MUREN-BENCH: EVALUATING LOCALITY IN REAL-WORLD MULTIMODAL UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning requires a trained model to remove designated knowledge while leaving the rest of its capability intact, a property known as locality. For multimodal large language models, real-world requests such as right-to-be-forgotten claims, copyright takedowns, and trademark disputes make this task concrete: the model must stop recognizing one specific real-world entity and answering questions about it, while preserving the knowledge surrounding that entity. Existing multimodal unlearning benchmarks fall short of this setting in two ways: they typically construct forget targets by injecting fictitious entities through fine-tuning, so the model lacks organic knowledge of the target and its real-world surroundings; and they measure locality with a single averaged retain score over entities far from the target, so failures to preserve closely related knowledge remain invisible. To address both problems, we introduce **MUREN-Bench** (**M**ultimodal **U**nlearning of **R**eal-world **E**ntity **N**eighborhoods), a benchmark built entirely from real-world entities across six categories: artifacts, biological entities, celebrities, IP characters, landmarks, and trademarks. The benchmark contains 90 groups, each pairing one forget target with a hierarchical retain set ordered by relatedness to the target: a distant anchor, an intermediate neighbor, and the closest neighbor. This structure turns locality from a single averaged score into a profile over proximity to the target. Every entity is probed through image-only, image-text, and text-only routes, yielding 360 entities and 57,240 probes in total. Experiments with seven unlearning methods on the LLaVA-1.5-7B and Qwen3-VL-8B backbones, compared against the original models, reveal a shared limitation: methods that suppress the target often fail to preserve nearby entities, and none achieves reliable locality within the target's neighborhood.

1 INTRODUCTION

Recent multimodal large language models (MLLMs) can identify people, landmarks, trademarks, fictional characters, artworks, artifacts, and biological entities from visual and textual evidence (Liu et al., 2023; Bai et al., 2025). This broad knowledge makes MLLMs useful, but it also allows them to retain private, copyrighted, outdated, or hazardous content long after training (Li et al., 2024; Shi et al., 2025; Liu et al., 2025a). Machine unlearning seeks to remove designated information without rebuilding the model from scratch (Cao & Yang, 2015; Bourtoule et al., 2021). For a deployed MLLM, however, removal is only half of the requirement: a valid update must forget the requested entity while preserving the surrounding knowledge that should remain available, a property known as *locality*. Locality is hardest to maintain precisely where it matters most: entities that share visual or semantic cues with the target are the ones most likely to be damaged by its removal.

Existing benchmarks provide important but incomplete views of this problem. TOFU, MLLMU-Bench, CLEAR, and PEBench create controllable forget targets by synthesizing authors, identities, or person and event concepts and injecting them before unlearning (Maini et al., 2024; Liu et al., 2025a; Dontsov et al., 2025; Xu et al., 2025). Although useful for controlled evaluation, this setting primarily tests the removal of recently implanted information rather than knowledge of real entities already encoded through pretraining or subsequent training. Other benchmarks evaluate pretrained hazardous knowledge, public-figure facts, attack robustness, or private information entangled with public image content (Li et al., 2024; Kawakami et al., 2025; Patil et al., 2025; Guang et al., 2026; Zhang et al., 2026). However, they typically measure locality with broad retain or utility sets drawn far from the target, making it difficult to determine how locality degrades as related entities approach

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Forgetting a target is not enough: its neighbors must remain recognizable.** (a) Six families of real-world entities already known to the base model. (b) Each target is paired with three retain entities of decreasing relatedness (here, by shared artist and subject; layout is schematic); held-out T1 and T2 test whether locality extends to unseen neighbors. (c) Each entity is probed through three routes, and locality is measured as level-wise retention relative to the original model.

the target. Across both settings, coverage also remains concentrated on people or a single application domain, leaving structured neighborhoods of real-world entities underexplored.

To address these limitations, we introduce **MUREN-Bench** (**M**ultimodal **U**nlearning of **R**eal-world **E**ntity **N**eighborhoods), illustrated in Figure 1. Rather than treating multimodal entity unlearning as an isolated deletion test, MUREN-Bench formulates it as a structured neighborhood problem in a realistic setting: forget targets and their neighbors are real-world entities already encoded in the base model's pretrained knowledge. The benchmark contains 90 groups spanning six entity families: artifacts, biological entities, celebrities, IP characters, landmarks, and trademarks. Each group pairs one forget target with a hierarchical retain set of three entities ordered by relatedness to the target: the closest neighbor (T1), an intermediate neighbor (T2), and a distant anchor (T3). Only T3 provides retain-training examples; T1 and T2 are held out from the group's unlearning training and serve to diagnose collateral damage as proximity to the target increases. This design turns locality from a single averaged score into a profile over proximity to the target. We probe all 360 entities through image-only, image-text, and text-only routes, yielding 57,240 probes in total. MUREN-Bench reports target accuracy before and after unlearning, retain accuracy at each level, and an overall score combining forgetting and locality.

We evaluate seven representative unlearning methods (Jang et al., 2023; Zhang et al., 2024; Huo et al., 2025; Ding et al., 2025) on LLaVA-1.5-7B (Liu et al., 2023) and Qwen3-VL-8B (Bai et al., 2025). The results reveal a common limitation: methods that preserve surrounding knowledge often fail to remove the target, whereas methods that forget more effectively can substantially degrade neighboring entities. None consistently balances target forgetting with locality across models, access routes, and entity families, indicating that neighborhood-level locality remains an open challenge for multimodal unlearning. Our contributions are summarized as follows:

- We introduce MUREN-Bench, to the best of our knowledge the first benchmark that evaluates locality in multimodal unlearning over structured neighborhoods of real-world entities already encoded in pretrained models. It covers six entity families and 90 groups, each pairing one forget target with three retain entities.
- We design a hierarchical retain set ordered by real-world relatedness, in which only the distant anchor is used for retain training and the two closer neighbors are held out. Combined with image-only, image-text, and text-only routes, this yields 57,240 probes that measure target leakage and level-wise locality.
- We benchmark seven representative unlearning methods on two MLLM backbones and find a persistent trade-off: existing methods either fail to forget the target sufficiently or cause substantial damage to neighboring entities, and none achieves reliable locality across models, access routes, and entity families.

2 Related Work

Machine Unlearning. Machine unlearning removes the influence of designated data without full retraining (Cao & Yang, 2015; Ginart et al., 2019). Exact approaches retain sufficient statistics or isolate training into retrainable shards, while certified removal bounds the effect of deletion for restricted model classes (Ginart et al., 2019; Bourtoule et al., 2021; Guo et al., 2019). For deep networks, approximate methods instead update parameters using gradients, curvature, or Fisher information (Golatkar et al., 2020; Foster et al., 2024). Language-model unlearning extends the goal from individual samples to knowledge elicited by many prompts: common objectives include gradient ascent, substitute labels, retain or KL regularization, negative preference optimization, second-order updates, and representation redirection (Jang et al., 2023; Eldan & Russinovich, 2023; Yao & Xu, 2024; Zhang et al., 2024; Jia et al., 2024; Li et al., 2024). Benchmarks such as TOFU, WMDP, and MUSE expose residual extraction, relearning, and utility loss across synthetic profiles, hazardous knowledge, books, and news; OpenUnlearning further unifies algorithms and metrics (Maini et al., 2024; Li et al., 2024; Shi et al., 2025; Dorna et al., 2026). Their language-only interfaces, however, cannot diagnose cross-modal leakage or localized damage to visually related entities.

Unlearning in Multimodal Large Language Models. MLLM methods either adapt language-model objectives or intervene in visual and language representations. MMUnlearner constrains updates with a geometry-aware saliency mask, MANU prunes influential neurons by modality, and SafeEraser uses prompt decoupling to reduce excessive refusal and over-forgetting (Huo et al., 2025; Liu et al., 2025b; Chen et al., 2025). Other methods act at different locations: SMFA masks a forgetting adapter with retain anchors, MIP-Editor edits influential cross-modal neuron paths, AIM aligns identity features under a Fisher constraint, AUVIC isolates visual concepts through adversarial perturbations, and MLLMEraser performs input-aware activation steering at inference time (Zeng et al., 2025; Li et al., 2026; Lee et al., 2026; Chen et al., 2026; Ding et al., 2025). We evaluate representative objectives and intervention locations under a common group-local protocol, focusing on where their side effects occur. Existing multimodal benchmarks cover complementary settings. MLLMU-Bench combines fictitious and celebrity profiles, CLEAR and PEBench study fictitious people and their associated events, and UnLOK-VQA adds proximity-controlled VQA and attack and defense evaluation (Liu et al., 2025a; Dontsov et al., 2025; Xu et al., 2025; Patil et al., 2025). SafeEraser targets harmful content, whereas PULSE examines pretrained knowledge and sequential deletion (Chen et al., 2025; Kawakami et al., 2025). Recent work also evaluates partial facts about public figures and private targets co-occurring with public figures or landmarks; AUVIC's VCUBench tests preservation of similar visual concepts (Guang et al., 2026; Zhang et al., 2026; Chen et al., 2026). MUREN-Bench complements these efforts by jointly combining pretrained real-world targets, six entity families, three access routes, and a hierarchical cross-entity retain neighborhood. It thereby turns preservation from a single utility score into a spatially resolved diagnosis of ripple damage.

3 MUREN-Bench

MUREN-Bench jointly evaluates target forgetting and locality, exposing damage near the target that a single averaged retain score can obscure. It combines three components: real-world entity neighborhoods ordered by relatedness, a role-specific split between training and held-out entities, and multimodal probes through three access routes. Dual-model screening further ensures that every image, for every role, recognizably depicts its entity. Table 1 situates this design among representative multimodal unlearning benchmarks. MUREN-Bench extends coverage from people, or people and events, to six entity families; it uses real-world forget targets; and it is the only benchmark in the comparison that evaluates locality hierarchically across entity neighborhoods. With 57,240 template-expanded probes, it also provides more QA/VQA instances than the listed benchmarks under their respective counting protocols.

3.1 Benchmark Composition

We select six entity families (artifacts, biological entities, celebrities, IP characters, landmarks, and trademarks) to cover varied forms of visual recognition and entity knowledge. These families span objects, organisms, people, fictional characters, places, and brands, allowing us to examine whether locality failures generalize across domains with different visual cues and semantic relationships. We

Table 1: Comparison of multimodal unlearning benchmarks in entity coverage, real-world forget targets, retention structure, and evaluation scale. MUREN-Bench combines diverse real-world targets with hierarchical retention to evaluate locality within entity neighborhoods.

Benchmark	Target types	Real-world targets	Hierarchical retention	QA/VQA pairs
MLLMU-Bench (Liu et al., 2025a)	People	×	×	20,754
CLEAR (Dontsov et al., 2025)	People	×	×	7,770
PEBench (Xu et al., 2025)	People, events	×	×	16,000
UnLOK-VQA (Patil et al., 2025)	Open-domain facts	✓	×	500
PULSE (Kawakami et al., 2025)	People	✓	×	700
PPU-Bench (Guang et al., 2026)	People	✓	×	24,334
MUREN-Bench (ours)	6 entity families	✓	✓	57,240

include 15 groups per family, yielding 90 groups: multiple neighborhoods capture variation within each family, while equal group counts keep any single family from dominating the benchmark.

