MMDU-Bench: Multi-modal Deep Unlearning Benchmark

Ziyang Zhang
Shandong University
ziyangzhang648@gmail.com

Abstract

Large Vision-Language Models (LVLMs) trained on web-scale data risk memorizing private, harmful, or outdated information, making machine unlearning increasingly important. Prior work mainly targets unimodal settings and isolated fact removal, overlooking the reality that knowledge is often deeply interconnected across modalities like text and images. We introduce **MMDU-Bench**, the first benchmark for **multi-modal deep unlearning**, where models must forget both explicit facts and implicit inferences made through cross-modal reasoning. Built on a large-scale synthetic knowledge graph with over 30k relations and 166k QA pairs, MMDU-Bench enables fine-grained evaluation of forgetting and retention. Experiments across five representative methods show that the majority achieve 30% Deep Forget Quality, revealing difficulty in removing entangled knowledge. We also observe large performance gaps between text-only and multi-modal unlearning, as well as a trade-off where stronger forgetting often leads to loss of related knowledge. MMDU-Bench highlights these overlooked challenges and provides a foundation for developing more effective and reliable unlearning methods.

1 Introduction

Large Vision-Language Models (LVLMs), trained on large-scale internet data, have demonstrated remarkable performance in contextual understanding [Brown *et al.*, 2020; Zhu *et al.*, 2024c], question answering [Kamalloo *et al.*, 2023; Arefeen *et al.*, 2024], and reasoning [Wei *et al.*, 2022; DeepSeek-AI *et al.*, 2025]. However, their powerful capabilities also raise significant concerns: these models can inadvertently memorize and generate private [Kim *et al.*, 2023; Staab *et al.*, 2024], harmful [Li *et al.*, 2024a; Gong *et al.*, 2025], or misleading content [Dhingra *et al.*, 2022; Mousavi *et al.*, 2024]. Such risks underscore the growing importance of **machine unlearning** [Liu *et al.*, 2024b], which aims to selectively forget undesirable information while retaining useful knowledge—aligning with GDPR[1] and other legal or ethical requirements.

Recent works have proposed several unlearning methods to enable LVLMs to forget specific knowledge by fine-tuning on data that needs to be removed, such as Gradient Ascent [Jang *et al.*, 2023; Yao *et al.*, 2024] and Negative Preference Optimization [Zhang *et al.*, 2024]. Although these methods have achieved notable success in forgetting targeted knowledge, we found that this is far from sufficient. In real-world applications, knowledge is rarely isolated—it is often deeply interconnected and spans across multiple modalities. For instance, as shown in Figure 1, simply forgetting the explicit fact "Bob works at OpenAI" is not enough, as the model may still infer it from related clues such as "Bob is Amy's colleague" and "Amy works at OpenAI".

To this end, we introduce a new and more realistic challenge: **multi-modal deep unlearning**. This task goes beyond simply removing surface-level facts—it requires LVLMs to also forget the hidden connections and reasoning paths that could still lead to the target knowledge, across both textual and visual modalities.

However, existing benchmarks fall short in evaluating this deeper level of forgetting, particularly in multi-modal scenarios. To address this gap, we propose the **M**ulti-**M**odal **D**eep **U**nlearning **Bench**mark (**MMDU-Bench**), a new benchmark specifically designed to assess unlearning methods in the multi-modal deep unlearning setting.

Our contributions are summarized as follows:

- We introduce **MMDU-Bench**, a benchmark for evaluating multi-modal deep unlearning. It includes a large-scale synthetic knowledge graph with over **30k** relations and **166k** QA pairs, covering both explicit facts and reasoning paths across textual and visual modalities.
- We design structured forget and retain sets that support fine-grained evaluation. The forget set targets both *single-fact* and *multi-fact* scenarios, while the retain set includes neighboring knowledge, globally unrelated facts, and utility sets to assess the model's general knowledge.
- We conduct extensive experiments on five representative unlearning methods and two widely used LVLMs, evaluated under both text-only and multi-modal settings.

[1] https://gdpr-info.eu/

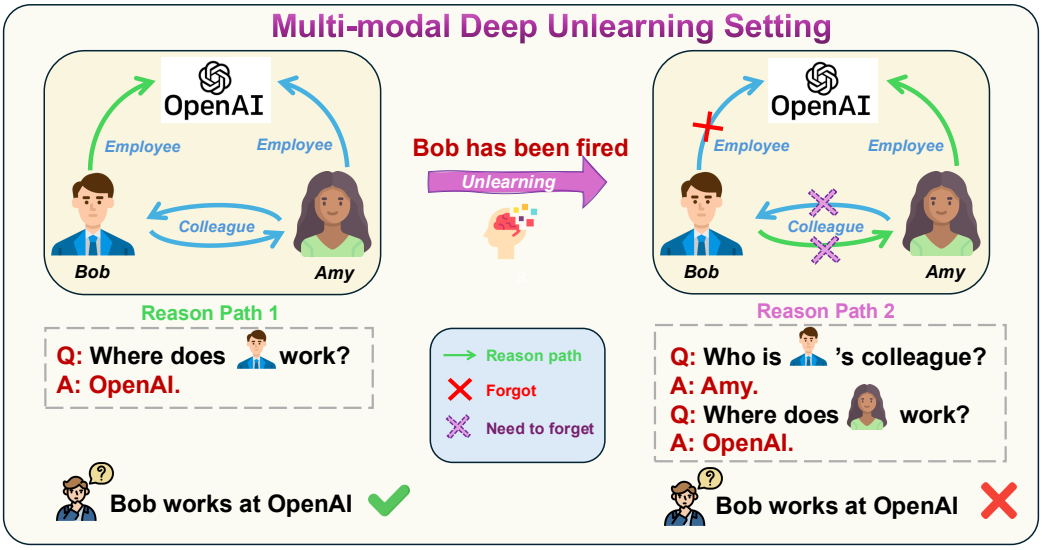

Figure 1: An example of multi-modal deep unlearning where knowledge is interconnected and spans across modalities.

Our analysis reveals key challenges, including the difficulty of forgetting intermediate knowledge involved in reasoning, and notable performance gaps across modalities.

2 Related Work

Large Vision-Language Models Large Vision-Language Models (LVLMs) combine visual encoders with large language models to jointly process image and text inputs. Recent models, such as MiniGPT-4 [Zhu *et al.*, 2024a], BLIP [Li *et al.*, 2022; Li *et al.*, 2023], LLaVA [Liu *et al.*, 2023; Liu *et al.*, 2024a], GPT-4V [OpenAI, 2023], and Qwen-VL [Bai *et al.*, 2023; Wang *et al.*, 2024; Bai *et al.*, 2025] have demonstrated increasing capabilities across diverse multimodal tasks, including image captioning [Agrawal *et al.*, 2019], visual question answering [Goyal *et al.*, 2019], and visual grounding [Kazemzadeh *et al.*, 2014].

Knowledge Unlearning for LVLMs Machine unlearning aims to selectively remove specific knowledge—such as personal information, copyrighted content, or harmful material—from a trained model while retaining unrelated capabilities. Prior work has explored unlearning in both the visual modality (e.g., removing specific images from classifiers) and the textual modality (e.g., forgetting content from books or documents). Techniques such as gradient ascent [Jang *et al.*, 2023; Yao *et al.*, 2024], preference optimization [Rafailov *et al.*, 2023; Zhang *et al.*, 2024], and task arithmetic [Ilharco *et al.*, 2023] have shown promising results in these settings. With the rise of LVLMs, unlearning across modalities becomes increasingly important. While a few recent studies have begun to explore unlearning in LVLMs [Ma *et al.*, 2024; Dontsov *et al.*, 2024], it remains unclear whether these methods can effectively handle cross-modal and entangled knowledge, highlighting the need for deeper investigation.

Unlearning Benchmarks for LVLMs Benchmarks are essential for systematically evaluating the effectiveness of unlearning methods. Several have been proposed for specific use cases: MUSE [Shi *et al.*, 2024] focuses on forgetting knowledge and corpora from real-world news and books, and introduces six evaluation criteria; TOFU [Maini *et al.*, 2024] targets fictional author biographies; CLEAR [Dontsov *et al.*, 2024] extends TOFU to the multi-modal setting; MLLMU [Liu *et al.*, 2025] targets unlearning in personal profiles using synthetic multimodal QA; RWKU [Jin *et al.*, 2024] addresses the unlearning of real-world knowledge embedded in LLMs; and WMDP [Li *et al.*, 2024b] focuses on safety-sensitive domains such as biosecurity, cybersecurity, and chemical safety;

While these benchmarks are well-designed for their respective tasks, they largely treat knowledge as isolated units to be forgotten. In practice, however, knowledge is often interconnected—forgetting a single fact is not sufficient if related context allows it to be reconstructed. A recent study [Wu *et al.*, 2024] introduces the concept of deep unlearning and proposes EDU-RELAT, a small-scale dataset designed to assess unlearning of relational knowledge. However, it is limited to the text modality, supports only a few relation types, and lacks diverse question formats. A concurrent work, FaithUn [Yang *et al.*, 2025], also explores unlearning interconnected knowledge, but it is again restricted to text and uses only multiple-choice QA to verify knowledge removal—an evaluation method that may be insufficient for the complexities of LVLMs and prone to bias [Shostack, 2024]. A detailed comparison with those benchmarks is listed in Table 1.

