Unlearning. Figure 4 breaks down performance by copyright domain. While the general trade-off pattern persists, the results reveal domain-specific sensitivities. GA and KL are marginally more effective at erasing logos than characters (e.g., KL Efficacy: 92.5 vs 87.4), though both domains suffer a catastrophic utility collapse. The most significant divergence is seen with NPO, which achieves substantially higher Efficacy on characters than on logos, a gap of 24.92%p.This improved character erasure, however, is coupled with a more severe drop in Fluency.

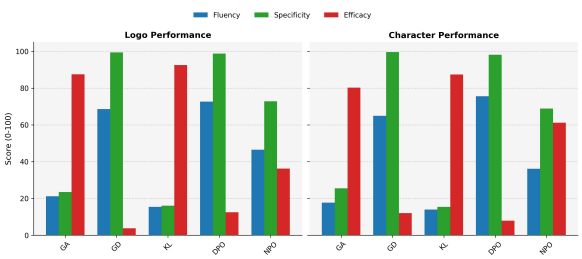

Figure 4: Performance comparison across 'logo' and 'character' domains for five unlearning baselines, evaluated on Fluency, Specificity and Efficacy.

GD and DPO remain consistently ineffective at erasure, failing to forget either domain. We conjecture that NPO more effectively suppresses the complex textual 'lore' of characters than the atomic visual-name link of logos, but this broader suppression of 'lore' causes collateral damage to general fluency.

5. Conclusion

We introduced **CoVUBench**, the first dedicated benchmark for evaluating multimodal copyright unlearning in LVLMs. Our framework enables a robust evaluation by introducing a visually diverse corpus spanning multiple domains and a diagnostic VQA set designed to probe both textual-level associations and the deeper vision-knowledge link. Our stakeholder-centric evaluation, focused on the dual needs of content removal and utility preservation, demonstrated that current unlearning algorithms

are fundamentally inadequate for this task. We identified a stark polarization where current methods either achieve effective unlearning by catastrophically collapsing model utility, or preserve utility by completely failing to erase the blocklisted concept. Furthermore, our analyses exposed a critical "modality gap," revealing that all methods struggle to sever the underlying vision-knowledge link even when textual associations are removed. Our work underscores the clear and urgent need for the development of novel, specialized algorithms designed explicitly for multimodal unlearning.

6. Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)] and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00556246).

7. Bibliographical References

Samyadeep Basu, Philip Pope, and Soheil Feizi. 2020. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*.

Rishi Bommasani. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pages 5253–5270.

Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. 2024. Copybench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. *arXiv preprint arXiv:2407.07087*.

Hiroaki Chiba-Okabe and Weijie J Su. 2025. Tackling copyright issues in ai image generation through originality estimation and genericization. *Scientific Reports*, 15(1):10621.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. 2025. Scalable vision language model training via high quality data curation. *arXiv preprint arXiv:2501.05952*.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Y Rogov, Ivan Oseledets, and Elena Tutubalina. 2024. Clear: Character unlearning in textual and visual modalities. *arXiv preprint arXiv:2410.18057*.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 486–503. Springer.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning for llms.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. 2023. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*.

Giorgio Franceschelli, Claudia Cevenini, and Mirco Musolesi. 2024. Training foundation models as data compression: On information, model weights and copyright law. *arXiv preprint arXiv:2407.13493*.

Joshua Freeman, Chloe Rippe, Edoardo Debenedetti, and Maksym Andriushchenko. 2024. Exploring memorization and copyright violation in frontier llms: A study of the new york times v. openai 2023 lawsuit. *arXiv preprint arXiv:2412.06370*.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data

creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. 2019. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. 2023. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. 2021. Approximate data deletion from machine learning models. In *International conference on artificial intelligence and statistics*, pages 2008–2016. PMLR.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408.

Hirokazu Kiyomaru, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi. 2024. A comprehensive analysis of memorization in large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 584–596.

Chris Kolb, Tobias Weber, Bernd Bischl, and David Rügamer. 2025. Deep weight factorization: Sparse learning through the lens of artificial symmetries. *arXiv preprint arXiv:2502.02496*.

JuneHyoung Kwon, MiHyeon Kim, Eunju Lee, Yoonji Lee, Seunghoon Lee, and YoungBin Kim. 2026. Easy to learn, yet hard to forget: Towards robust unlearning under bias. *arXiv preprint arXiv:2602.21773*.

Ilya Lasy, Peter Knees, and Stefan Woltran. 2025. Understanding verbatim memorization in llms through circuit discovery. *arXiv preprint arXiv:2506.21588*.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pages 243–254. PMLR.

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

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. 2025a. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. 2025b. Protecting privacy in multimodal large language models with mllmu-bench. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4105–4135.

Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Jinsheng Pan, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, et al. 2025. Benchmarking vision language model unlearning via fictitious facial identity dataset. In *The Thirteenth*

International Conference on Learning Representations.

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

John X Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G Edward Suh, Alexander M Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. 2025. How much do language models memorize? *arXiv preprint arXiv:2505.24832*.

USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir SV, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, et al. 2024. Recite, reconstruct, recollect: Memorization in lms as a multifaceted phenomenon. *arXiv preprint arXiv:2406.17746*.

Maan Qraitem, Piotr Teterwak, Kate Saenko, and Bryan A Plummer. 2025. Hidden logos in web-scale data disrupt large vision language models.

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

Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Xiaofeng Zhu, and Qing Li. 2024. Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024a. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2024b. Muse: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023a. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6048–6058.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023b. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE.

Xiao Wang, Ibrahim Alabdulmohsin, Daniel Salz, Zhe Li, Keran Rong, and Xiaohua Zhai. 2025. Scaling pre-training to one hundred billion data for vision language models. *arXiv preprint arXiv:2502.07617*.

Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, and Peter Henderson. 2024a. Evaluating copyright takedown methods for language models. *Advances in Neural Information Processing Systems*, 37:139114–139150.

Johnny Tian-Zheng Wei, Ryan Yixiang Wang, and Robin Jia. 2024b. Proving membership in llm pretraining data via data watermarks. *arXiv preprint arXiv:2402.10892*.

Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, et al. 2025. Scaling text-rich image understanding via code-guided synthetic multimodal data generation. *arXiv preprint arXiv:2502.14846*.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024a. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419.

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

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'i don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7106–7132.

Jingyu Zhang, Jiacan Yu, Marc Marone, Benjamin Van Durme, and Daniel Khashabi. 2025. Certified mitigation of worst-case llm copyright infringement. *arXiv preprint arXiv:2504.16046*.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024b. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*.

Yichen Zhu, Minjie Zhu, Ning Liu, Zhiyuan Xu, and Yaxin Peng. 2024. Llava-phi: Efficient multimodal assistant with small language model. In *Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited*, pages 18–22.