Each group pairs one forget target with three retain entities, giving four roles per group and 360 entity roles in total. We order the retain entities by visual and real-world relatedness to the target: T1 is the closest neighbor, T2 an intermediate neighbor, and T3 a distant anchor. This hierarchy separates damage to close neighbors from changes at intermediate and distant levels, turning locality into a profile over proximity to the target rather than a single averaged score.

Figure 1(b) shows an example. With Mona Lisa as the target, Lady with an Ermine (T1) shares its artist and portrait format; The Virgin and Child with St. Anne (T2) shares the artist but depicts a different composition; and The Thinker (T3) is a more distant artwork by a different artist in a different medium. We check visual proximity by comparing target and neighbor image representations from CLIP and DINOv2 (Radford et al., 2021; Oquab et al., 2023), using higher similarity as evidence of a closer relationship.

The benchmark separates the roles used for optimization from those used for diagnosis. In every group, the target contributes forget examples and T3 contributes retain examples, while T1 and T2 are held out from unlearning training. Because no method receives direct supervision on the closest neighbors, their post-update behavior provides a genuine test of collateral damage. This split also lets us compare methods under the same amount of forget and retain supervision, independent of the number of images available for a particular entity.

Each entity is queried through three access routes: image-only, image-text, and text-only. The image-only route pairs an image with a generic identification question and tests visual recognition without naming cues. The image-text route adds a short description, testing whether visual and textual evidence together still elicit the entity. The text-only route uses the same description without the image, measuring knowledge accessible through language alone. Because the image-text and text-only routes share the same description, differences across routes reflect modality access rather than content. We expand a common set of templates over all roles, images, and routes, and keep answers open-ended, scoring them against accepted aliases rather than forcing a closed-set choice.

3.2 Data Collection

To obtain clear images for the entity neighborhoods in Section 3.1, we first use GPT to propose forget targets and candidate retain entities across the six families. As illustrated in Figure 2(a), each entity name serves as a search query; for example, "Giant Panda" retrieves candidate images from Google Images, Bing Images, and Wikimedia Commons. We keep only images whose copyright and usage terms permit research use, and remove images with watermarks, visible blur, or other quality defects that could confound recognition.

To ensure that each image recognizably depicts its assigned entity, we apply the dual-model screening in Figure 2(b). GPT-4o-mini and Gemini-2.5-Flash-Lite independently identify the entity in every candidate image, and an image is kept only when both models identify it correctly. Requiring agreement reduces dependence on the recognition behavior of any single model. This criterion applies to all target, T1, T2, and T3 images used for training or evaluation. Experts then review ambiguous samples and neighborhood relationships, checking both entity correspondence and the coherence of each group.

To produce reliable annotations while keeping manual effort manageable, we combine automated drafting with human verification (Figure 2(c)). The automated stage drafts entity labels, identification questions, and entity descriptions; human reviewers check these drafts against the images,

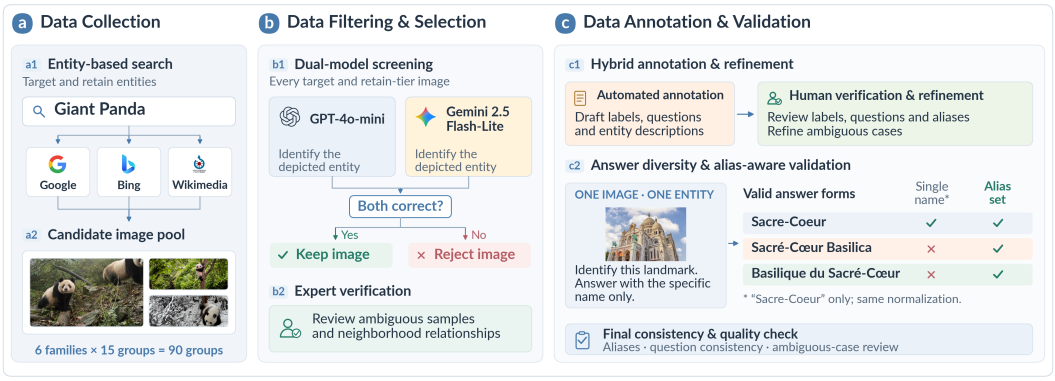

Figure 2: **Data collection, screening, and annotation.** (a) Entity-based queries gather candidate images from multiple web sources across six families and 90 groups. (b) Every target and retain image is kept only when both GPT-4o-mini and Gemini-2.5-Flash-Lite identify its entity correctly, followed by expert verification. (c) Automated annotation is refined by human review and consistency checks. The landmark example shows how an accepted alias set recognizes valid name variants that a single reference name would miss.

correct inaccurate or ambiguous wording, and refine the accepted aliases. Each entity is associated with a set of acceptable answers so that valid name variants are not penalized. For example, "Basilique du Sacré-Cœur" and "Sacré-Cœur" refer to the same building, but if the reference name is the former, a response giving only the latter would be missed. Finally, we remove duplicate descriptions and check consistency across images, labels, questions, descriptions, and aliases.

3.3 Evaluation Metrics

We evaluate target forgetting and locality through recognition accuracy. A response is counted as correct when its normalized text contains an accepted name or alias of the expected entity. We compute accuracy for each entity family and access route, then average with equal weights.

For target forgetting, we report Unlearning Accuracy (UA), the recognition accuracy on forget targets, before and after unlearning; lower post-unlearning UA indicates stronger forgetting. For locality, we report Retain Accuracy separately for T1, T2, and T3, denoted RA_1, RA_2, and RA_3; higher RA indicates better locality at that level. We refer to the drop in RA from the original model as collateral damage at that level.

The overall Score weights forgetting and locality equally, with locality weighted 3:2:1 across T1, T2, and T3 to emphasize closer neighbors:

$$\text{Score} = 0.5F + 0.5\,\frac{3\mathrm{RA}_1 + 2\mathrm{RA}_2 + \mathrm{RA}_3}{6}, \qquad F = 100 \times \frac{\mathrm{UA}_{\text{orig}} - \mathrm{UA}_{\text{unl}}}{\mathrm{UA}_{\text{orig}}}.$$

F is the relative reduction in UA; for example, a decrease from 80% to 20% gives $F = 75$. All RA values are percentages, and higher Scores indicate a better balance between target forgetting and locality.

In result tables, subscripts give changes from the original model in percentage points. Green indicates improvement (UA down or RA up) and red indicates degradation; bold and underlining mark the best and second-best unlearning methods within each model and column.

4 Experiments

4.1 Experiment Setup

Baseline Method. We compare seven unlearning baselines: gradient ascent (GA) (Jang et al., 2023), gradient difference (GD) (Maini et al., 2024), KL-regularized forgetting (KLMin) (Maini et al., 2024), negative preference optimization (NPO) (Zhang et al., 2024), direct preference optimization (DPO) (Rafailov et al., 2023), MMUnlearner (Huo et al., 2025), and MLLMEraser (Ding et al., 2025). The original model provides the reference before unlearning.

Table 2: Overall performance of seven unlearning methods on two multimodal models. Joint evaluation of target unlearning and hierarchical retention reveals whether target suppression preserves neighboring knowledge, with the overall Score summarizing this balance.

LLaVA-1.5-7B					
Method	UA ↓	RA_1 ↑	RA_2 ↑	RA_3 ↑	Score ↑
Original	64.7	34.8	54.4	72.2	—
GA	**35.5**$_{(-29.2)}$	32.4$_{(-2.4)}$	49.3$_{(-5.1)}$	65.9$_{(-6.3)}$	**44.4**
GD	43.2$_{(-21.5)}$	34.4$_{(-0.4)}$	52.9$_{(-1.5)}$	83.1$_{(+10.9)}$	41.0
KLMIN	43.2$_{(-21.5)}$	34.5$_{(-0.3)}$	52.9$_{(-1.5)}$	83.1$_{(+10.9)}$	41.0
NPO	44.1$_{(-20.7)}$	34.9$_{(+0.1)}$	53.4$_{(-1.0)}$	84.3$_{(+12.1)}$	40.6
DPO	52.5$_{(-12.2)}$	**36.5**$_{(+1.6)}$	**56.4**$_{(+2.0)}$	**90.9**$_{(+18.7)}$	35.5
MMUnlearner	47.3$_{(-17.4)}$	34.4$_{(-0.4)}$	54.5$_{(+0.1)}$	87.1$_{(+14.9)}$	38.4
MLLMEraser	50.9$_{(-13.9)}$	33.6$_{(-1.2)}$	48.3$_{(-6.1)}$	71.1$_{(-1.1)}$	33.1
Qwen3-VL-8B					
Method	UA ↓	RA_1 ↑	RA_2 ↑	RA_3 ↑	Score ↑
Original	85.1	65.5	76.6	87.7	—
GA	50.2$_{(-34.9)}$	48.0$_{(-17.5)}$	63.0$_{(-13.6)}$	80.1$_{(-7.6)}$	49.7
GD	46.1$_{(-39.1)}$	54.0$_{(-11.5)}$	62.8$_{(-13.8)}$	93.3$_{(+5.6)}$	54.7
KLMIN	50.1$_{(-35.1)}$	56.5$_{(-9.0)}$	66.7$_{(-9.9)}$	93.8$_{(+6.1)}$	53.6
NPO	49.0$_{(-36.1)}$	59.7$_{(-5.8)}$	68.1$_{(-8.5)}$	**95.9**$_{(+8.3)}$	55.5
DPO	**15.1**$_{(-70.0)}$	50.1$_{(-15.4)}$	58.9$_{(-17.6)}$	95.9$_{(+8.2)}$	**71.5**
MMUnlearner	59.2$_{(-26.0)}$	60.8$_{(-4.7)}$	69.3$_{(-7.2)}$	95.4$_{(+7.8)}$	50.0
MLLMEraser	79.5$_{(-5.6)}$	**61.5**$_{(-4.0)}$	**74.5**$_{(-2.0)}$	86.4$_{(-1.3)}$	38.3

Implementation Details. We evaluate LLaVA-1.5-7B and Qwen3-VL-8B(Liu et al., 2023; Bai et al., 2025), starting each method and group independently from the same original checkpoint. Each group has 10 forget examples from the target and 10 retain examples from T3; T1 and T2 are used only for evaluation. The six LoRA-based methods use rank 16, alpha 16, dropout 0.05, learning rate 2×10^{-5}, five epochs, gradient accumulation 1, bfloat16 precision, maximum gradient norm 1.0, and seed 42 (Hu et al., 2021). Only decoder LoRA adapters are updated; the backbone, vision encoder, and multimodal projector remain frozen. We use AdamW with linear learning-rate decay and no warmup. KLMin, NPO, and DPO use the base model with adapters disabled as the frozen reference. For MMUnlearner, we compute masks before LoRA injection; for MLLMEraser, we normalize steering directions and cap their magnitude relative to each layer's hidden-state norm. Appendix A details the objectives, adapter configuration, and steering procedure.