3 Preliminary

3.1 Multi-modal Knowledge Graph

Following prior works [Liu *et al.*, 2019; Zhu *et al.*, 2024b; Chen *et al.*, 2024], we formulate a **multi-modal knowledge graph** (MMKG) as a directed labeled graph $\mathcal{G} = (V, E, R)$, where $V = V_{\text{text}} \cup V_{\text{visual}}$ denotes the set of entities, with each entity categorized as either textual or visual. The set R

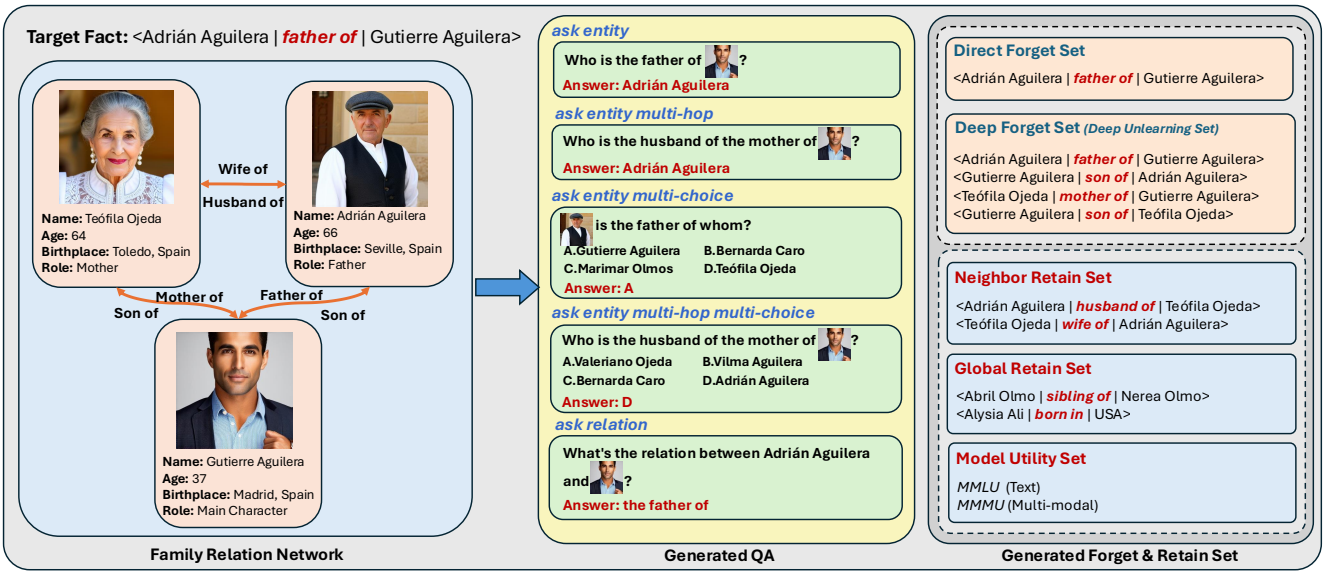

Figure 2: Example of MMDU-Bench QA generation and the construction of forget and retain sets.

contains all relation types, and $E \subseteq V \times R \times V$ represents the collection of directed edges (i.e., factual triples) of the form $\langle e_s, r, e_t \rangle$, indicating that a source entity e_s is connected to a target entity e_t via relation r.

3.2 Reasoning Paths

A **reasoning path** in the MMKG from entity e_s to e_t is defined as a sequence of directed or reversed edges:

$$p = \langle (e_s, r_1, e_1),\ (e_1, r_2, e_2),\ \ldots,\ (e_{n-1}, r_n, e_t) \rangle,$$

A path is valid if it satisfies the following conditions:

(1) *Traversability*: Each edge must appear in the MMKG either as a forward triple $(e_i, r_{i+1}, e_{i+1}) \in E$ or as a reversed triple $(e_{i+1}, r_{i+1}^{-1}, e_i) \in E$, where r_{i+1}^{-1} denotes the inverse of r_{i+1}.

(2) *Composability*: The relation sequence $(r'_1, \ldots, r'_n)$ must map to a composed relation $r_c \in R$ according to a rule in the composition set $\mathcal{R}_{\text{comp}}$, denoted as $\pi(p) = r_c$.

To avoid semantic ambiguity and inference loops, we restrict reasoning paths to be acyclic. Our benchmark uses a deterministic composition set $\mathcal{R}_{\text{comp}}$, ensuring that each valid path corresponds to at most one composed relation.

3.3 Multi-modal Deep Unlearning

Multi-modal deep unlearning aims to remove specific knowledge from LVLMs, ensuring it cannot be recovered by the model via alternative reasoning paths. To systematically study and evaluate this task, we represent the knowledge within LVLMs as structured triples in the previously defined multi-modal knowledge graph.

Formally, given a target fact $\tau = \langle e_a, r_{\text{target}}, e_b \rangle$, multi-modal deep unlearning seeks to eliminate not only τ itself but also any semantically equivalent inference that may reconstruct it through cross-modal reasoning.

Let $\mathcal{G}^*$ denote the **deductive closure** of the knowledge graph $\mathcal{G}$ under composition rules $\mathcal{R}_{\text{comp}}$, where:

$$\langle e_s, r_c, e_t \rangle \in \mathcal{G}^* \iff \exists\, p \in \mathcal{P}_{\mathcal{G}}(e_s, e_t) \text{ s.t. } \pi(p) = r_c,$$

with $\mathcal{P}_{\mathcal{G}}(e_s, e_t)$ denoting all valid reasoning paths from e_s to e_t, and $\pi(p)$ the composed relation derived from path p.

The unlearning operation $\mathcal{U}(\mathcal{G}, \tau)$ returns a modified graph $\mathcal{G}' = (V, E')$ such that $\tau \notin \mathcal{G}'^*$. This ensures the target fact cannot be recovered—either explicitly or via indirect multi-hop reasoning—in the deductive closure of the modified graph.

3.4 Deep Unlearning Set

We define a **deep unlearning set** as a subset of edges $\mathcal{F} \subseteq E$ such that, for a given target fact $\tau = \langle e_a, r_{\text{target}}, e_b \rangle$, the fact can no longer be inferred from the graph after removing $\mathcal{F}$; formally, $\tau \notin (\mathcal{G} \setminus \mathcal{F})^*$. In other words, a deep unlearning set effectively eliminates all valid reasoning paths that could lead to the inference of τ.

Among the many possible deep unlearning sets, we aim to find one that removes fewer edges while minimizing disruption to the model's unrelated knowledge. Specifically, we prioritize edges directly involved in the inference of τ, and avoid those that represent general or widely shared knowledge.

In practice, we employ Algorithm 1 to compute such a deep unlearning set. A formal proof of the algorithm's correctness is provided in Appendix A.

4 MMDU-Bench

To systematically study multi-modal deep unlearning, it is essential to construct a dataset as a multi-modal knowledge graph that enables structured reasoning and explicit evaluation of knowledge removal. However, existing datasets built from sources like Wikipedia [Auer *et al.*, 2007; Suchanek *et al.*, 2007] or Freebase [Bordes *et al.*, 2013; Toutanova *et al.*,

2015] are text-only and require costly annotation for multimodal alignment. Although MMKG [Liu *et al.*, 2019] augments these graphs with visual signals, it remains noisy and lacks explicit composition rules, making it unsuitable for controlled evaluation of unlearning.

To address these limitations, we introduce **MMDU-Bench**, a dataset built upon structured multi-modal knowledge graphs, specifically designed to support the study of multimodal deep unlearning in LVLMs. It avoids contamination from pretraining data, enables deterministic composition rules, and allows fine-grained control over graph structure and distribution, providing a basis for more systematic evaluation. An illustrative example is shown in Figure 2.

4.1 Dataset Construction

Our dataset construction follows the pipeline outlined below:

Name and Company Generation. We compiled an initial list of over 10,000 names from the *Behind the Name* database[2], removed duplicates, and ensured gender balance, yielding a final set of 4,000 unique names. In parallel, we employed GPT-4 to generate 200 fictional companies, each described by geographic location, primary industry (e.g., education, finance), and a concise description. These companies cover a broad spectrum of industries to ensure generalizability for subsequent analytical tasks.

Knowledge Graph Construction. We randomly selected 200 names as *main characters*. Each main character is associated with two social sub-networks: a family network and a workplace network, collectively comprising approximately 20 entities. Family networks contain 7–10 individuals characterized by consistent linguistic backgrounds, realistic demographic distributions (e.g., gender ratios of 0.4–0.6), and logically assigned family roles with appropriate age relationships. Each main character was assigned to one of the 200 companies, with organizational roles (e.g., leadership, senior, junior) determined by age, ensuring realistic workplace hierarchies. Additional coworkers were sampled from the broader name set to introduce geographic diversity.

Entity Attributes and Multimodal Representation. Structured attributes, including role, age, birthplace, and appearance, were generated for each entity using GPT-4. We validated these attributes for internal consistency, for example, ensuring logical organizational hierarchies within companies. Culturally appropriate surnames were subsequently assigned: familial relationships dictated surname inheritance (e.g., father's surname), whereas workplace entities received surnames aligned with geographic origin.

To enrich entities with visual representations, we utilized Stable Diffusion 3 [Esser *et al.*, 2024] to generate four candidate images per entity based on textual descriptions. These images were ranked using CLIPScore [Hessel *et al.*, 2021], and the highest-ranked image was selected as the final visual representation.

QA Generation. To enhance the diversity and realism of our dataset, we design five question types reflecting varied reasoning requirements: *ask_relation*, *ask_entity*, *ask_entity_multi_hop*, *ask_entity_multi_choice*, and *ask_entity_multi_hop_multi_choice*. For each relation in the knowledge graph, we generate both single-hop and multi-hop questions, ensuring unique answers. This setup enriches the dataset with varied reasoning paths, allowing more faithful evaluation of unlearning performance in realistic QA scenarios.

[2]https://www.behindthename.com/

4.2 Split Dataset

Forget and Retain Sets. To evaluate unlearning effectiveness, we construct explicit *forget* and *retain* sets. For each main character, we randomly select one fact from their family or workplace network as the *direct forget set*. Algorithm 1 is then applied to identify additional supporting facts along reasoning paths that could indirectly infer this target fact, forming the *deep forget set*. Removing these ensures the goal of deep unlearning.

To comprehensively assess retention capabilities, we design the retain set across three progressively broader contexts. First, the *neighbor retain set* comprises remaining facts within the same local network. Next, the *global retain set* includes facts sampled from unrelated character networks, providing broader domain coverage. Finally, we introduce a *model utility set*, based on the general-knowledge benchmarks MMLU [Hendrycks *et al.*, 2021] and MMMU [Yue *et al.*, 2024], enabling the evaluation of broader model utility.