4.2 Main Results

Overall Performance. Table 2 shows that method rankings depend on the model and the retention objective. GA achieves the highest Score on LLaVA, whereas DPO leads on Qwen under the same training budget. DPO also exhibits different retention behavior across the two models: on LLaVA, it improves all three retain accuracies but suppresses target recognition less effectively than GA; on Qwen, it achieves the lowest UA while reducing T1 and T2 accuracy. Stronger unlearning does not invariably require greater retention loss. On Qwen, GA and KLMin reach similar UA, around 50%, but KLMin preserves substantially more knowledge at every retain level. NPO, meanwhile, ranks above GD in Score despite its higher UA, because it retains more knowledge at all three levels. The methods also differ in how retention gains extend beyond the supervised tier. Although retain training uses only T3, GD, KLMin, NPO, DPO, and MMUnlearner all alter recognition at the untrained T1 and T2 levels: on Qwen, they improve T3 accuracy while degrading both closer levels. DPO raises T3 accuracy to 95.9%, yet T1 and T2 fall by 15.4 and 17.6 points.

Category-wise Unlearning and Retention. Table 3 and Figure 3 reveal differences across entity families that are obscured by overall accuracy. On LLaVA, GA achieves the lowest UA in five categories, while MLLMEraser is strongest on trademarks. On Qwen, DPO leads across all six categories, but its trademark UA remains 57.7%, compared with at most 10.4% for every other category. Original recognition also affects the interpretation of low final accuracy. LLaVA starts with artifact UA of 23.5% and IP-character UA of 83.2%. GA's artifact UA of 8.1% represents a 15.4-point reduction, whereas its higher IP-character UA of 30.5% follows a much larger 52.7-point reduction. Retention loss likewise varies across categories and does not uniformly decrease with

Table 3: Category-wise unlearning across six real-world entity families and two models. Comparison with Original reveals how target suppression varies across categories, highlighting differences that can be obscured by a single overall accuracy.

LLaVA-1.5-7B

Method	Artifact ↓	Biology ↓	Celebrity ↓	IP Char. ↓	Landmark ↓	Trademark ↓
Original	23.5	54.9	78.2	83.2	62.2	86.3
GA	**8.1**$_{(-15.4)}$	**7.8**$_{(-47.1)}$	**53.8**$_{(-24.4)}$	**30.5**$_{(-52.7)}$	**28.4**$_{(-33.9)}$	84.3$_{(-2.0)}$
GD	11.2$_{(-12.3)}$	14.6$_{(-40.3)}$	66.6$_{(-11.7)}$	47.8$_{(-35.4)}$	33.7$_{(-28.5)}$	85.4$_{(-0.9)}$
KLMIN	11.2$_{(-12.3)}$	15.2$_{(-39.7)}$	66.6$_{(-11.7)}$	47.9$_{(-35.3)}$	33.1$_{(-29.1)}$	85.4$_{(-0.9)}$
NPO	13.2$_{(-10.3)}$	16.1$_{(-38.8)}$	64.0$_{(-14.2)}$	47.6$_{(-35.6)}$	37.9$_{(-24.3)}$	85.6$_{(-0.8)}$
DPO	17.1$_{(-6.4)}$	28.9$_{(-26.0)}$	73.5$_{(-4.8)}$	65.3$_{(-17.9)}$	46.0$_{(-16.2)}$	84.2$_{(-2.1)}$
MMUnlearner	14.4$_{(-9.1)}$	17.8$_{(-37.1)}$	73.6$_{(-4.6)}$	55.0$_{(-28.2)}$	37.5$_{(-24.7)}$	85.7$_{(-0.6)}$
MLLMEraser	9.6$_{(-13.9)}$	36.2$_{(-18.7)}$	55.2$_{(-23.0)}$	74.6$_{(-8.6)}$	55.0$_{(-7.2)}$	**74.6**$_{(-11.8)}$

Qwen3-VL-8B

Method	Artifact ↓	Biology ↓	Celebrity ↓	IP Char. ↓	Landmark ↓	Trademark ↓
Original	81.1	73.6	83.6	92.5	90.1	90.0
GA	38.1$_{(-43.0)}$	46.1$_{(-27.5)}$	33.7$_{(-49.9)}$	52.5$_{(-39.9)}$	42.7$_{(-47.4)}$	88.1$_{(-1.8)}$
GD	36.0$_{(-45.1)}$	28.9$_{(-44.7)}$	42.4$_{(-41.2)}$	50.9$_{(-41.6)}$	31.4$_{(-58.7)}$	86.7$_{(-3.2)}$
KLMIN	42.6$_{(-38.5)}$	39.6$_{(-34.0)}$	46.9$_{(-36.7)}$	41.8$_{(-50.7)}$	43.9$_{(-46.1)}$	85.6$_{(-4.4)}$
NPO	39.9$_{(-41.2)}$	43.9$_{(-29.8)}$	46.9$_{(-36.8)}$	46.3$_{(-46.2)}$	30.9$_{(-59.2)}$	86.4$_{(-3.6)}$
DPO	**7.3**$_{(-73.8)}$	**10.4**$_{(-63.2)}$	**5.2**$_{(-78.4)}$	**9.8**$_{(-82.6)}$	**0.1**$_{(-90.0)}$	**57.7**$_{(-32.3)}$
MMUnlearner	62.4$_{(-18.7)}$	41.2$_{(-32.4)}$	54.6$_{(-29.0)}$	58.8$_{(-33.7)}$	50.8$_{(-39.2)}$	87.2$_{(-2.8)}$
MLLMEraser	72.6$_{(-8.5)}$	60.1$_{(-13.5)}$	85.8$_{(+2.1)}$	86.8$_{(-5.7)}$	85.6$_{(-4.5)}$	86.3$_{(-3.6)}$

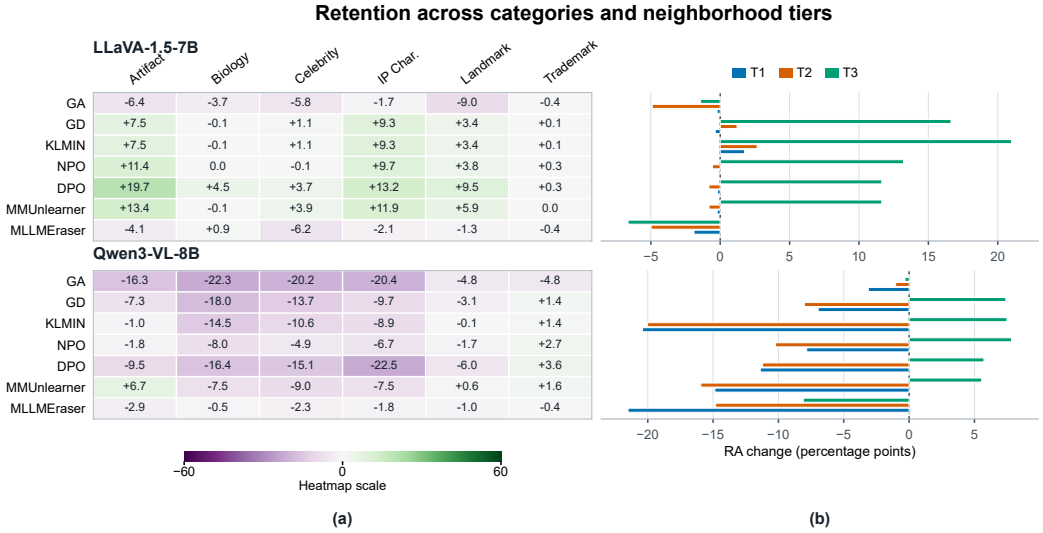

Figure 3: Retention changes from Original across six entity categories and three neighborhood levels for all seven methods. The two model panels reveal how preservation varies with entity relatedness, highlighting whether gains on distant neighbors extend to closer ones.

distance. Figure 3 reports changes in probe-weighted accuracy within each category and tier. Under Qwen's DPO, biological entities and IP characters lose more than 37 points at T1 while gaining accuracy at T3; trademarks instead gain at T1 as well as T3. LLaVA's GA exhibits the opposite distance pattern for artifacts and landmarks, with greater losses at T3 than at T1.

Unlearning across Modalities. Figure 4 compares UA reductions from Original under image-only (I), image-text (I+T), and text-only (T) queries. Suppression through one input route can leave another largely intact. For Qwen's trademarks, GD, KLMin, NPO, and MMUnlearner reduce text-only UA while leaving image-based recognition almost unchanged. DPO is an exception, reducing trademark UA across all three modalities, although the category remains its largest source of residual recognition. The most persistent input route also varies across settings. On LLaVA, GA reduces celebrity UA by about 33 points for image-only queries and 27 points for image-text queries,

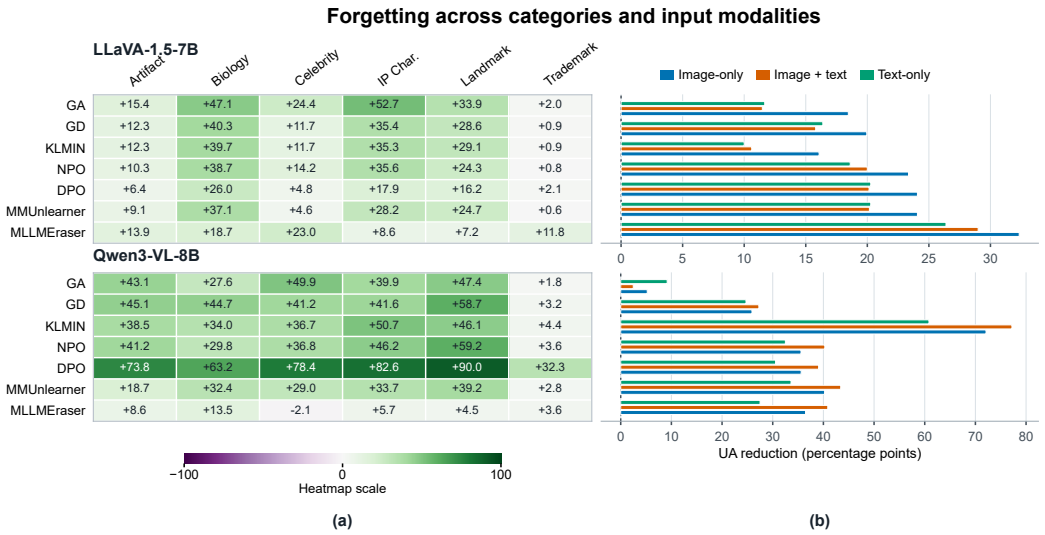

Figure 4: Unlearning across six entity categories and three input modalities for all seven methods. Colors show UA reductions from Original, revealing uneven suppression across query formats. Red outlines identify cases where target recognition increases.