Training and Evaluation Sets. To mimic real-world applications—where models are typically trained on limited question types but evaluated under broader conditions—we adopt a selective training strategy. For each fact, only two QA types are included in the training set, limiting exposure to linguistic variations. The evaluation set, by contrast, includes all QA types, thus testing the model's ability to generalize to unseen formulations. This setup imposes a more rigorous criterion for both retention and forgetting.

Algorithm 1 Get Forget Set

Require: Target fact $f(A \xrightarrow{r} B)$, knowledge graph $\mathcal{G}$
Ensure: Forget fact set $\mathcal{F}$ or $\emptyset$

function GETFORGETSET($f, \mathcal{G}$)
  $\mathcal{P} \leftarrow$ GETREASONINGPATHS(f)
  if $\mathcal{P} = \emptyset$ **then**
    return $\emptyset$
  end if
  $\mathcal{F} \leftarrow \emptyset$
  for $p \in \mathcal{P}$ **do**
    $\mathcal{F} \leftarrow \mathcal{F} \cup \{p[0], p[0]^{-1}\}$
  end for
  return $\mathcal{F}$ ▷ Return the forget fact set
end function

4.3 Statistics

The dataset comprises more than 30k relations, 166k QA pairs, and 4,800 entities. The detailed statistics are shown in Appendix D.

Benchmark	TOFU [Maini *et al.*, 2024]	CLEAR [Dontsov *et al.*, 2024]	RWKU [Jin *et al.*, 2024]	KLUE [Yang *et al.*, 2025]	**MMDU (ours)**
Knowledge Source	Synthetic	Synthetic	Real-world	Real-world	Synthetic
Deep Unlearning	✗	✗	✗	✓	✓
Multi-modal Data	✗	✓	✗	✗	✓
Multi-hop QA	✗	✗	✗	✓	✓
Model Utility	✓	✓	✗	✗	✓
Multi-fact Unlearning	✗	✗	✗	✗	✓
Unlearning Targets	200	200	200	200	1,000
Forget Probes	4,000	4,000	13,131	8,377	38,356

Table 1: A comparison between current unlearning benchmarks and MMDU-Bench.

4.4 Evaluation Metrics

Since we cannot directly verify whether a fact has been completely forgotten, we evaluate forgetting based on the model's inability to answer queries related to that fact or its reasoning path. Specifically, if a model can answer any QA query derived from the forget set, we consider the fact to be retained. This stringent criterion reflects realistic adversarial scenarios in which users might probe a model using varied questioning strategies.

Based on this framework, we define the following evaluation metrics:

Forget Quality. To assess how well target knowledge is forgotten, we define two forgetting metrics as follows:

Direct Forget Quality (DirectFQ): Measures whether the model forgets the specific target fact. Let $\mathbb{I}$ be an indicator function that returns 1 if the model answers correctly and 0 otherwise. Then,

$$\text{DirectFQ} = 1 - \frac{1}{|\mathcal{F}_d|} \sum_{f \in \mathcal{F}_d} \frac{\sum_{q \in Q_t(f)} \mathbb{I}(q)}{|Q_t(f)|}$$

where $\mathcal{F}_d$ is the *direct forget set* and $Q_t(f)$ denotes the set of training questions associated with fact f.

Deep Forget Quality (DeepFQ): Assesses whether the model forgets inferred facts along the reasoning path:

$$\text{DeepFQ} = 1 - \frac{1}{|\mathcal{F}_i|} \sum_{f \in \mathcal{F}_i} \frac{\sum_{q \in Q_e(f)} \mathbb{I}(\exists q \in Q_e)}{|Q_e(f)|}$$

where $\mathcal{F}_i$ is the *deep forget set* and $Q_e(f)$ denotes the set of evaluation questions associated with fact f.

Retain Quality. To assess whether unrelated knowledge is retained, we define retention metrics at three levels:

Neighbor Retain Quality (NRQ): Measures retention of neighboring facts in the same network that are not in the forget set:

$$\text{NRQ} = \frac{1}{|\mathcal{F}_n|} \sum_{f \in \mathcal{F}_n} \mathbb{I}\left(\exists q \in Q_e(f)\right)$$

where $\mathcal{F}_n$ is the *neighbor retain set*.

Global Retain Quality (GRQ): Measures retention of facts from unrelated networks:

$$\text{GRQ} = \frac{1}{|\mathcal{F}_g|} \sum_{f \in \mathcal{F}_g} \mathbb{I}\left(\exists q \in Q_e(f)\right)$$

where $\mathcal{F}_g$ is the *global retain set*.

Model Utility: Evaluates the model's general knowledge after unlearning. We measure this via MMLU (text) and MMMU (multi-modal).

Forgetting-Retention Trade-off Score (FRTS). We introduce a normalized metric to evaluate the balance between forgetting effectiveness and knowledge retention. It is defined as:

$$\text{FRTS} = \frac{1}{2}\Bigg(\underbrace{\frac{\text{DirectFQ} + \text{DeepFQ}}{2}}_{\text{ForgetScore}} \cdot \underbrace{\frac{1}{2}\left(\frac{\text{NRQ}_{\text{unlearn}}}{\text{NRQ}_{\text{base}}} + \frac{\text{GRQ}_{\text{unlearn}}}{\text{GRQ}_{\text{base}}}\right)}_{\text{RetainRatio}} + \underbrace{\frac{\text{Utility}_{\text{unlearn}}}{\text{Utility}_{\text{base}}}}_{\text{UtilityRatio}}\Bigg)$$

All components are scaled to the range $[0, 1]$, and thus FRTS also lies in $[0, 1]$. A higher score indicates more effective forgetting with minimal loss of unrelated or general-purpose knowledge.

5 Experiments

5.1 Setup

Models. We adopt LLaVA-1.5_{7B} [Liu *et al.*, 2024a] and Qwen2.5-VL_{3B} [Bai *et al.*, 2025], two widely used open-source LVLMs with strong multimodal capabilities, as base models for evaluating unlearning methods.

Data. Due to computational constraints, we randomly sample 50 *main characters* from the full dataset and reconstruct a corresponding subset using the same pipeline described in the dataset section. To help the model retain general knowledge during unlearning, we augment the training data with data from the ScienceQA [Lu *et al.*, 2022] dataset and MMLU [Hendrycks *et al.*, 2021] training set.

Unlearning Baselines We evaluate five mainstream unlearning methods: Task Vector (TV)[Ilharco *et al.*, 2023], Gradient Ascent (GA)[Jang *et al.*, 2023], Direct Preference Optimization (DPO)[Rafailov *et al.*, 2023], Negative Preference Optimization (NPO)[Zhang *et al.*, 2024], and Who's Harry Potter (WHP)[Eldan and Russinovich, 2023]. We consider two unlearning settings: *single-fact*, where the model is trained to forget a single target fact, and *multi-fact*, where five related facts are removed simultaneously. These settings allow us to evaluate both fine-grained unlearning precision and the robustness of each method when handling multiple interconnected facts. Each method is trained and evaluated

Model	Method	Forget Quality		Retain Quality			FRTS↑
		DirectFQ ↑	*DeepFG*↑	*NRQ*↑	*GRQ* ↑	*Utility*↑	
LLaVA-1.5$_{7B}$	NONE	7.33/9.33	0.00/0.00	**94.43/91.56**	**94.50/92.50**	**50.50/32.25**	N/A
	DPO$_{text}$	64.00/50.67	2.14/1.32	89.93/91.11	91.00/90.50	48.50/30.12	0.6186
	DPO$_{multi\text{-}modal}$	52.00/98.67	3.12/4.45	92.21/88.20	90.50/89.00	48.50/29.75	0.6637
	GA$_{text}$	**100.00/100.00**	**34.11/31.87**	74.45/80.17	90.50/86.50	45.38/29.75	0.7495
	GA$_{multi\text{-}modal}$	**100.00/100.00**	19.55/24.20	82.55/81.21	91.50/83.00	47.12/31.62	**0.7521**
	NPO$_{text}$	45.33/77.33	4.55/10.83	86.45/88.41	93.00/87.00	48.25/28.50	0.6279
	NPO$_{multi\text{-}modal}$	62.67/96.67	6.70/13.39	92.09/86.88	92.00/90.00	49.50/30.63	0.7012
	TV$_{text}$	72.00/72.00	2.50/3.12	94.28/87.48	93.00/87.00	48.75/31.25	0.6648
	TV$_{multi\text{-}modal}$	62.67/62.67	3.12/2.50	90.80/89.70	93.00/91.00	47.62/29.62	0.6267
	WHP$_{text}$	**100.00/100.00**	8.67/5.71	84.76/81.28	90.00/88.00	47.50/30.50	0.7185
	WHP$_{multi\text{-}modal}$	78.00/**100.00**	7.41/4.29	92.75/87.78	90.50/91.00	47.50/28.12	0.6871
Qwen2.5-VL$_{3B}$	NONE	0.00/0.00	0.00/0.00	**100.00/99.73**	**100.00/99.50**	57.50/42.50	N/A
	DPO$_{text}$	**100.00/100.00**	**13.75**/8.25	95.43/96.74	**100.00**/99.00	56.75/42.25	**0.7669**
	DPO$_{multi\text{-}modal}$	75.00/75.00	4.35/4.55	**100.00**/98.91	**100.00**/99.00	57.50/42.75	0.6992
	GA$_{text}$	77.33/72.00	3.75/5.63	100.00/99.46	98.50/98.00	56.00/41.75	0.6855
	GA$_{multi\text{-}modal}$	26.67/54.67	2.50/3.13	99.07/98.53	99.00/98.50	56.25/42.25	0.6001
	NPO$_{text}$	75.62/50.67	6.75/3.33	**100.00**/99.07	99.25/99.00	56.00/40.50	0.6521
	NPO$_{multi\text{-}modal}$	44.08/78.90	3.33/**11.82**	99.07/99.07	99.50/99.25	56.25/41.75	0.6617
	TV$_{text}$	45.00/43.05	0.00/0.00	**100.00**/98.53	**100.00**/99.00	56.25/42.75	0.6046
	TV$_{multi\text{-}modal}$	25.00/35.00	0.00/0.00	99.07/99.07	**100.00**/99.00	56.50/**43.50**	0.5746
	WHP$_{text}$	**100.00**/78.00	4.38/3.12	98.15/98.53	**100.00**/99.00	**58.00**/42.75	0.7336
	WHP$_{multi\text{-}modal}$	77.33/**100.00**	2.42/5.56	**100.00**/98.53	**100.00**/99.00	55.25/39.75	0.7057

Table 2: Results on *single-fact* unlearning setting. In the performance formulated as "a/b", "a" denotes the text-only performance and "b" is the multi-modal performance. Results on *multi-fact* unlearning setting can be found in Table 3.

separately on both textual and multi-modal QA datasets. Appendix E includes implementation and hyperparameter details.

5.2 Results

The results for *single-fact* unlearning are presented in Table 2, while those for *multi-fact* unlearning are shown in Table 3. We highlight three key observations from the experimental outcomes.

Current Methods Struggle with Deep Forgetting. In the *single-fact* unlearning setting, several methods exhibit promising results in removing target facts while preserving unrelated knowledge. For example, GA$_{multi_modal}$ on LLaVA-1.5$_{7B}$ achieves 100% DirectFQ and the highest FRTS of 0.7521, suggesting it can effectively and precisely forget the intended fact without broadly degrading model behavior. However, even the strongest method in this setting achieves only 34.11% and 31.87% DeepFQ on LLaVA-1.5$_{7B}$ and Qwen2.5-VL$_{3B}$, respectively, revealing that current methods remain largely ineffective at forgetting inferred knowledge along reasoning paths.

In the more realistic *multi-fact* unlearning setting, overall forgetting improves. For instances, GA$_{text}$ achieves a DeepFQ of 59.20% on text-only QA and 47.74% on multi-modal QA. However, this comes at the cost of reduced NRQ, GRQ, and Utility, highlighting a stronger trade-off. Furthermore, for models with stronger memorization capacity, such as Qwen2.5-VL$_{3B}$, deep unlearning becomes even more difficult. While DPO-text achieves 83.16% DeepFQ on text-only QA and 78.47% on multi-modal QA in the *multi-fact* setting, most other methods yield much lower scores even under the simpler *single-fact* setting. These findings confirm that current unlearning techniques are primarily surface-level, lacking robustness when facing multi-hop, modality-linked reasoning or stronger models, underscoring the urgent need for more principled and scalable deep unlearning approaches.

Large Performance Gaps Between Modalities. We find that the modality of data used during unlearning has a significant impact on both forgetting effectiveness and the side effects on retained knowledge. First, using different modality data for unlearning leads to varying levels of forgetting. In several cases, training on text-only data results in more thorough removal of facts, especially when evaluated on text-based queries. For example, on the *single-fact* dataset with LLaVA-1.5$_{7B}$, GA trained on text-only data shows stronger forgetting compared to when trained on multi-modal data.

Second, unlearning tends to affect the modality-aligned knowledge more directly. When training is performed on text-only data, we observe higher forgetting scores on textual evaluations (e.g., higher DirectFQ), but this also results in greater loss of nearby knowledge within the text modality. For instance, in LLaVA-1.5$_{7B}$, GA trained on text-only data yields lower NRQ (74.45% compared to 80.17%), showing that aggressively forgetting in one modality can come at the cost of retaining useful knowledge in the same modality. Conversely, training with multi-modal data tends to produce more moderate forgetting effects, while better preserving broader context. These results highlight that the choice of training data in unlearning must be aligned with the intended deployment scenario and that modality-specific for-

Model	Method	Forget Quality		Retain Quality			FRTS↑
		DirectFQ↑	*DeepFG*↑	*NRQ*↑	*GRQ*↑	*Utility*↑	
LLaVA-1.5$_{7B}$	NONE	32.50/34.44	8.85/7.81	**94.57/91.72**	**95.50/92.50**	**50.50/32.25**	N/A
	DPO$_{text}$	89.44/78.61	23.78/19.62	90.08/89.66	93.50/90.50	49.00/30.75	0.7387
	DPO$_{multi-modal}$	75.83/83.61	23.26/23.78	94.22/86.98	94.00/89.50	49.00/30.75	0.7334
	GA$_{text}$	**100.00/100.00**	**59.20**/47.74	61.51/62.65	86.00/86.50	45.38/30.87	0.7646
	GA$_{multi-modal}$	**100.00**/97.22	45.31/**67.53**	74.03/51.17	91.50/83.00	47.12/28.62	**0.7678**
	NPO$_{text}$	83.89/81.11	22.40/20.83	88.54/85.12	92.50/90.00	48.62/29.88	0.7220
	NPO$_{multi-modal}$	75.83/81.11	26.56/28.82	92.26/86.68	93.50/91.00	48.62/30.75	0.7373
	TV$_{text}$	75.83/73.06	19.44/20.66	91.07/87.95	94.50/92.00	48.75/30.63	0.7103
	TV$_{multi-modal}$	75.83/81.39	19.44/22.05	90.10/85.22	94.50/90.50	49.00/31.00	0.7225
	WHP$_{text}$	97.50/94.72	32.12/31.08	79.71/81.13	90.00/91.00	47.50/31.12	0.7666
	WHP$_{multi-modal}$	84.17/94.72	18.06/26.56	89.31/82.49	91.50/88.00	49.62/29.62	0.7410
Qwen2.5-VL$_{3B}$	NONE	10.28/2.50	0.00/0.00	**100.00/99.06**	**100.00/99.50**	57.50/42.50	N/A
	DPO$_{text}$	**100.00/100.00**	**83.16/78.47**	44.05/47.52	81.00/82.50	52.50/41.25	**0.7580**
	DPO$_{multi-modal}$	74.17/79.17	23.96/22.40	83.95/82.04	99.00/98.50	56.88/**43.00**	0.7270
	GA$_{text}$	83.33/83.33	66.67/48.44	58.78/55.93	73.50/72.00	40.25/33.25	0.5974
	GA$_{multi-modal}$	89.44/91.94	31.77/42.88	74.48/64.19	95.50/84.50	43.50/35.00	0.6484
	NPO$_{text}$	65.00/49.44	3.12/4.69	97.42/95.96	99.50/98.50	55.75/41.37	0.6357
	NPO$_{multi-modal}$	38.89/41.39	2.14/3.03	98.96/97.57	99.75/**99.50**	56.50/42.88	0.6030
	TV$_{text}$	34.17/31.67	0.00/0.00	98.61/98.02	99.50/**99.50**	**57.62**/41.88	0.5792
	TV$_{multi-modal}$	23.33/28.89	0.00/0.00	98.44/98.02	99.50/99.00	**57.62**/42.88	0.5672
	WHP$_{text}$	94.72/81.39	19.97/11.63	91.02/90.77	99.50/99.00	55.38/41.75	0.7334
	WHP$_{multi-modal}$	76.39/97.22	14.06/17.88	98.26/96.03	99.25/99.00	55.00/42.25	0.7393

Table 3: Results on *multi-fact* unlearning.

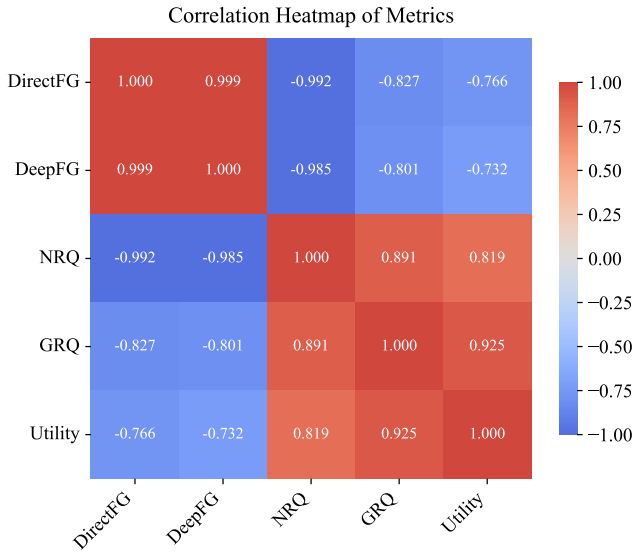

Figure 3: Metric correlation based on averaged baseline results.

getting dynamics should be carefully considered to balance effectiveness and stability.

Unlearning Disproportionately Affects Local Knowledge. Although these methods aim to remove specific facts, they often cause collateral damage to related knowledge. NRQ captures this effect, showing substantial impact on structurally related but unremoved facts. For example, in the *multi-fact* setting with LLaVA-1.5$_{7B}$, the GA-text method reduces NRQ to 61.51% on text-only QA and 62.65% on multi-modal QA, a stark drop from the baseline scores of 94.57% and 91.72%. This observation aligns with prior findings [Qin *et al.*, 2024], suggesting that interconnected knowledge often shares similar gradient directions, making isolated removal inherently difficult. In contrast, GRQ and Utility are only slightly reduced. GRQ drops from 94.00% to 86.00% on text-only QA and from 91.50% to 86.50% on multi-modal QA. Meanwhile, performance on general knowledge remains stable. This discrepancy highlights a core insight: unlearning impacts are spatially biased, with proximal knowledge bearing the brunt of interference. Figure 3 further supports this pattern, revealing clear stratified correlations among forget and retain metrics. The metric correlation of each baseline is shown in Appendix F.1. Conservative approaches like TV and NPO better preserve both NRQ and GRQ but exhibit reduced forgetting, underscoring the fundamental trade-off between unlearning efficacy and local retention. Together, these findings call for more fine-grained strategies that can disentangle target facts from their surrounding context.