Method	Target	T1 · closest	T2 · intermediate	T3 · distant
GA	**Ostrich** B: ostrich **A: rhea**	**Emu** B: Emu **A: rhea**	**Peacock** B: peacock **A: rhea**	**Horse** B: Horse **A: horse**
GD	**Polar bear** B: Polar bear **A: bear**	**Arctic fox** B: Arctic fox **A: Fox**	**Walrus** B: Walrus **A: Walrus**	**Camel** B: Camel **A: camel**
KLMin	**Crocodile** B: Crocodile **A: Caiman**	**Alligator** B: Alligator **A: Caiman**	**Komodo dragon** B: Komodo dragon **A: Caiman**	**Turtle** B: Turtle **A: Turtle**
NPO	**Goku** B: Goku **A: Gogeta**	**Gohan** B: Gohan **A: Goten**	**Vegeta** B: Vegeta **A: Vegeta**	**Iron Man** B: Iron Man **A: Iron Man**
DPO	**Cheetah** B: Cheetah **A: I don't know.**	**Leopard** B: Leopard **A: I don't know.**	**Lion** B: Lion **A: Lion**	**Zebra** B: Zebra **A: Zebra**
MMUnlearner	**Audi** B: Audi **A: Audi**	**Olympic rings** B: Olympic Rings **A: Audi**	**BMW** B: BMW **A: BMW**	**McDonald's** B: McDonald's **A: McDonalds logo**
MLLMEraser	**Crocodile** B: Crocodile **A: gharial**	**Alligator** B: Alligator **A: alligator**	**Komodo dragon** B: Komodo dragon **A: Komodo dragon**	**Turtle** B: Turtle **A: Turtle**

Figure 5: Image-only Qwen3-VL-8B(Bai et al., 2025) cases across seven methods and Target/T1–T3. B/A denote responses before/after unlearning. Green marks correct neighbors, red errors or refusal, and amber retained targets, broader answers, or typos.

compared with only 13 points for text-only queries; text-only UA remains 76.7%. Here, suppression is weaker through text, whereas Qwen's trademark recognition under GD, KLMin, NPO, and MMUnlearner changes primarily through text. Average suppression can also coexist with local increases in recognition. MLLMEraser reduces Qwen's overall UA, but increases its celebrity UA in all three modalities, as marked by the red outlines in Figure 4. Image-only UA rises from 87.9% to 89.5%, image-text UA from 93.0% to 96.1%, and text-only UA from 70.0% to 71.7%.

Method	Target	T1 · closest	T2 · intermediate	T3 · distant
GA	 **Moose** B: Moose **A: Giraffe**	 **Reindeer** B: Reindeer **A: Giraffe**	 **Camel** B: Camel **A: Camel**	 **Tiger** B: Tiger **A: Tiger**
GD	 **Link** B: Link **A: Zelda**	**Zelda** B: Link **A: Zelda**	 **Ganon** B: Ganon **A: Ganon**	 **Hello Kitty** B: Hello kitty **A: Hello Kitty**
KLMin	 **Ostrich** B: Ostrich **A: Bird**	 **Emu** B: Ostrich **A: Bird**	 **Peacock** B: Peacock **A: Bird**	 **Horse** B: Horse **A: Horse**
NPO	 **Koala** B: Koala **A: Kangaroo**	 **Sloth** B: Sloth **A: Sloth**	 **Orangutan** B: Orangutan **A: Monkey**	 **Penguin** B: Penguin **A: Penguin**
DPO	 **Mario** B: Mario **A: Mario**	 **Luigi** B: Mario **A: Luigi**	 **Bowser** B: Mario **A: Bowser**	 **Superman** B: Superman **A: Superman**
MMUnlearner	 **Cheetah** B: Cheetah **A: Leopard**	 **Leopard** B: Leopard **A: Leopard**	 **Lion** B: Lion **A: Lion**	 **Zebra** B: Zebra **A: Zebra**
MLLMEraser	 **McDonald's** B: Mcdonald's **A: Mcdonald't**	 **Burger King** B: Burger king **A: Burger king**	 **KFC** B: Kfc **A: Kfc**	 **Sony** B: Sony **A: Sony**

Figure 6: Image-only LLaVA-1.5-7B(Liu et al., 2023) cases across seven methods and Target/T1–T3. B/A denote responses before/after unlearning. Green marks correct neighbors, red errors or refusal, and amber retained targets, broader answers, or typos.

4.3 Visual Analysis

Figures 5 and 6 show selected image-only cases for Qwen and LLaVA across all seven methods; Appendix B provides additional cases for both models. Original and updated responses are paired on identical images and questions within each row. The Qwen examples expose different ways in which target suppression affects the neighborhood. GA maps the ostrich, emu, and peacock to the same answer, "rhea", while preserving the horse at T3. KLMin similarly replaces crocodile, alligator, and Komodo dragon with "Caiman", leaving the turtle unchanged. These responses lose distinctions among the target and nearby entities rather than selectively withholding the target name. DPO instead returns "I don't know." for both the cheetah target and the leopard at T1, while still identifying the lion and zebra. GD produces broader names, "bear" and "Fox", for polar bear and Arctic fox; NPO substitutes related character names for Goku and Gohan while retaining Vegeta and Iron Man. Target recognition and neighbor damage can also move independently: MMUnlearner still identifies Audi but newly calls the Olympic rings "Audi". In the crocodile case, MLLMEraser changes only the target answer to "gharial" and preserves all three neighbors, whereas KLMin alters T1 and T2 on the same displayed images.

5 Conclusion

Selective multimodal unlearning should remove requested knowledge while preserving the related knowledge that remains useful. MUREN-Bench makes this requirement explicit through real-world entity neighborhoods and hierarchical retention evaluation. Our experiments show that suppressing a target can also affect its neighbors, motivating joint assessment of target forgetting and neighborhood preservation. We hope this benchmark supports methods that control the scope of forgetting and preserve surrounding knowledge more reliably.

Limitations. Our evaluation is limited to selected models, English prompts, and a fixed training budget. Name-based scoring measures recognition behavior; refusal can reduce target accuracy without establishing knowledge erasure. Overlap between target training and evaluation images, repeated entities, and imperfect neighborhood ordering also limit generalization. Broader coverage and stronger semantic validation remain directions for future work.

Ethics Statement

MUREN-Bench includes public figures, copyrighted characters, and trademarks, using only public images whose terms permit research use. Celebrities are probed solely for identity recognition, and no private information is collected. Low target accuracy reflects recognition behavior, not verified knowledge erasure.

Reproducibility Statement

Section 3 and Appendix A detail the data pipeline and training setup; source code is in the supplementary material.

AI Use Statement

Generative AI proposed candidate entities, screened images, and drafted annotations (Section 3); human annotators reviewed and corrected all outputs. AI also assisted with writing. The authors take full responsibility for all content.

References

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

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Haokun Chen, Jianing Li, Yao Zhang, Jinhe Bi, Yan Xia, Jindong Gu, and Volker Tresp. Auvic: Adversarial unlearning of visual concepts for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30218–30225, 2026.

Junkai Chen, Zhijie Deng, Kening Zheng, Yibo Yan, Shuliang Liu, PeiJun Wu, Peijie Jiang, Jia Liu, and Xuming Hu. Safeeraser: Enhancing safety in multimodal large language models through multimodal machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14194–14224, 2025.

Chenlu Ding, Jiancan Wu, Leheng Sheng, Fan Zhang, Yancheng Yuan, Xiang Wang, and Xiangnan He. Mllmeraser: Achieving test-time unlearning in multimodal large language models through activation steering. *arXiv preprint arXiv:2510.04217*, 2025.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Clear: Character unlearning in textual and visual modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20582–20603, 2025.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zico Kolter, Zachary Lipton, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *Advances in Neural Information Processing Systems*, 38, 2026.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*, 2023.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12043–12051, 2024.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9301–9309. IEEE, 2020.

Jiahui Guang, Zexun Zhan, Zhenlin Xu, Cuiyun Gao, Haiyan Wang, Jing Li, Zhaoquan Gu, and Yanchun Zhang. Ppu-bench: Real world benchmark for personalized partial unlearning in vision language models. *arXiv preprint arXiv:2605.08800*, 2026.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, and Xuming Hu. Mmunlearner: Reformulating multimodal machine unlearning in the era of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7190–7206, 2025.

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

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. Soul: Unlocking the power of second-order optimization for llm unlearning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4276–4292, 2024.

Tatsuki Kawakami, Kazuki Egashira, Atsuyuki Miyai, Go Irie, and Kiyoharu Aizawa. Pulse: Practical evaluation scenarios for large multimodal model unlearning. *arXiv preprint arXiv:2507.01271*, 2025.

Wonjun Lee, Jaehyuk Jang, Kangwook Ko, Hee-Seon Kim, and Changick Kim. Aim: Anchor identity features, then match for multimodal large language model unlearning. *arXiv preprint arXiv:2608.28312*, 2026.

Kunhao Li, Wenhao Li, Di Wu, Lei Yang, Jun Bai, Ju Jia, and Jason Xue. Cross-modal unlearning via influential neuron path editing in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35589–35597, 2026.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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 *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)*, pp. 4105–4135, 2025a.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. Modality-aware neuron pruning for unlearning in multimodal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5913–5933, 2025b.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Vaidehi Patil, Yi-Lin Sung, Peter Hase, Jie Peng, Tianlong Chen, and Mohit Bansal. Unlearning sensitive information in multimodal llms: Benchmark and attack-defense evaluation. *arXiv preprint arXiv:2505.01456*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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

Zhaopan Xu, Pengfei Zhou, Weidong Tang, Jiaxin Ai, Wangbo Zhao, Kai Wang, Xiaojiang Peng, Wenqi Shao, Hongxun Yao, and Kaipeng Zhang. Pebench: A fictitious dataset to benchmark machine unlearning for multimodal large language models. *arXiv preprint arXiv:2503.12545*, 2025.

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

Zhen Zeng, Leijiang Gu, Zhangling Duan, Feng Li, Cees GM Snoek, Meng Wang, and Zenglin Shi. Towards benign memory forgetting for selective multimodal large language model unlearning. *arXiv preprint arXiv:2511.20196*, 2025.

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

Xianren Zhang, Delvin Ce Zhang, Dongwon Lee, and Suhang Wang. Ppe-bench: A benchmark for evaluating mllm unlearning under private-public entanglement. *arXiv preprint arXiv:2607.02897*, 2026.

A Implementation Details

This appendix describes the implementations used for the seven baseline methods, including their shared optimization settings and method-specific adaptations. All interventions are constructed independently for each group from the original checkpoint, so updates for one target do not carry over to another.

Shared training configuration. GA, GD, KLMin, NPO, DPO, and MMUnlearner update LoRA adapters in the language decoder. We attach adapters to the query, key, value, and output projections of attention and the gate, up, and down projections of the feed-forward blocks. The underlying model weights, vision encoder, and multimodal projector remain frozen during optimization. Each group provides five target images and five T3 images, with two questions per image, yielding 10 forget and 10 retain examples. T1 and T2 do not enter training. Images are resized with their aspect ratio preserved and maximum side length 1,024 pixels. We use each model's conversation template and mask prompt and image positions in the supervised answer loss.