6 Conclusion and Future Work

In this work, we introduce MMDU-Bench, the first benchmark for multi-modal deep unlearning, which evaluates models' ability to forget both explicit facts and cross-modal reasoning paths. Our findings show that deep unlearning remains a challenging task—existing methods often fail to remove entangled knowledge, especially when it spans modalities. Moreover, we observe a significant performance gap depending on the modality of the unlearning data, revealing new challenges in multi-modal unlearning. We hope MMDU-

Bench provides a foundation and useful insights for future research into more robust, modality-aware unlearning methods that better align with real-world safety and compliance demands.

A Proof of Valid Deep Unlearning Set

Proposition. Given a target fact $\tau = \langle e_a, r_{\text{target}}, e_b \rangle$, the forget set $\mathcal{F}$ returned by Algorithm 1 constitutes a valid *deep unlearning set*. That is, removing $\mathcal{F}$ from $\mathcal{G}$ yields a new graph $\mathcal{G}'$ such that:

$$\tau \notin \mathcal{G}'^{*},$$

where $\mathcal{G}'^{*}$ is the deductive closure of $\mathcal{G}'$ under the composition rules, and no valid reasoning path remains that implies r_{target} from e_a to e_b.

Proof. Let $\mathcal{P}$ denote the set of all valid acyclic reasoning paths p in $\mathcal{G}$ such that $\pi(p) = r_{\text{target}}$ and p starts at e_a and ends at e_b. By definition of the deductive closure, $\tau \in \mathcal{G}^{*}$ if and only if such a $p \in \mathcal{P}$ exists.

Algorithm 1 selects, for each $p \in \mathcal{P}$, the first edge $p[0]$ (along with its inverse, if applicable), and includes it in the forget set $\mathcal{F}$. After removing all such edges from $\mathcal{G}$, none of the paths in $\mathcal{P}$ remain traversable, and thus:

$$\forall p \in \mathcal{P}, \quad p \notin \mathcal{P}_{\mathcal{G}'}(e_a, e_b).$$

Hence, $\tau \notin \mathcal{G}'^{*}$ by construction.

To verify sufficiency, suppose for contradiction that $\tau \in \mathcal{G}'^{*}$. Then there must exist a new path $p' \in \mathcal{P}_{\mathcal{G}'}(e_a, e_b)$ such that $\pi(p') = r_{\text{target}}$. However, since all known paths in $\mathcal{P}$ were disrupted at their first edge, and composition rules $\mathcal{R}_{\text{comp}}$ are assumed deterministic and acyclic, any new path p' must either coincide with a removed prefix or require recombination of partial segments invalid under $\mathcal{R}_{\text{comp}}$. This contradicts the completeness of the edge removal over $\mathcal{P}$, so no such p' exists.

Furthermore, the selection of the first edge $p[0]$ in each path is motivated by its typically higher specificity to the target fact. For example, in the path

$$\texttt{Alice} \xrightarrow{\textit{born\_in}} \texttt{California} \xrightarrow{\textit{located\_in}} \texttt{USA},$$

removing $\langle \texttt{Alice}, \textit{born\_in}, \texttt{California} \rangle$ invalidates the inference "Alice was born in the USA" while leaving broader geographic knowledge intact. In contrast, removing $\langle \texttt{California}, \textit{located\_in}, \texttt{USA} \rangle$ would affect many unrelated facts.

Thus, although the algorithm does not guarantee minimality in terms of edge count, it yields a valid deep unlearning set that reduces disruption to unrelated facts.

B Additional Algorithms

B.1 Reasoning Path Extraction

Algorithm 2 Reasoning Path Extraction

Require: Target fact $f(A \xrightarrow{r} B)$, graph $\mathcal{G}$
Ensure: Set of valid reasoning paths $\mathcal{P}$ or $\emptyset$

function GETREASONINGPATHS($f, \mathcal{G}$)
  $s \leftarrow f.A$ ▷ source
  $t \leftarrow f.B$ ▷ target
  $\mathcal{P}_{cand} \leftarrow$ FINDPATHS$(s, t, \mathcal{G})$ ▷ DFS search
  $\mathcal{P}_{valid} \leftarrow \emptyset$
  for $p \in \mathcal{P}_{cand}$ **do**
    $e \leftarrow$ CONVERTEDGES(p) ▷ Apply Composition Rules
    if $e.rel = f.rel$ **then**
      $\mathcal{P}_{valid} \leftarrow \mathcal{P}_{valid} \cup \{p\}$
    end if
  end for
  return $\mathcal{P}_{valid}$ if non-empty else $\emptyset$
end function

B.2 Multi-Fact Forget Set Generation

Algorithm 3 Get Multi-Fact Forget Set

Require: Set of target facts $\mathcal{R} = \{f_1, f_2, ..., f_n\}$, knowledge graph $\mathcal{G}$
Ensure: Forget fact set $\mathcal{F}$

function GETMULTIFACTFORGETSET($\mathcal{R}, \mathcal{G}$)
  $\mathcal{F} \leftarrow \emptyset$
  for $f \in \mathcal{R}$ **do**
    $\mathcal{F}_f \leftarrow$ GETFORGETSET$(f, \mathcal{G})$
    if $\mathcal{F}_f \neq \emptyset$ **then**
      for $e \in \mathcal{F}_f$ **do**
        $\mathcal{F} \leftarrow \mathcal{F} \cup \{e\}$
        $\mathcal{G} \leftarrow \mathcal{G} \setminus \{e\}$ ▷ Remove edge e from the graph
      end for
    end if
  end for
  return $\mathcal{F}$
end function

C Training Details.

To ensure the models effectively memorize the target facts prior to unlearning, we conduct full-parameter fine-tuning. Training is performed for 4 epochs with a learning rate of 5×10^{-5}, an effective batch size of 256, a gradient norm clipping of 1, and a warm-up of 100 steps. All experiments are conducted on 4 A800-80G GPUs.

D Details of MMDU-Bench

D.1 Relation Types

MMDU-Bench defines a total of 24 distinct relation types, categorized into three domains: *Family Relation*, *Work Relation*, and *Geography Relation*, as illustrated in Figure 4.

Family Relation		
Father of	Mother of	Husband of
Wife of	Grandfather of	Grandmother of
Grandson of	Granddaughter of	Son of
Daughter of	Sibling of	Cousin of
Aunt of	Uncle of	Niece of
Nephew of		
Work Relation		
Colleague of	Manager of	Subordinate of
Employee of	Employer of	
Geography Relation		
Born in	Locates in	Contains

Table 4: Relation Types in MMDU-Bench.

D.2 QA Statistics

Split	AE.	AE MC.	AEMH.	AEMH MC.	AR.	AB.	AE Neg.	Total
Train	3,742	16,808	3,897	21,913	9,256	577	4,200	56,193
Eval	15,688	56,822	12,030	51,207	29,806	577	0	166,130

Table 5: QA type distribution across Train and Eval sets. **AE.**: ask_entity, **AE MC.**: ask_entity_multi_choice, **AEMH.**: ask_entity_multi_hop, **AEMH MC.**: ask_entity_multi_hop_multi_choice, **AR.**: ask_relation, **AB.**: ask_birthplace, **AE Neg.**: ask_entity_negative.

D.3 Character Country Distribution

The distribution of character nationalities and regions is illustrated in Figure 4.

E Implementation of Unlearning Baselines

E.1 Method Descriptions

We use the following unlearning methods as our baseline:

Task Vector (TV): Task Vector [Ilharco *et al.*, 2023] represents the difference between a fine-tuned model and its base model in parameter space, enabling modular knowledge transfer and modification. Mathematically, given a base model θ_{base} and a fine-tuned model θ_{task}, the task vector is defined as $\tau_{\text{task}} = \theta_{\text{task}} - \theta_{\text{base}}$. For unlearning, we can negate the task vector and add it to the model's parameters as $\theta_{\text{unlearned}} = \theta_{\text{learned}} - \lambda\tau_{\text{task}}$.

Gradient Ascend (GA): Gradient ascent [Jang *et al.*, 2023] maximizes the loss by updating model parameters in the direction of the gradient $\theta' = \theta + \eta\nabla_\theta\mathcal{L}(\theta)$. Although it can effectively erase knowledge from the model, it may cause **catastrophic forgetting**, degrading performance on unrelated tasks.

Direct Preference Optimization (DPO): DPO [Rafailov *et al.*, 2023] is a preference-based fine-tuning method that improves model alignment by directly optimizing the difference in likelihood between preferred and dispreferred outputs, without relying on reinforcement learning. The loss function is defined as:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}\left[\log\sigma\left(\beta\left(\log\frac{\pi_\theta(z_+)}{\pi_{\text{ref}}(z_+)} - \log\frac{\pi_\theta(z_-)}{\pi_{\text{ref}}(z_-)}\right)\right)\right]$$

where z_+ and z_- denote the preferred and dispreferred responses, and $\pi_{\text{ref}}(z)$ is the reference model's probability. The hyperparameter β controls the sharpness of the preference margin. In our unlearning setup, we treat the target fact as the dispreferred response and use an "I don't know" (IDK) style response as the preferred output. A full list of IDK templates is provided in Table 6.

Negative Preference Optimization (NPO): NPO [Zhang *et al.*, 2024] extends DPO by reinforcing negative preference signals to facilitate unlearning. The optimization objective is defined as:

$$\mathcal{L}_{\text{NPO}} = -\frac{2}{\beta}\log\sigma\left(-\beta\log\frac{\pi_\theta(z)}{\pi_{\text{ref}}(z)}\right)$$

where $\pi_\theta(z)$ is the model's probability of generating a target response z, and $\pi_{\text{ref}}(z)$ is a reference model's probability. By pushing the probability of certain responses below their reference likelihood, NPO forces the model to forget specific knowledge while preserving general capabilities.