The adapters use rank 16, scaling alpha 16, and dropout 0.05. AdamW starts at learning rate 2×10^{-5} with linear decay and no warmup. Training runs for five epochs with one example per loss term, gradient accumulation 1, bfloat16 precision, gradient checkpointing, gradient clipping at norm 1.0, and seed 42. Forget and retain examples are shuffled independently each epoch. With 10 examples in each set, this gives 50 optimizer updates per group. All methods with a retain cross-entropy term use coefficient 1.0. Reference-based losses use the same base model with adapters disabled, evaluated without gradient updates. Sequence log-probabilities for NPO and DPO sum the supervised answer-token log-probabilities, accounting for image-token expansion when aligning labels and model outputs.

GA. Gradient ascent (Jang et al., 2023) minimizes the negative cross-entropy of the correct target answer on forget examples, reducing its likelihood. It uses no retain loss. This provides a direct target-suppression baseline against which methods with explicit preservation objectives can be compared.

GD. Gradient difference (Maini et al., 2024) combines negative cross-entropy on a forget example with ordinary cross-entropy on a T3 retain example. Both terms have unit weight and contribute to the same optimizer update. The retain term encourages correct answers on the available T3 examples while the forget term discourages correct target answers.

KLMin. KLMin (Maini et al., 2024) adds a reference-distribution constraint to GD. On retain examples, we compute the token-level KL divergence from the frozen reference distribution to the current model distribution, summing over the vocabulary and averaging over sequence positions. This term has coefficient 1.0 and is applied alongside retain cross-entropy. Unlike the supervised answer loss, the KL term uses all aligned sequence positions. It encourages preservation of the reference model's output distribution beyond the probability assigned to the single correct answer.

NPO. Negative preference optimization (Zhang et al., 2024) treats the correct target answer as an undesirable response. Its loss uses the answer's log-probability relative to the frozen reference, with $\beta = 0.4$ and the standard $2/\beta$ scaling. As the answer becomes less likely than under the reference, the loss reduces the strength of further suppression. We add the T3 retain cross-entropy term to this objective; NPO does not require a preferred response for the forget examples.

DPO. Direct preference optimization (Rafailov et al., 2023) uses a response pair for each forget example: the chosen response is *I don't know.* and the rejected response is the correct entity answer. We optimize the reference-adjusted log-probability margin between these responses with $\beta = 0.1$, together with T3 retain cross-entropy. This explicitly trains a preference for withholding the target answer. Accordingly, lower UA under this method measures reduced target recognition in the evaluated responses, rather than establishing that the underlying knowledge has been erased.

MMUnlearner. Our LoRA implementation adapts the saliency-based masking of MMUnlearner (Huo et al., 2025). Before attaching adapters, we separately accumulate squared gradients of the

supervised forget and retain losses on the relevant base-model linear weights. Averaging along each weight matrix's rows and columns yields output-axis and input-axis saliency estimates. We retain axes whose forget-to-retain saliency ratio is at least 1.0, then map the selected input axes to columns of the LoRA A matrix and the selected output axes to rows of the LoRA B matrix. During training, the mask is applied to the gradient from negative forget cross-entropy. The retain cross-entropy gradient is then added without masking, and the combined gradient is clipped before one optimizer update. All 10 examples from each training set contribute to mask construction.

MLLMEraser. MLLMEraser (Ding et al., 2025) uses activation steering without optimizing model parameters. For each decoder layer, we average hidden states over the answer tokens for the refusal and correct-answer responses on the 10 forget examples. Their difference, refusal minus correct answer, is normalized to form a steering direction. Separately, last-prompt-token states from the forget and T3 retain examples define two centroids. At inference, an initial unsteered forward pass computes a gate for each layer by applying a sigmoid, with scale 25, to the difference between the query's cosine similarities to these centroids. The resulting gates are fixed for that query during generation. We add the gated direction to each decoder layer's output, using strength 0.03 for LLaVA and 0.025 for Qwen times the layer's mean calibration hidden-state norm. The strength is capped at 3.0 before gating. These settings bound the intervention while making it more active for inputs closer to the forget centroid.

B Additional Visualization Results

We show three complete Target/T1–T3 groups for each of seven methods, separately for Qwen3-VL-8B and LLaVA-1.5-7B under Image-only, Image+Text, and Text-only inputs: 21 groups per panel and 126 groups in total. Each pair uses the same probe before and after its group-local update. The cases are selected to make target suppression, neighboring recognition, and input-modality differences inspectable; they do not estimate how often each behavior occurs. All groups differ from the main-text cases. Groups may recur across methods and modalities to aid comparison.

B.1 Image-only: target suppression and neighborhood damage can dissociate

Figures 7–8 retain the original Image-only panels. For Qwen under GA, Gorilla remains recognizable as "gorilla gorilla," while the correctly recognized T1 Chimpanzee changes to that same answer; Monkey and Dog remain recognizable. KLMin likewise retains Kangaroo but changes Wallaby to "kangaroo" and Wombat to "wallaby," with Ostrich unchanged. These groups expose damage around a target that is still identifiable. In contrast, Qwen DPO changes Homer Simpson, Bart Simpson, and Peter Griffin from correct names to "I don't know." while retaining Naruto at T3. Target suppression here accompanies losses at both nearby tiers. LLaVA MLLMEraser retains Panda but changes Raccoon at T2 to "Opossum." Together, the cases explain why target UA and distant retention alone cannot characterize the local cost of an update.

B.2 Image+Text: descriptive input does not ensure local preservation

Figures 9–10 include both the displayed image and a role-specific description. Qwen DPO changes the correctly named Kangaroo, Wallaby, and Wombat to refusals while preserving Ostrich, showing that near-neighbor losses remain visible with descriptive input. Under MLLMEraser, Naruto and Boruto stay recognizable, but Sasuke changes from "Sasuke Uchiha" to "Itachi Uchiha"; Winnie the Pooh remains intact. LLaVA GD and KLMin produce "Lizard" for Chameleon, Iguana, and Snake, while Frog stays correct. This combines broader labels for two lizards with a new misidentification of Snake; it is not three equivalent identity errors. Some groups remain stable: Qwen identifies Shell and its neighbors under GD and KLMin. Such retained recognition provides a useful comparison with the Text-only cases below, rather than evidence that every method fails in the same way.

B.3 Text-only: suppression can differ by modality and miss the target

In Figures 11–12, Qwen GD and KLMin change Shell to "ExxonMobil," although Shell is retained in their Image+Text groups. Qwen NPO and MLLMEraser similarly change "Marvel Comics" to "DC Comics" in text while retaining Marvel in Image+Text. Conversely, Qwen DPO still identifies the Kangaroo target as "Red kangaroo" in text but changes the correctly named Wombat to "Bilby." For LLaVA, DPO and MMUnlearner retain BMW while changing Volkswagen at T1 to Mercedes-Benz; MLLMEraser instead changes Mercedes-Benz at T2 to Volkswagen. Google at T3 stays correct in all three groups. This is neighboring damage without suppression of the displayed target. The panels also separate variants and baseline errors: GA's Batman $\rightarrow$ Bruce Wayne remains the same identity, and Qwen DPO already misnames Homer and Peter before its Text-only refusals. Those refusals cannot establish newly lost recognition.

Reading the panels. Each cell gives the reference, original answer, and post-update answer in bold; answers are reproduced in full. Green marks recognizable neighbors or corrections; amber marks recognizable targets, broader labels, or minor name variants; red marks new identity changes or refusals; black marks pre-existing misses that remain misses. These qualitative annotations do not replace benchmark alias scores. Input codes resolve to the exact descriptions and templates in Figures 13–14. Since question wording and some images differ across modalities, these are observed response contrasts, not an isolated causal test of adding an image. The benchmark's contribution is to make these target, tier, and modality distinctions visible alongside aggregate results.

B Additional visualization: Image-only / Qwen3-VL-8B

Method		Target	T1 · closest	T2 · intermediate	T3 · distant
GA	Q1	**Gorilla** Gorilla **gorilla gorilla**	**Chimpanzee** Chimpanzee **gorilla gorilla**	**Monkey** Monkey **monkey**	**Dog** dog **dog**
	Q2	**Iron Man** Iron Man **Tony Stark**	**War Machine** War Machine **Okoye**	**Capt. America** Captain America **Captain America**	**Link** Link **Zelda**
	Q3	**Mastercard** Mastercard **Master Card**	**Flickr** Flickr **flickr**	**Visa** Visa **VISA**	**Ford** Ford **Ford**
GD	Q1	**Honey Bee** Honey bee **Bee**	**Wasp** Ichneumon wasp **Wasps**	**Butterfly** butterfly **Butterfly**	**Spider** Spider **Spider**
	Q2	**Naruto** Naruto Uzumaki **ninja**	**Boruto** Boruto Uzumaki **boruto**	**Sasuke** Sasuke Uchiha **Itachi**	**Winnie the Pooh** Winnie the Pooh **winnie the pooh**
	Q3	**BMW** BMW **bmw**	**Volkswagen** Volkswagen **vw**	**Mercedes-Benz** Mercedes-Benz **mercedes benz**	**Google** Google **Google**
KLMin	Q1	**Kangaroo** Kangaroo **kangaroo**	**Wallaby** Wallaby **kangaroo**	**Wombat** Wombat **wallaby**	**Ostrich** Ostrich **Ostrich**
	Q2	**Mickey Mouse** Mickey Mouse **Mickey Mouse**	**Oswald** Oswald **Oswald**	**Donald Duck** Daisy Duck **Donald Duck**	**Goku** Goku **Goku**
	Q3	**Puma** PUMA **PUMA**	**Jaguar** Jaguar **JAGUAR**	**Reebok** Reebok **REEBOK**	**Coca-Cola** Coca-Cola **COCA-COLA**
NPO	Q1	**Swan** Swan **Mute Swan**	**White Goose** Goose **Muscovy Duck**	**Pelican** Pelican **Brown Pelican**	**Bald Eagle** Bald eagle **Bald Eagle**
	Q2	**Batman** Batman **BATMAN**	**Robin** Robin **Robin**	**Black Panther** Black Panther **Black Panther**	**Homer Simpson** Homer Simpson **Homer Simpson**
	Q3	**IBM** IBM **Intel**	**HP** HP **hp**	**Intel** Intel **Intel**	**Disney** Disney+ **Disney+**
DPO	Q1	**Chameleon** Chameleon **I don't know.**	**Green Iguana** Iguana **Iguana**	**Snake** Snake **Black mamba**	**Frog** Frog **Frog**
	Q2	**Homer Simpson** Homer Simpson **I don't know.**	**Bart Simpson** Bart Simpson **I don't know.**	**Peter Griffin** Peter Griffin **I don't know.**	**Naruto** Naruto **Naruto Uzumaki**
	Q3	**Shell** Shell **I don't know.**	**Chupa Chups** Chupa Chups **Chupa Chups**	**ExxonMobil** ExxonMobil **ExxonMobil**	**Meta / Facebook** Meta **Meta**
MMUnlearner	Q1	**Golden retriever** Golden retriever **Domestic Dog**	**Labrador** Dog **Domestic Dog**	**German shep.** German Shepherd **Domestic Dog**	**Domestic Cat** Kitten **Domestic Cat**
	Q2	**Spider-Man** Spider-Man **Spider-Man**	**Deadpool** Deadpool **Deadpool**	**Venom** Venom **Venom**	**Mario** Mario **Mario**
	Q3	**Nike** Nike **Nike**	**Adidas** Adidas **Adidas**	**Under Armour** Under Armour **Under Armour**	**Rolex** Rolex **Rolex crown**
MLLMEraser	Q1	**Giant Panda** Panda **Panda**	**Brown Bear** Brown bear **bear**	**Raccoon** raccoon **raccoon**	**Kangaroo** Kangaroo **Kangaroo**
	Q2	**Darth Vader** Darth Vader **Darth Vader**	**Kylo Ren** Kylo Ren **Kylo Ren**	**Luke Skywalker** Luke Skywalker **Luke Skywalker**	**Capt. America** Captain America **Captain America**
	Q3	**Rolex** Rolex **Rolls-Royce**	**Hallmark** Hallmark **Hallmark**	**Omega** Omega **Omega**	**Spotify** Spotify **Spotify**

Each cell: reference name / before / after (bold).
Q1. What species or animal is shown in this image? Answer with the common name only.
Q2. Which fictional character is shown in this image? Answer with the name only.
Q3. What logo, brand, or company is shown in this image? Answer with the name only.