Who's Harry Potter (WHP) [Eldan and Russinovich, 2023] is an unlearning method that mitigates knowledge traces by aligning the model's logits with those of a baseline model. Specifically, it computes a generic logit representation as follows:

$$v_{\text{generic}} = v_{\text{baseline}} - \alpha\,\text{ReLU}(v_{\text{reinforced}} - v_{\text{baseline}})$$

where $v_{\text{reinforced}}$ denotes the logits after fine-tuning, and v_{baseline} corresponds to the original model's logits. During fine-tuning, WHP replaces answer-critical tokens with generic alternatives, encouraging the model to produce non-informative outputs while preserving general capabilities. In our implementation, we approximate this effect by randomly replacing answers with alternative values.

E.2 Hyperparameter Settings

We set the effective batch size to 16 for all baseline methods. Below, we report the hyperparameter configurations used for each unlearning method under both the `single_fact` and `multi_fact` unlearning settings, and across two vision-language models: LLaVA-1.5$_{7B}$ and Qwen2.5-VL$_{3B}$. Table 7 summarizes the learning rate and number of training epochs for all methods and settings. This unified table facilitates direct comparison between models and unlearning configurations. Each row corresponds to an unlearning method, while the columns report the learning rate and epoch values under both single-fact and multi-fact settings for each model.

F Additional Results

F.1 Metric Correlation Across Baselines

Figures 5–9 present Pearson correlation heatmaps between evaluation metrics across unlearning methods under the *multi-fact* unlearning setting using the Qwen2.5-VL$_{3B}$ model. We observe consistently strong correlations between DirectFQ and DeepFQ, indicating that effective removal of

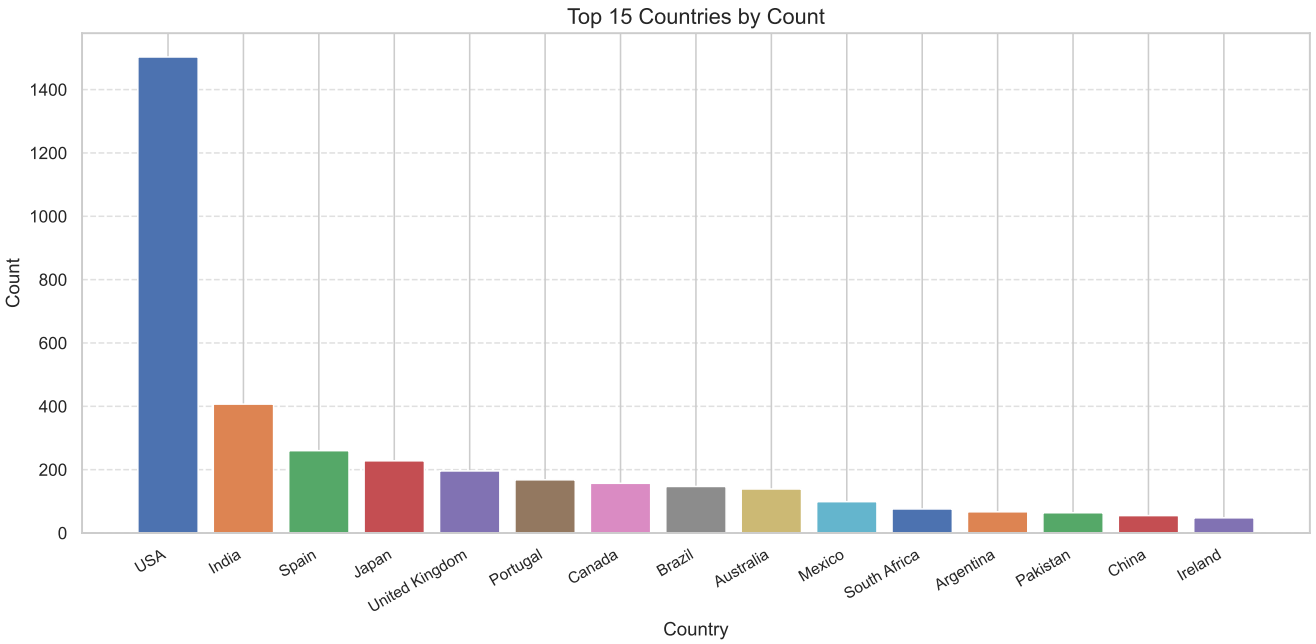

Figure 4: Character nationality and regional distribution

IDK Responses	
I'm not sure about that.	Sorry, I really don't know.
I have no idea at all.	I'm not certain of that.
That's beyond my knowledge.	I can't provide an answer.
I don't have that info.	I'm not familiar with it.
That's not something I know.	I'm not informed on this.
I have no clue about it.	I don't have enough information about that.

Table 6: List of IDK-style responses used as preferred outputs in DPO training.

target facts generally accompanies removal of related supporting facts along reasoning paths. The retention metrics—Neighbor Retain Quality (NRQ), Global Retain Quality (GRQ), and Model Utility—form a distinct three-level hierarchy, reflecting progressively broader categories of retained knowledge.

Notably, DPO and GA exhibit higher DirectFQ-DeepFQ correlations and weaker correlations between forgetting and retention metrics compared to more conservative methods (e.g., NPO and TV). This indicates that DPO and GA achieve more targeted unlearning by effectively eliminating both the specified facts and their associated reasoning chains, without substantially compromising the retention of unrelated information.

F.2 Forget&Retain Trend over Epoch

Figures 10–14 illustrate the trends of each evaluation metric across five unlearning baselines as the number of training epochs increases.

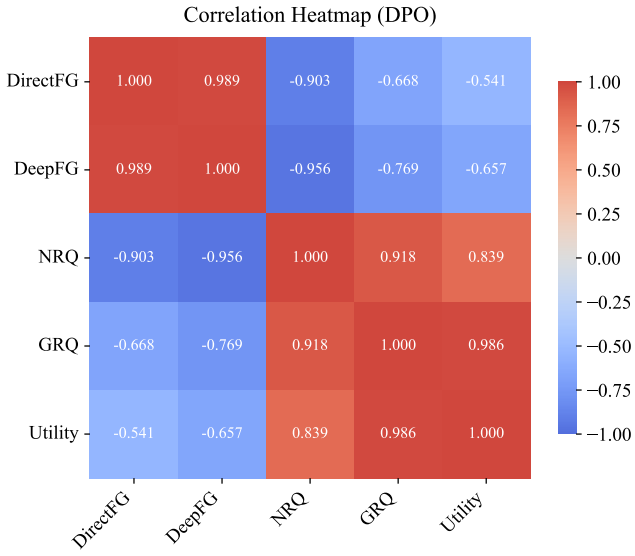

Figure 5: Correlation Heatmap of DPO

Method	LLaVA-1.5$_{7B}$ (Single)		Qwen2.5-VL$_{3B}$ (Single)		LLaVA-1.5$_{7B}$ (Multi)		Qwen2.5-VL$_{3B}$ (Multi)	
	LR	Epochs	LR	Epochs	LR	Epochs	LR	Epochs
NPO	5×10^{-6}	20	1×10^{-5}	20	5×10^{-6}	20	1×10^{-5}	20
GA	1×10^{-5}	30	2×10^{-5}	30	1×10^{-5}	30	2×10^{-5}	30
TV	1×10^{-5}	30	1×10^{-5}	30	1×10^{-5}	30	1×10^{-5}	30
DPO	5×10^{-6}	25	1×10^{-5}	20	5×10^{-6}	25	1×10^{-5}	30
WHP	5×10^{-6}	25	1×10^{-5}	20	5×10^{-6}	25	1×10^{-5}	20

Table 7: Unified hyperparameter settings (learning rate and number of epochs) for all unlearning methods under both single-fact and multi-fact scenarios across LLaVA-1.5$_{7B}$ and Qwen2.5-VL$_{3B}$.

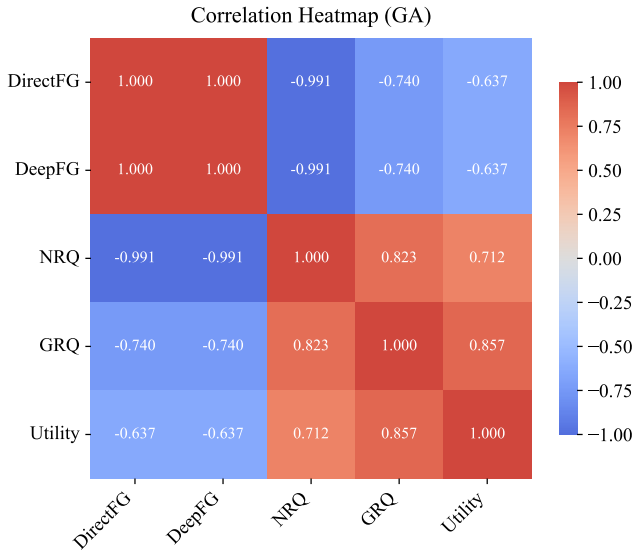

Figure 6: Correlation Heatmap of GA

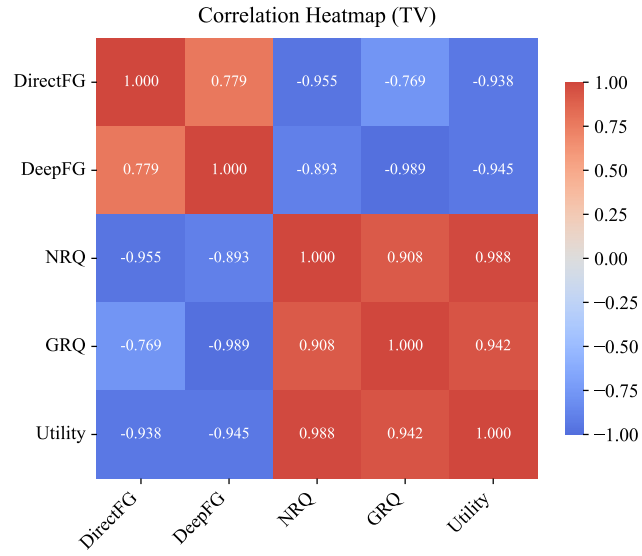

Figure 8: Correlation Heatmap of TV

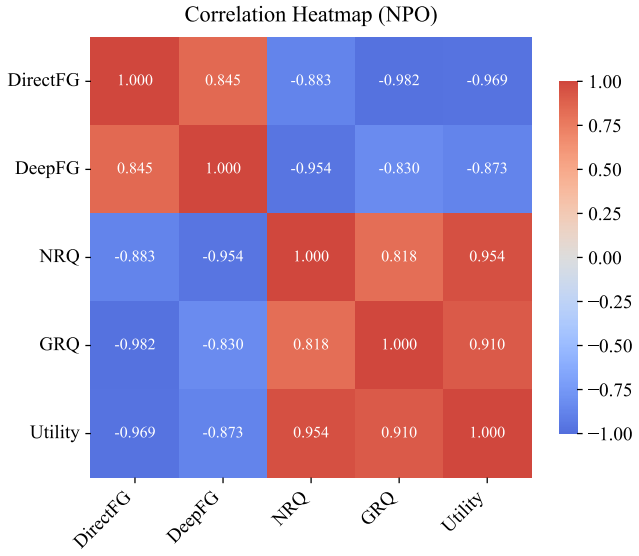

Figure 7: Correlation Heatmap of NPO

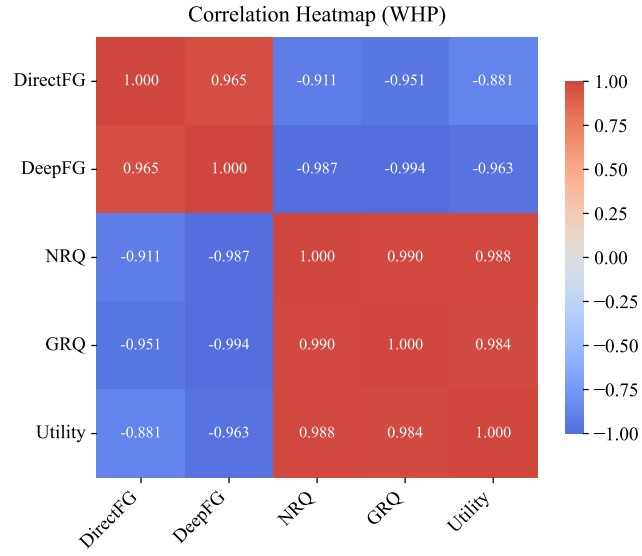

Figure 9: Correlation Heatmap of WHP

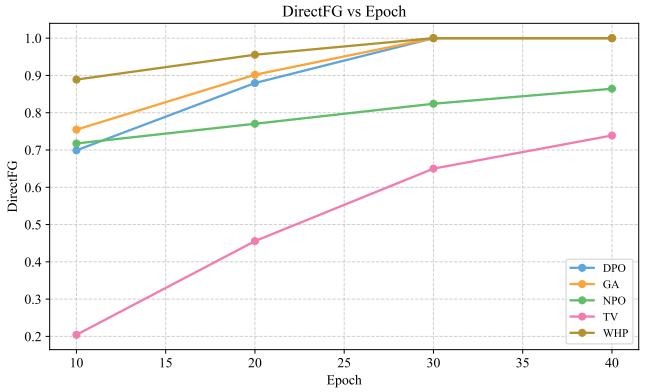

Figure 10: Direct Forget Quality Trend

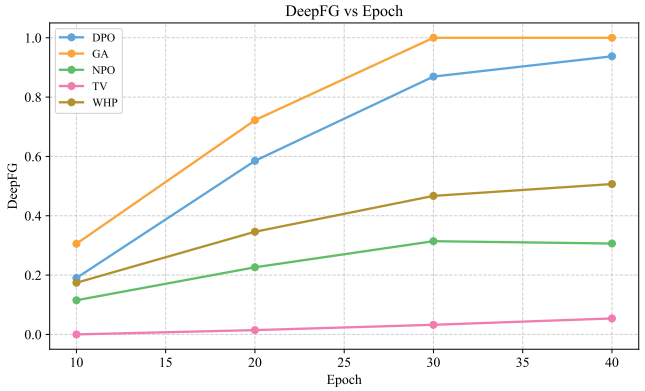

Figure 11: Deep Forget Quality Trend

G Prompt Used in MMDU-Bench Construction

Figure 15 and Figure 16 present the prompts provided to GPT-4 for assigning attributes to entities within the family and work relation networks. Figure 17 and Figure 18 show the prompts used to guide GPT-4 in generating character appearance and company logo descriptions.

Ethical Statement

There are no ethical issues.

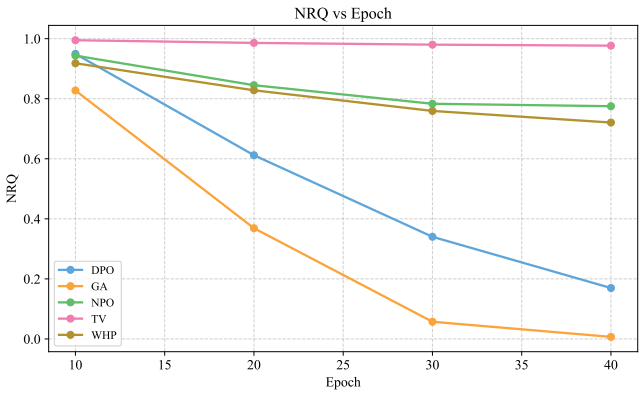

Figure 12: Neighbor Retain Quality Trend

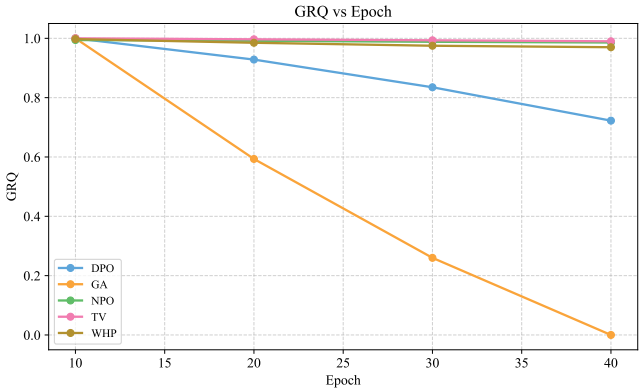

Figure 13: Global Retain Quality Trend

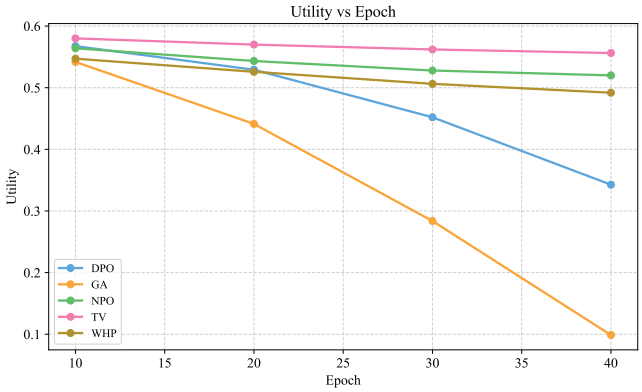

Figure 14: Model Utility Trend

Prompt

```
You are tasked with constructing a virtual family relationship knowledge graph based on
    the provided JSON data. Follow these instructions:
1. Identify the main character as specified in the input and build family roles for each
    individual in relation to this main character.
2. For the main character and each family member, add the following details:
   - Age: Ensure it is appropriate for their role in the family.
   - Place of Birth: Provide a detailed location in the format City, State/Province,
      Country.
3. Role Options: [grandfather,grandmother,father,mother,uncle,aunt,husband,wife,son,
    daughter,sibling,cousin,nephew,niece,granddaughter,grandson]
Output the enhanced family data, including information about the main character, in JSON
    format.
Input Example:
{
  "main_character": {
    "name": "Alice",
    "gender": "Female",
    "role": "main character",
    "age": 30
  }
  "family_members": [
    {"name": "John", "gender": "Male"},
    {"name": "Mary", "gender": "Female"},
    {"name": "Emily", "gender": "Female"}
  ]
}
Output Example:
{
  "main_character": {
    "name": "Alice",
    "role": "main character",
    "gender": "Female",
    "age": 30,
    "birthplace": "Seattle, Washington, USA",
  },
  "family_members": [
    {
      "name": "John",
      "gender": "Male",
      "role": "father",
      "age": 55,
      "birthplace": "Springfield, Illinois, USA",
    },
    {
      "name": "Mary",
      "gender": "Female",
      "role": "mother",
      "age": 52,
      "birthplace": "Denver, Colorado, USA",
    },
    {
      "name": "Emily",
      "gender": "Female",
      "role": "sibling",
      "age": 18,
      "birthplace": "Seattle, Washington, USA",
    }
  ]
}
Notes:
- Include the main character in the output with the role explicitly labeled as "main
    character".
- Maintain logical consistency in family roles, ages, and relationships.
- The family relationships need to align with normal family structures.
- All roles need to be within the range of Role Options.
Input:
$input_json
```

Figure 15: Prompt used for generating family relations.