Figure 7: Qwen3-VL-8B, Image-only: three complete Target/T1–T3 groups per method (21 groups). Cases differ from the main text; colors are explained in Appendix B.

B Additional visualization: Image-only / LLaVA-1.5-7B

Method		Target	T1 · closest	T2 · intermediate	T3 · distant
GA	Q1	**Swan** Swan **Duck**	**White Goose** Duck **Duck**	**Pelican** Pelican **Duck**	**Bald Eagle** Eagle **Bird**
	Q2	**Batman** Batman **Cat**	**Robin** Robin **R**	**Black Panther** Black panther **Panther**	**Homer Simpson** Homer simpson **Homer**
	Q3	**Mastercard** Mastercard **Visa**	**Flickr** Flickr **Flickr**	**Visa** Visa **Visa**	**Ford** Ford **Ford**
GD	Q1	**Honey Bee** Bee **Bee**	**Wasp** Bee **Bee**	**Butterfly** Butterfly **Butterfly**	**Spider** Spider **Spider**
	Q2	**Mickey Mouse** Mickey mouse **Minnie mouse**	**Oswald** Oswald **Oswald**	**Donald Duck** Donald duck **Donald duck**	**Goku** Vegeta **Goku**
	Q3	**Puma** Puma **Humana**	**Jaguar** Jaguar **Jaguar**	**Reebok** Reed **Reebok**	**Coca-Cola** Coca cola **Coca-Cola**
KLMin	Q1	**Chameleon** Chameleon **Lizard**	**Green Iguana** Lizard **Lizard**	**Snake** Snake **Lizard**	**Frog** Frog **Frog**
	Q2	**Spider-Man** Spider-man **Spiderman**	**Deadpool** Deadpool **Deadpool**	**Venom** Venom **Spiderman**	**Mario** Mario **Mario**
	Q3	**Nike** Nike **Nike**	**Adidas** Adidas **Nike**	**Under Armour** Under armour **Under armour**	**Rolex** Rolex **Crown**
NPO	Q1	**Gorilla** Gorilla **Monkey**	**Chimpanzee** Chimp **Monkey**	**Monkey** Monkey **Monkey**	**Dog** Dog **Dog**
	Q2	**Homer Simpson** Homer simpson **Bart simpson**	**Bart Simpson** Homer simpson **Bart simpson**	**Peter Griffin** Pig **Peter griffin**	**Naruto** Naruto **Naruto**
	Q3	**Marvel** Marvel **Marvel**	**Supreme** Supreme **Supreme**	**Zara** Zara **ZAR**	**PlayStation** Sony **PlayStation**
DPO	Q1	**Kangaroo** Kangaroo **Koala**	**Wallaby** Kangaroo **Wallaby**	**Wombat** Kangaroo **Wombat**	**Ostrich** Ostrich **Ostrich**
	Q2	**Naruto** Naruto **Goku**	**Boruto** Boruto **Boruto**	**Sasuke** Naruto **Anime**	**Winnie the Pooh** Winnie pooh **Winnie the Pooh**
	Q3	**Porsche** Porsche **Porsche**	**Ferrari** Ferrari **Ferrari**	**Lamborghini** Lamborghini **Lamborghini**	**Samsung** Samsung **Samsung**
MMUnlearner	Q1	**African elephant** Elephant **Elephant**	**Asian elephant** Elephant **Elephant**	**Rhinoceros** Rhino **Rhino**	**Giraffe** Giraffe **Giraffe**
	Q2	**Darth Vader** Darth vader **Darth vader**	**Kylo Ren** Darth vader **Darth vader**	**Luke Skywalker** Luke skywalker **Princess leia**	**Capt. America** Captain america **Captain America**
	Q3	**Google** Google **Google**	**eBay** Ebay **eBay**	**Yahoo** Yahoo **Hulu**	**Louis Vuitton** Louis vuitton **Louis Vuitton**
MLLMEraser	Q1	**Giant Panda** Panda **Panda**	**Brown Bear** Bear **Bear**	**Raccoon** Raccoon **Opossum**	**Kangaroo** Kangaroo **Kangaroo**
	Q2	**Iron Man** Iron man **Iron man**	**War Machine** Iron man **Robocop**	**Capt. America** Captain america **Captain america**	**Link** Link **Link**
	Q3	**Rolex** Rolex **Gucci**	**Hallmark** Hallmark **Hallmark**	**Omega** Omega **Alpha**	**Spotify** Spotify **Spotify**

Each cell: reference name / before / after (bold).
Q1. What species or animal is shown in this image? Answer with the common name only.
Q2. Which fictional character is shown in this image? Answer with the name only.
Q3. What logo, brand, or company is shown in this image? Answer with the name only.

Figure 8: LLaVA-1.5-7B, Image-only: three complete Target/T1–T3 groups per method (21 groups). Cases differ from the main text; colors are explained in Appendix B.

B Additional visualization: Image+Text / Qwen3-VL-8B

Method / input	Target	T1 · closest	T2 · intermediate	T3 · distant
GA 1 D05/B	**Gorilla** Gorilla **gorilla gorilla**	**Chimpanzee** Chimpanzee **gorilla gorilla**	**Monkey** monkey **monkey**	**Dog** dog **dog**
2 D09/A	**Iron Man** Iron Man **Tony Stark**	**War Machine** War Machine **Okoye**	**Capt. America** Captain America **Steve Rogers**	**Link** Link **Link**
3 D14/A	**Puma** PUMA **PUMA**	**Jaguar** Jaguar **JAGUAR**	**Reebok** Reebok **Reebok**	**Coca-Cola** Coca-Cola **Coca-Cola**
GD 1 D02/A	**Kangaroo** Kangaroo **kangaroo**	**Wallaby** Wallaby **kangaroo**	**Wombat** Wombat **wombat**	**Ostrich** Ostrich **emua**
2 D10/A	**Naruto** Naruto Uzumaki **ninja**	**Boruto** Boruto Uzumaki **boruto**	**Sasuke** Sasuke Uchiha **Itachi**	**Winnie the Pooh** Winnie the Pooh **winnie the pooh**
3 D16/A	**Shell** Shell **SHELL**	**Chupa Chups** Chupa Chups **Chupa Chups**	**ExxonMobil** ExxonMobil **ExxonMobil**	**Meta / Facebook** Meta **Meta**
KLMin 1 D02/A	**Kangaroo** Kangaroo **kangaroo**	**Wallaby** Wallaby **kangaroo**	**Wombat** Wombat **wombat**	**Ostrich** Ostrich **Ostrich**
2 D07/A	**Batman** Batman **Batman**	**Robin** Robin **Robin**	**Black Panther** Black Panther **Black Panther**	**Homer Simpson** Homer Simpson **Homer Simpson**
3 D16/A	**Shell** Shell **SHELL**	**Chupa Chups** Chupa Chups **Chupa Chups**	**ExxonMobil** ExxonMobil **ExxonMobil**	**Meta / Facebook** Meta **Meta**
NPO 1 D05/A	**Gorilla** Gorilla **gorilla**	**Chimpanzee** Chimpanzee **Chimpanzee**	**Monkey** macaque **monkey**	**Dog** Bearded Collie **Dog**
2 D07/A	**Batman** Batman **BATMAN**	**Robin** Robin **Robin**	**Black Panther** Black Panther **Black Panther**	**Homer Simpson** Homer Simpson **Homer Simpson**
3 D15/A	**Marvel** MARVEL **MARVEL**	**Supreme** Supreme **Supreme**	**Zara** ZARA **ZARA**	**PlayStation** PlayStation **PlayStation logo**
DPO 1 D02/A	**Kangaroo** Kangaroo **I don't know.**	**Wallaby** Wallaby **I don't know.**	**Wombat** Wombat **I don't know.**	**Ostrich** Ostrich **Ostrich**
2 D08/A	**Homer Simpson** Homer Simpson **I don't know.**	**Bart Simpson** Bart Simpson **Bart Simpson**	**Peter Griffin** Peter Griffin **I don't know.**	**Naruto** Naruto Uzumaki **Naruto Uzumaki**
3 D12/A	**Mastercard** Mastercard **I don't know.**	**Flickr** Flickr **Flickr**	**Visa** Visa **Visa**	**Ford** Ford **Ford oval logo**
MMUnlearner 1 D04/A	**Chameleon** Chameleon **Chameleon**	**Green Iguana** Iguana **Iguana**	**Snake** Snake **Snake**	**Frog** Frog **Frog**
2 D06/A	**Spider-Man** Spider-Man **Spider-Man**	**Deadpool** Deadpool **Deadpool**	**Venom** Venom **Venom**	**Mario** Mario **Mario**
3 D14/A	**Puma** PUMA **PUMA**	**Jaguar** Jaguar **JAGUAR**	**Reebok** Reebok **Reebok**	**Coca-Cola** Coca-Cola **Coca-Cola logo**
MLLMEraser 1 D04/A	**Chameleon** Chameleon **Chameleon**	**Green Iguana** Iguana **Iguana**	**Snake** Snake **snake**	**Frog** Frog **Frog**
2 D10/A	**Naruto** Naruto Uzumaki **Naruto Uzumaki**	**Boruto** Boruto Uzumaki **Boruto Uzumaki**	**Sasuke** Sasuke Uchiha **Itachi Uchiha**	**Winnie the Pooh** Winnie the Pooh **Winnie the Pooh**
3 D15/A	**Marvel** MARVEL **MARVEL**	**Supreme** Supreme **Supreme**	**Zara** ZARA **ZARA**	**PlayStation** PlayStation **PlayStation**

Each cell: reference name / before / after (bold); answers are verbatim.
Input code = description group / template. Each role has its own description.
See the exact input key (D01-D16; templates A-F) following the case figures.

Figure 9: Qwen3-VL-8B, Image+Text: three complete Target/T1–T3 groups per method (21 groups). Input codes identify exact description groups and question templates (Figures 13–14); each role has its own description.

B Additional visualization: Image+Text / LLaVA-1.5-7B

Method / input	Target	T1 · closest	T2 · intermediate	T3 · distant
GA	**Swan** Swan **Duck**	**White Goose** Duck **Duck**	**Pelican** Pelican **Duck**	**Bald Eagle** Eagle **Eagle**
1 D03/A				
	Batman Batman **Bruce wayne**	**Robin** Robin **Robin**	**Black Panther** Black panther **Black panther**	**Homer Simpson** Homer simpson **Homer**
2 D07/B				
	Marvel Marvel **Marvel**	**Supreme** Supreme **Supreme**	**Zara** Zara **Zar**	**PlayStation** Playstation **Playstation**
3 D15/B				
GD	**Chameleon** Chameleon **Lizard**	**Green Iguana** Iguana **Lizard**	**Snake** Snake **Lizard**	**Frog** Frog **Frog**
1 D04/B				
	Homer Simpson Homer simpson **Bart simpson**	**Bart Simpson** Homer simpson **Bart simpson**	**Peter Griffin** Peter griffin **Peter griffin**	**Naruto** Naruto **Naruto uzumaki**
2 D08/B				
	Marvel Marvel **Marvel**	**Supreme** Supreme **Supreme**	**Zara** Zara **Zara**	**PlayStation** Sony **PlayStation**
3 D15/A				
KLMin	**Chameleon** Chameleon **Lizard**	**Green Iguana** Iguana **Lizard**	**Snake** Snake **Lizard**	**Frog** Frog **Frog**
1 D04/B				
	Spider-Man Spider-man **Spiderman**	**Deadpool** Deadpool **Deadpool**	**Venom** Venom **Venom**	**Mario** Mario **Mario**
2 D06/A				
	Nike Nike **Nike**	**Adidas** Adidas **Adidas**	**Under Armour** Under armour **Under armour**	**Rolex** Rolex **Crown**
3 D11/A				
NPO	**Kangaroo** Kangaroo **Koala**	**Wallaby** Kangaroo **Koala**	**Wombat** Kangaroo **Koala**	**Ostrich** Ostrich **Ostrich**
1 D02/A				
	Batman Batman **Batman**	**Robin** Robin **Robin**	**Black Panther** Black panther **Black Panther**	**Homer Simpson** Homer simpson **Homer Simpson**
2 D07/A				
	Marvel Marvel **Marvel**	**Supreme** Supreme **Supreme**	**Zara** Zara **Zara**	**PlayStation** Sony **PlayStation**
3 D15/A				
DPO	**Kangaroo** Kangaroo **Giraffe**	**Wallaby** Kangaroo **Wallaby**	**Wombat** Kangaroo **Wombat**	**Ostrich** Ostrich **Ostrich**
1 D02/A				
	Homer Simpson Homer simpson **Homer simpson**	**Bart Simpson** Homer simpson **Bart simpson**	**Peter Griffin** Peter griffin **Peter griffin**	**Naruto** Naruto **Naruto Uzumaki**
2 D08/A				
	BMW Bmw **Bmw**	**Volkswagen** Volkswagen **Volkswagen**	**Mercedes-Benz** Mercedes **Mercedes**	**Google** Google **Google logo**
3 D13/A				
MMUnlearner	**Kangaroo** Kangaroo **Koala**	**Wallaby** Kangaroo **Koala**	**Wombat** Kangaroo **Koala**	**Ostrich** Ostrich **Ostrich**
1 D02/A				
	Naruto Naruto **Naruto**	**Boruto** Naruto **Naruto**	**Sasuke** Anime **Anime**	**Winnie the Pooh** Winnie pooh **Winnie the Pooh**
2 D10/A				
	BMW Bmw **Bmw**	**Volkswagen** Volkswagen **Volkswagen**	**Mercedes-Benz** Mercedes **Mercedes**	**Google** Google **Google**
3 D13/A				
MLLMEraser	**Giant Panda** Panda **Panda**	**Brown Bear** Bear **Bear**	**Raccoon** Raccoon **Opossum**	**Kangaroo** Kangaroo **Kangaroo**
1 D01/A				
	Spider-Man Spider-man **Spiderman**	**Deadpool** Deadpool **Deadpool**	**Venom** Venom **Venom**	**Mario** Mario **Mario**
2 D06/A				
	BMW Bmw **Bmw**	**Volkswagen** Volkswagen **Volkswagen**	**Mercedes-Benz** Mercedes **Mercedes**	**Google** Google **Google**
3 D13/A				

Each cell: reference name / before / after (bold); answers are verbatim.
Input code = description group / template. Each role has its own description.
See the exact input key (D01-D16; templates A-F) following the case figures.

Figure 10: LLaVA-1.5-7B, Image+Text: three complete Target/T1–T3 groups per method (21 groups). Input codes identify exact description groups and question templates (Figures 13–14); each role has its own description.

B Additional visualization: Text-only / Qwen3-VL-8B

Method / input	Target	T1 · closest	T2 · intermediate	T3 · distant
GA	**Gorilla**	**Chimpanzee**	**Monkey**	**Dog**
1 D05/C	Gorilla	Chimpanzee	Monkey	Dog
	gorilla gorilla	**gorilla gorilla**	**monkey**	**dog**
2 D09/D	**Iron Man**	**War Machine**	**Capt. America**	**Link**
	Iron Man	Iron Man	Captain America	Link
	Tony Stark	**Grayson**	**Captain America**	**Link**
3 D14/C	**Puma**	**Jaguar**	**Reebok**	**Coca-Cola**
	Puma	Jaguar	Delta Force	Coca-Cola
	CATS	**JAGUAR**	**NIKE**	**COCA-COLA**
GD	**Kangaroo**	**Wallaby**	**Wombat**	**Ostrich**
1 D02/C	Kangaroo	Kangaroo rat	Wombat	Ostrich
	kangaroo	**kangaroo**	**wombat**	**Ostrich**
2 D10/C	**Naruto**	**Boruto**	**Sasuke**	**Winnie the Pooh**
	Naruto Uzumaki	Naruto Uzumaki	Itachi Uchiha	Winnie-the-Pooh
	Naruto	**Naruto**	**Itachi**	**Winnie the Pooh**
3 D16/C	**Shell**	**Chupa Chups**	**ExxonMobil**	**Meta / Facebook**
	Shell	Dali	ExxonMobil	Instagram
	ExxonMobil	**Dali**	**Exxon**	**Meta Platforms, Inc.**
KLMin	**Kangaroo**	**Wallaby**	**Wombat**	**Ostrich**
1 D02/C	Kangaroo	Kangaroo rat	Wombat	Ostrich
	kangaroo	**kangaroo**	**wombat**	**Ostrich**
2 D07/C	**Batman**	**Robin**	**Black Panther**	**Homer Simpson**
	Batman	Robin	T'Challa	Bart Simpson
	Batgirl	**Robin**	**T'Challa**	**Homer Simpson**
3 D16/C	**Shell**	**Chupa Chups**	**ExxonMobil**	**Meta / Facebook**
	Shell	Dali	ExxonMobil	Instagram
	ExxonMobil	**Dali**	**Exxon**	**Meta Platforms, Inc.**
NPO	**Gorilla**	**Chimpanzee**	**Monkey**	**Dog**
1 D05/F	Gorilla gorilla	Pan troglodytes	New World monkey	Canis lupus familiaris
	Pan troglodytes	**Pan troglodytes**	**monkey**	**Dog**
2 D07/F	**Batman**	**Robin**	**Black Panther**	**Homer Simpson**
	Batman	Robin	T'Challa	Bart Simpson
	BATMAN	**Robin**	**T'Challa**	**Homer Simpson**
3 D15/E	**Marvel**	**Supreme**	**Zara**	**PlayStation**
	Marvel Comics	Box Mark	Zara	PlayStation
	DC Comics	**BOXMARK**	**ZARA**	**PlayStation logo**
DPO	**Kangaroo**	**Wallaby**	**Wombat**	**Ostrich**
1 D02/C	Kangaroo	Kangaroo rat	Wombat	Ostrich
	Red kangaroo	**Bettongia gaimardi**	**Bilby**	**Ostrich**
2 D08/C	**Homer Simpson**	**Bart Simpson**	**Peter Griffin**	**Naruto**
	Bart Simpson	Bart Simpson	Burt Reynolds	Naruto Uzumaki
	I don't know.	**Bart Simpson**	**I don't know.**	**Naruto Uzumaki**
3 D12/C	**Mastercard**	**Flickr**	**Visa**	**Ford**
	Visa	Instagram	Visa	Ford
	Visa	**Instagram**	**Visa**	**Ford oval logo**
MMUnlearner	**Chameleon**	**Green Iguana**	**Snake**	**Frog**
1 D04/E	Chameleon	Iguana	Snake	Frog
	Chameleon	**Iguana**	**Lizard**	**Frog**
2 D06/D	**Spider-Man**	**Deadpool**	**Venom**	**Mario**
	Spider-Man	Mystique	Venom	Mario
	Spider-Man	**Deadpool**	**Venom**	**Mario**
3 D14/C	**Puma**	**Jaguar**	**Reebok**	**Coca-Cola**
	Puma	Jaguar	Delta Force	Coca-Cola
	CATSPEED	**JAGUAR**	**Delta Force**	**Coca-Cola**
MLLMEraser	**Chameleon**	**Green Iguana**	**Snake**	**Frog**
1 D04/C	Chameleon	Iguana	Lizard	Frog
	Chameleon	**Iguana**	**Snake**	**Frog**
2 D10/D	**Naruto**	**Boruto**	**Sasuke**	**Winnie the Pooh**
	Naruto Uzumaki	Naruto Uzumaki	Itachi Uchiha	Winnie-the-Pooh
	Naruto	**Kaito Orange**	**Kakashi Hatake**	**Winnie-the-Pooh**
3 D15/C	**Marvel**	**Supreme**	**Zara**	**PlayStation**
	Marvel Comics	Box Mark	Zara	PlayStation
	DC Comics	**Box Mark**	**Zara**	**PlayStation**

Each cell: reference name / before / after (bold); answers are verbatim.
Input code = description group / template. Each role has its own description.
See the exact input key (D01-D16; templates A-F) following the case figures.

Figure 11: Qwen3-VL-8B, Text-only: three complete Target/T1–T3 groups per method (21 groups). Input codes identify exact description groups and question templates (Figures 13–14); each role has its own description.