Prompt

```
I am constructing a knowledge graph of virtual company relationships. The structure should
     begin with a main_character that is at the top of the hierarchy. The main character
    will have three fields: manager, colleague, and subordinate, which represent the
    respective relationships in the company.
For each person under the manager, colleague, and subordinate fields:
manager: Each manager will have a field for age, birthplace and a manager field to
    indicate their superior. If a person has no manager, the manager field should be 'null
    '.
colleague: Colleagues will have age, birthplace, and a colleague field to represent other
    colleagues at the same level. If a person has no colleagues, the colleague field should
     be 'null'.
subordinate: Subordinates will have age, birthplace, and a subordinate field to represent
    the subordinates under them. If a person has no subordinates, the subordinate field
    should be 'null'.
The birthplace should provide a detailed location in the format City, State/Province,
    Country.
The hierarchy of roles in the company is categorized from highest to lowest as ['
    leadership', 'management', 'senior', 'mid-level', 'junior', 'entry-level'].
Each person should be assigned a role that corresponds to their position in the hierarchy,
     along with age that fit their level.
The hierarchical relationship between the main_character and other persons in the company
    must STRICTLY align with the main_character's assigned level.
Example Input (JSON format):
{
   "main_character": {
      "name": "Alice",
      "age": 45
      "level": "leadership"
   }
   "company_members": [
      {"name": "John", "gender": "Male"},
      {"name": "Smith", "gender": "Male"},
      {"name": "Bob", "gender": "Male"}
   ]
}
Example Output:
{
   "Alice": {
      "manager": null
      "subordinate": {
         "John": {
               "age": 38,
               "subordinate": {
                  "Smith": {
                     "age": 28,
                     "subordinate": null,
                     "birthplace": "Springfield, Illinois, USA"
                   }
               },
               "birthplace": "Seattle, Washington, USA"
         }
      }
      "colleague": {
         "Bob": {
            "age": 36,
            "colleague": null,
            "birthplace": "Denver, Colorado, USA"
         }
      }
   }
}
Input:
$input_json
```

Figure 16: Prompt used for generating work relations.

Prompt

```
Input: A JSON data containing character relationships, including details such as name,
    role, gender, age, and birthplace.
Output: Using keywords that suitable for Stable Diffusion image generation to describe a
    detailed appearance for each character, covering:
- Gender and age range (e.g., young, adult, middle-aged, elderly)
- Ethnicity/Regional traits (based on birthplace, e.g., Caucasian, African American, East
    Asian)
- Facial features (e.g., sharp jawline, round face, high cheekbones)
- Hair style and color (e.g., short black hair, long wavy blonde hair)
- Eye color and shape (e.g., deep-set blue eyes, almond-shaped brown eyes)
- Skin tone (e.g., fair, tan, dark)
- Clothing style (if applicable, based on cultural background or profession)
- Accessories (if relevant, such as glasses, earrings, or hats)
- The appearance description should be realistic, diverse, and coherent, reflecting the
    person's background.
Output should format in JSON and only output the answer.
Example:
Input:
{
  "main_character": {
    "name": "Alice",
    "role": "main character",
    "gender": "Female",
    "age": 54,
    "birth_place": "Boston, Massachusetts, USA"
  },
  "family_members": [
    {
      "name": "Mike",
      "gender": "Male",
      "role": "husband",
      "age": 56,
      "birth_place": "New York City, New York, USA"
    },
    {
      "name": "Alan",
      "gender": "Male",
      "role": "son",
      "age": 30,
      "birth_place": "Boston, Massachusetts, USA"
    },
    {
      "name": "John",
      "gender": "Male",
      "role": "son",
      "age": 28,
      "birth_place": "Boston, Massachusetts, USA"
    }
  ]
}
Output:
{
  "Alice": "Female, middle-aged, Caucasian, round face, high cheekbones, soft jawline,
      short light brown hair, slightly wavy, green almond-shaped eyes, fair skin tone,
      casual modern attire, light colors, no accessories.",
  "Mike": "Male, middle-aged, Caucasian, square jawline, broad forehead, prominent nose,
      short dark brown hair, neatly combed, deep-set blue eyes, fair skin tone, casual
      button-up shirts and khaki pants, watch.",
  "Alan": "Male, young adult, Caucasian, sharp jawline, straight nose, youthful appearance,
       medium-length wavy blonde hair, hazel round-shaped eyes, fair skin tone, modern
      casual t-shirts and jeans, no accessories.",
  "John": "Male, young adult, Caucasian, soft jawline, narrow face, light stubble, short
      brown hair, slightly messy, brown almond-shaped eyes, fair skin tone, casual hoodies
      and jeans, no accessories."
}
Input:
$input_json
```

Figure 17: Prompt used for generating appearance for each character.

Prompt

```
Input: a company name and a brief description of the company
Output: keywords for stable diffusion to generate the company logo.

For example:
Input:
Name: GrapeWine Company
Description: A company that produces grape wines with great flavor.
Output: logo,Minimalist,A bunch of grapes and a wine glass

Input:
Name: $name
Description: $description
```

Figure 18: Prompt used to generate logo descriptions for each company.

References

[Agrawal *et al.*, 2019] Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956. IEEE, 2019.

[Arefeen *et al.*, 2024] Md. Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. Leancontext: Cost-efficient domain-specific question answering using llms. *Nat. Lang. Process. J.*, 7:100065, 2024.

[Auer *et al.*, 2007] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2007.

[Bai *et al.*, 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. *CoRR*, abs/2308.12966, 2023.

[Bai *et al.*, 2025] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[Chen *et al.*, 2024] Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Huajun Chen. Knowledge graphs meet multi-modal learning: A comprehensive survey. *CoRR*, abs/2402.05391, 2024.

[DeepSeek-AI *et al.*, 2025] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025.

[Dhingra *et al.*, 2022] Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-aware language models as temporal knowledge bases. *Trans. Assoc. Comput. Linguistics*, 10:257–273, 2022.

[Dontsov *et al.*, 2024] Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Y. Rogov, Ivan V. Oseledets, and Elena Tutubalina. CLEAR: character unlearning in textual and visual modalities. *CoRR*, abs/2410.18057, 2024.

[Eldan and Russinovich, 2023] Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *CoRR*, abs/2310.02238, 2023.

[Esser *et al.*, 2024] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[Gong *et al.*, 2025] Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In Toby Walsh, Julie Shah, and Zico Kolter, editors, *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 23951–23959. AAAI Press, 2025.

[Goyal *et al.*, 2019] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis.*, 127(4):398–414, 2019.

[Hendrycks *et al.*, 2021] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[Hessel *et al.*, 2021] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics, 2021.

[Ilharco *et al.*, 2023] Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[Jang *et al.*, 2023] 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 L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14389–14408. Association for Computational Linguistics, 2023.

[Jin *et al.*, 2024] Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. RWKU: benchmarking real-world knowledge unlearning for large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[Kamalloo *et al.*, 2023] Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5591–5606. Association for Computational Linguistics, 2023.

[Kazemzadeh *et al.*, 2014] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL, 2014.

[Kim *et al.*, 2023] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. Propile: Probing privacy leakage in large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023.

[Li *et al.*, 2024a] Mukai Li, Lei Li, Yuwei Yin, Masood Ahmed, Zhenguang Liu, and Qi Liu. Red teaming visual language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 3326–3342. Association for Computational Linguistics, 2024.

[Li *et al.*, 2024b] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih,

Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Kiran Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[Liu *et al.*, 2019] Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. MMKG: multi-modal knowledge graphs. In Pascal Hitzler, Miriam Fernández, Krzysztof Janowicz, Amrapali Zaveri, Alasdair J. G. Gray, Vanessa López, Armin Haller, and Karl Hammar, editors, *The Semantic Web - 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2-6, 2019, Proceedings*, volume 11503 of *Lecture Notes in Computer Science*, pages 459–474. Springer, 2019.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE, 2024.

[Liu *et al.*, 2024b] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1817–1829. Association for Computational Linguistics, 2024.

[Liu *et al.*, 2025] Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. Protecting privacy in multimodal large language models with mllmu-bench. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 4105–4135. Association for Computational Linguistics, 2025.

[Lu *et al.*, 2022] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[Ma *et al.*, 2024] Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, Yejin Choi, Muhao Chen, and Chaowei Xiao. Benchmarking vision language model unlearning via fictitious facial identity dataset. *CoRR*, abs/2411.03554, 2024.

[Maini *et al.*, 2024] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for llms. *CoRR*, abs/2401.06121, 2024.

[Mousavi *et al.*, 2024] Seyed Mahed Mousavi, Simone Alghisi, and Giuseppe Riccardi. Dyknow: Dynamically verifying time-sensitive factual knowledge in llms. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 8014–8029. Association for Computational Linguistics, 2024.

[OpenAI, 2023] OpenAI. Gpt-4v(ision) technical work and authors. https://cdn.openai.com/contributions/gpt-4v.pdf, 2023. Accessed: 2025-04-22.

[Qin *et al.*, 2024] Jiaxin Qin, Zixuan Zhang, Chi Han, Pengfei Yu, Manling Li, and Heng Ji. Why does new knowledge create messy ripple effects in llms? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 12602–12609. Association for Computational Linguistics, 2024.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[Shi *et al.*, 2024] Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: machine unlearning six-way evaluation for language models. *CoRR*, abs/2407.06460, 2024.

[Shostack, 2024] Adam Shostack. The boy who survived: Removing harry potter from an llm is harder than reported, 2024.

[Staab *et al.*, 2024] Robin Staab, Mark Vero, Mislav Balunovic, and Martin T. Vechev. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[Suchanek *et al.*, 2007] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In Carey L. Williamson, Mary Ellen Zurko,

Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 697–706. ACM, 2007.

[Toutanova *et al.*, 2015] Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. Representing text for joint embedding of text and knowledge bases. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1499–1509. The Association for Computational Linguistics, 2015.

[Wang *et al.*, 2024] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[Wu *et al.*, 2024] Ruihan Wu, Chhavi Yadav, Russ Salakhutdinov, and Kamalika Chaudhuri. Evaluating deep unlearning in large language models. *CoRR*, abs/2410.15153, 2024.

[Yang *et al.*, 2025] Nakyeong Yang, Minsung Kim, Seunghyun Yoon, Joongbo Shin, and Kyomin Jung. Faithun: Toward faithful forgetting in language models by investigating the interconnectedness of knowledge. *CoRR*, abs/2502.19207, 2025.

[Yao *et al.*, 2024] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[Yue *et al.*, 2024] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE, 2024.

[Zhang *et al.*, 2024] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *CoRR*, abs/2404.05868, 2024.

[Zhu *et al.*, 2024a] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[Zhu *et al.*, 2024b] Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. Multi-modal knowledge graph construction and application: A survey. *IEEE Trans. Knowl. Data Eng.*, 36(2):715–735, 2024.

[Zhu *et al.*, 2024c] Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. Can large language models understand context? In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 2004–2018. Association for Computational Linguistics, 2024.