B Additional visualization: Text-only / LLaVA-1.5-7B

Method / input	Target	T1 · closest	T2 · intermediate	T3 · distant
GA	**Swan**	**White Goose**	**Pelican**	**Bald Eagle**
	Swan	Duck	Pelican	Bald eagle
1 D03/F	**Duck**	**Duck**	**Goose**	**Bald eagle**
	Batman	**Robin**	**Black Panther**	**Homer Simpson**
	Batman	Robin	T'Challa	Homer Simpson
2 D07/C	**Bruce Wayne**	**Robin**	**T'Challa**	**Homer Simpson**
	Marvel	**Supreme**	**Zara**	**PlayStation**
	Marvel	Supreme	Zara	PlayStation
3 D15/E	**Marvel**	**Supreme**	**Zara**	**Sony**
GD	**Chameleon**	**Green Iguana**	**Snake**	**Frog**
	Chameleon	Iguana	Salamander	Frog
1 D04/E	**Lizard**	**Iguana**	**Lizard**	**Frog**
	Homer Simpson	**Bart Simpson**	**Peter Griffin**	**Naruto**
	Homer Simpson	Bart Simpson	Frank	Naruto Uzumaki
2 D08/C	**Homer simpson**	**Bart Simpson**	**Frank belcher**	**Naruto Uzumaki**
	Marvel	**Supreme**	**Zara**	**PlayStation**
	Marvel	Supreme	Zara	PlayStation
3 D15/C	**DC Comics**	**Supreme**	**Zara**	**PlayStation**
KLMin	**Chameleon**	**Green Iguana**	**Snake**	**Frog**
	Chameleon	Iguana	Salamander	Frog
1 D04/E	**Lizard**	**Iguana**	**Lizard**	**Frog**
	Spider-Man	**Deadpool**	**Venom**	**Mario**
	Spider-Man	Deadpool	Venom	Mario
2 D06/C	**Spiderman**	**Deadpool**	**Venom**	**Mario**
	Nike	**Adidas**	**Under Armour**	**Rolex**
	Nike	Adidas	Under Armour	Rolex
3 D11/C	**Nike**	**Adidas**	**Under Armour**	**Rolex**
NPO	**Kangaroo**	**Wallaby**	**Wombat**	**Ostrich**
	Kangaroo	Kangaroo	Kangaroo	Ostrich
1 D02/E	**Koala**	**Wallabies**	**Koala**	**Ostrich**
	Batman	**Robin**	**Black Panther**	**Homer Simpson**
	Batman	Robin	T'Challa	Homer Simpson
2 D07/E	**The Riddler**	**Robin**	**T'Challa**	**Homer Simpson**
	Marvel	**Supreme**	**Zara**	**PlayStation**
	Marvel	Supreme	Zara	PlayStation
3 D15/C	**DC Comics**	**Supreme**	**Zara**	**PlayStation**
DPO	**Kangaroo**	**Wallaby**	**Wombat**	**Ostrich**
	Kangaroo	Kangaroo	Kangaroo	Ostrich
1 D02/C	**Kangaroo**	**Kangaroo**	**Wallabee**	**Ostrich**
	Homer Simpson	**Bart Simpson**	**Peter Griffin**	**Naruto**
	Homer Simpson	Bart Simpson	Frank	Naruto Uzumaki
2 D08/C	**Homer Simpson**	**Bart Simpson**	**Frank Murphy**	**Naruto Uzumaki**
	BMW	**Volkswagen**	**Mercedes-Benz**	**Google**
	BMW	Volkswagen	Mercedes-Benz	Google
3 D13/C	**BMW**	**Mercedes-Benz**	**Mercedes-Benz**	**Google**
MMUnlearner	**Kangaroo**	**Wallaby**	**Wombat**	**Ostrich**
	Kangaroo	Kangaroo	Kangaroo	Ostrich
1 D02/E	**Koala**	**Wallabies**	**Koala**	**Ostrich**
	Naruto	**Boruto**	**Sasuke**	**Winnie the Pooh**
	Naruto Uzumaki	Naruto	Naruto Uzumaki	Winnie the Pooh
2 D10/C	**Naruto**	**Naruto**	**Kakashi Hatake**	**Winnie the Pooh**
	BMW	**Volkswagen**	**Mercedes-Benz**	**Google**
	BMW	Volkswagen	Mercedes-Benz	Google
3 D13/C	**Bmw**	**Mercedes-benz**	**Mercedes-benz**	**Google**
MLLMEraser	**Giant Panda**	**Brown Bear**	**Raccoon**	**Kangaroo**
	Panda	Ursus arctos	Raccoon	Kangaroo
1 D01/E	**Panda**	**Brown bear**	**Raccoon**	**Kangaroo**
	Spider-Man	**Deadpool**	**Venom**	**Mario**
	Spider-Man	Deadpool	Venom	Mario
2 D06/C	**Spider-Man**	**Deadpool**	**Symbiote**	**Mario**
	BMW	**Volkswagen**	**Mercedes-Benz**	**Google**
	BMW	Volkswagen	Mercedes-Benz	Google
3 D13/C	**BMW**	**Volkswagen**	**Volkswagen**	**Google**

Each cell: reference name / before / after (bold); answers are verbatim.
Input code = description group / template. Each role has its own description.
See the exact input key (D01-D16; templates A-F) following the case figures.

Figure 12: LLaVA-1.5-7B, Text-only: three complete Target/T1–T3 groups per method (21 groups). Input codes identify exact description groups and question templates (Figures 13–14); each role has its own description.

B Exact text input key (1/2)

Exact question templates A-B: Image + Text

A. Description: {description}
What specific entity is shown in the image? Answer with the name only.

B. Use the image and this description to identify the entity.
Description: {description}
Answer with the exact name only.

Replace {description} with the exact text in the corresponding role below.

D01 | biology01			
Target	**T1**	**T2**	**T3**
large black and white bamboo eating bear native to mountain forests of China	large omnivorous ursid species including grizzly and Eurasian populations	small North American mammal with ringed tail and black facial mask	Australian marsupial with powerful hind legs, pouch, and hopping locomotion
D02 | biology06			
Target	**T1**	**T2**	**T3**
Australian marsupial with powerful hind legs, pouch, and hopping locomotion	small to medium hopping marsupial related to larger Australian macropods	burrowing Australian marsupial with cube shaped droppings and sturdy body	large flightless African bird with long legs and very fast running
D03 | biology09			
Target	**T1**	**T2**	**T3**
large graceful waterbird with long curved neck and mostly white plumage	domesticated pale feathered waterfowl often found on farms	large waterbird with oversized throat pouch used for catching fish	North American raptor with white head and tail used as United States national symbol
D04 | biology12			
Target	**T1**	**T2**	**T3**
lizard known for independently moving eyes, long tongue, and color change	large arboreal lizard from the Americas with a dorsal crest	legless reptile with elongated body and forked tongue	tailless amphibian that jumps and often lives near water
D05 | biology14			
Target	**T1**	**T2**	**T3**
largest living great ape, ground dwelling and native to African forests	African great ape closely related to humans and known for tool use	primate with tail in many species, generally smaller than apes	domesticated canid kept as companion and working animal
D06 | ip_character03			
Target	**T1**	**T2**	**T3**
web slinging Marvel superhero from New York with red and blue costume	wisecracking masked Marvel antihero known for breaking the fourth wall	black alien symbiote antihero associated with a toothy monstrous appearance	red capped mustached plumber hero from Nintendo platform games
D07 | ip_character04			
Target	**T1**	**T2**	**T3**
masked billionaire vigilante from DC Comics who fights crime in Gotham City	young sidekick from Gotham superhero stories often wearing red, green, and yellow	Wakandan king and Marvel superhero wearing a black vibranium suit	bald yellow cartoon father from an animated Springfield family sitcom
D08 | ip_character09			
Target	**T1**	**T2**	**T3**
bald yellow cartoon father from an animated Springfield family sitcom	spiky haired mischievous boy from an animated Springfield family sitcom	heavyset cartoon father from an animated Rhode Island family sitcom	orange clad ninja protagonist from a Japanese manga series who seeks village leadership

Figure 13: Exact inputs for Image+Text and Text-only cases. The D code selects a group, the column selects a role, and the letter selects a template. Substitute that description verbatim for {description}; Image+Text also uses the image in the corresponding case cell.

B Exact text input key (2/2)

Exact question templates C-F: Text only

C. Description: {description}
What specific entity is being described? Answer with the name only.

D. Description: {description}
What is the name of this entity? Answer with the exact name only.

E. Description: {description}
Which entity does this description refer to? Answer with only the entity name.

F. Description: {description}
Identify the described entity. Give only the specific name.

Replace {description} with the exact text in the corresponding role below.

D09 | ip_character10			
Target	**T1**	**T2**	**T3**
armored billionaire superhero from Marvel Comics using powered metal suits	gray armored military superhero associated with powered suits and heavy weapons	patriotic shield carrying Marvel super soldier from the Second World War era	green clad Hylian hero from fantasy adventure games who wields a sword
D10 | ip_character11			
Target	**T1**	**T2**	**T3**
orange clad ninja protagonist from a Japanese manga series who seeks village leadership	young ninja protagonist who is the son of a previous orange clad hero	black haired rival ninja from a Japanese manga clan with special eye powers	honey loving teddy bear character from stories set in the Hundred Acre Wood
D11 | trademark05			
Target	**T1**	**T2**	**T3**
curved checkmark symbol used by a major athletic footwear and sportswear company	three parallel stripe emblem used by a German sportswear brand	interlocking U and A emblem used by an athletic apparel company	five pointed crown emblem used by a Swiss luxury watchmaker
D12 | trademark06			
Target	**T1**	**T2**	**T3**
two overlapping red and yellow circles mark used by a global payment card network	two dot pink and blue mark used by a photo sharing service	blue payment card wordmark used by a global electronic transaction network	blue oval automotive wordmark used by an American car manufacturer
D13 | trademark08			
Target	**T1**	**T2**	**T3**
round blue white and black automotive badge associated with a Bavarian manufacturer	stacked V over W roundel used by a German automobile manufacturer	three pointed star emblem inside a circle used by a German automobile brand	multicolored search engine wordmark using blue red yellow and green letters
D14 | trademark10			
Target	**T1**	**T2**	**T3**
leaping cat silhouette used by a German athletic footwear and apparel brand	leaping wild cat emblem used by a British luxury automobile marque	athletic footwear and apparel mark historically associated with a vector or delta shape	red and white cursive wordmark used by a classic cola soft drink brand
D15 | trademark13			
Target	**T1**	**T2**	**T3**
red rectangular white letter mark used by a superhero comics and film publisher	red rectangular white word streetwear box mark used by a New York skate fashion brand	high fashion black serif wordmark used by a Spanish clothing retailer	stylized overlapping P and S symbol for a Japanese video game console brand
D16 | trademark14			
Target	**T1**	**T2**	**T3**
yellow and red scallop emblem used by a petroleum company	daisy shaped yellow and red lollipop brand mark designed by Salvador Dali	red and blue oil company wordmark with crossing x letters	infinity shaped blue loop mark for a social networking parent company

Figure 14: Exact inputs for Image+Text and Text-only cases. The D code selects a group, the column selects a role, and the letter selects a template. Substitute that description verbatim for {description}; Image+Text also uses the image in the corresponding case cell.